OXFORD

CONCEPTUAL ENGINEERING *and* CONCEPTUAL ETHICS

EDITED BY

alexis BURGESS

herman CAPPELEN

david PLUNKETT

Conceptual Engineering and Conceptual Ethics

Conceptual Engineering and Conceptual Ethics

EDITED BY

Alexis Burgess, Herman Cappelen, and David Plunkett

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries

First Edition published in 2020
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2019946746

ISBN 978–0–19–880185–6

DOI: 10.1093/oso/9780198801856.003.0001

Printed and bound in Great Britain by
Clays Ltd, Elcograf S.p.A.

Contents

Note to Readers

No perfect partition of the table of contents emerged in the editing process. This reflects the fact that we have a rich field on our hands, with a variety of cross-cutting themes. The Introduction details many of these. But here we want to offer some brief guidance for readers eager to dig in. (An important proviso: the chapters listed below are obvious places to start; they aren't meant to imply that no other chapters contain material relevant to the indicated topics.)

If you're coming to the volume unsure about the status or nature of the whole field, you might start with, in addition to the Introduction, the chapters by Braddon-Mitchell, Cappelen, Pérez Carballo, Pettit, Richard, Thomasson, Sawyer, and Scharp. The chapters by Cappelen, Sawyer, Scharp, and Thomasson generally represent a pro-attitude toward the field, while chapters by Ball and Greenough develop some skepticism.

The chapters by Burgess, Dever, and Sterken take up various methodological challenges for doing conceptual engineering and conceptual ethics.

If you're coming to the volume with more applied interests, you might consult the chapters by Belleri, Eklund, and Thomasson for work on metaphysics, or the chapters by Brigandt and Rosario, Díaz-León, Haslanger, Jackson, McPherson and Plunkett, and Pettit for work on ethics and social/political philosophy.

Contributors

Derek Ball is a Senior Lecturer in Philosophy at the University of St Andrews.

Delia Belleri is a fixed-term Assistant Professor of Philosophy at the University of Vienna.

David Braddon-Mitchell is Professor of Philosophy at the University of Sydney.

Ingo Brigandt is Professor in philosophy and Canada Research Chair in philosophy of biology at the University of Alberta.

Alexis Burgess is an independent philosopher.

Herman Cappelen is a Professor of Philosophy at the University of Oslo, and at the University of St Andrews (1/5th time).

Josh Dever is a Professor of Philosophy at the University of Texas.

E. Díaz-León is an Associate Professor in the Department of Philosophy at the University of Barcelona.

Matti Eklund is Professor of Philosophy at Uppsala University.

Patrick Greenough is a Senior Lecturer in Logic and Metaphysics at the University of St Andrews.

Sally Haslanger is the Ford Professor of Philosophy in the Department of Linguistics and Philosophy at the Massachusetts Institute of Technology.

Frank Jackson is Emeritus Professor of Philosophy at the Australian National University, Canberra.

Tristram McPherson is Associate Professor in the Department of Philosophy at Ohio State University.

Alejandro Pérez Carballo is Assistant Professor in Philosophy at the University of Massachusetts, Amherst.

Philip Pettit is L.S. Rockefeller University Professor of Human Values at Princeton University and Distinguished University Professor of Philosophy at the Australian National University, Canberra.

David Plunkett is Associate Professor in the Department of Philosophy at Dartmouth College.

Mark Richard is Professor in the Department of Philosophy at Harvard University.

Esther Rosario is a PhD student in the Department of Philosophy at the University of Alberta.

Sarah Sawyer is a Senior Lecturer in Philosophy at the University of Sussex.

Kevin Scharp is Reader in Philosophy and Director of Arché at the University of St Andrews.

Rachel Katharine Sterken is an Associate Professor in the Department of Philosophy at the University of Oslo.

Amie L. Thomasson is Professor in the Department of Philosophy at Dartmouth College.

Acknowledgements

Many friends, colleagues, and institutions helped us put this volume together. ConceptLab at the University of Oslo was particularly important: in the spring of 2016, ConceptLab hosted a workshop where many of the papers were presented and discussed. Several of the contributions were also discussed in other events hosted by ConceptLab. We are particularly grateful to Derek Ball, David Braddon-Mitchell, Matti Eklund, Øystein Linnebo, Tristram McPherson, Olav Gjelsvik, Kevin Scharp, Rachel Sterken, Tim Sundell, and Amie Thomasson. We are also grateful to Matthew McKeever for his help throughout the process of putting this volume together, and to Ira Richardson and Max Frye for their help with the index. Finally, we got invaluable help from our editor at OUP, Peter Momtchiloff.

1

Introduction

A Guided Tour of Conceptual Engineering and Conceptual Ethics

Herman Cappelen and David Plunkett

Introduction

In *The Will to Power*, Nietzsche writes the following:

> Philosophers . . . have trusted in concepts as completely as they have mistrusted the senses: they have not stopped to consider that concepts and words are our inheritance from ages in which thinking was very modest and unclear. . . . What dawns on philosophers last of all: they must no longer accept concepts as a gift, nor merely purify and polish them, but first make and create them, present them and make them convincing. Hitherto one has generally trusted one's concepts as if they were a wonderful dowry from some sort of wonderland: but they are, after all, the inheritance from our most remote, most foolish as well as most intelligent ancestors. . . . What is needed above all is an absolute skepticism toward all inherited concepts.[1]

Nietzsche here articulates a radical skepticism about all inherited concepts. Philosophers should question whether the concepts we have are good enough and should engage in conceptual critique. What emerges, thinks Nietzsche, is the following: we should not just improve the concepts we've been given, reforming or "polishing" them in minor ways, but also create new ones—concepts not tainted by the "most foolish of our ancestors".

Even if you think Nietzsche's claim is more than a bit hyperbolic, you might think some more moderate version of his view is justified. For example: maybe *some* of the concepts we have inherited are defective, or at least not as good as they could be for our current purposes. In contrast, maybe you think Nietzsche is wildly off here in his radical stance. Maybe we have good reason to think that our current stock of concepts is just fine for the purposes at hand. Or maybe you think that, even if that stock of concepts could be better, it doesn't make sense to think about changing our concepts—or at least not the core concepts that really

[1] Nietzsche (1901/1968: 220–1, section 409). Thanks to Michael Beaney for pointing us to this passage.

Herman Cappelen and David Plunkett, *Introduction: A Guided Tour of Conceptual Engineering and Conceptual Ethics* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Herman Cappelen and David Plunkett.
DOI: 10.1093/oso/9780198801856.003.0001

matter in philosophy (e.g., TRUTH, MEANING, CONTENT, or VALUE).[2] The issues that this quote from Nietzsche brings up thus includes questions such as: What are the dimensions of assessment for concepts? Which philosophical concepts are defective and how can we improve them? How important are facts about the history (or "genealogy") of our use of concepts to the assessment of our current concepts? These are questions at the heart of the fields that we call "conceptual engineering" and "conceptual ethics".

If you care about these and related questions, this volume should be of interest to you. It is the first ever volume devoted entirely to conceptual engineering and conceptual ethics. Our hope is that it will help shape and promote what we (the editors) take to be an important, exciting, and underexplored part of philosophy. In this Introduction, we first try to delineate the field and explain why we are using two different expressions ('conceptual engineering' and 'conceptual ethics') to describe the topics in the book. We then turn to some of the central foundational issues that arise for conceptual engineering and conceptual ethics, and finally we outline various views one might have about their role in philosophy and inquiry more generally.

1. 'Conceptual Engineering' and 'Conceptual Ethics'

The title of this volume uses two expression to describe its topic: 'conceptual engineering' and 'conceptual ethics'. Why? The answer to this is not straightforward. We don't think these expressions come with fixed meanings. The previous literature has used them in different ways and so do the authors in this volume. These terms are often used without precise definitions by those working in the field. Moreover, when they are given more precise definitions by philosophers, these definitions often contradict those given by others. As editors, we could have played the terminology police for those contributing to this volume. But that would be an exercise in futility. Instead, we have decided to let a thousand (or at least a few) flowers bloom. Contributors use central terms, such as 'conceptual engineering', 'conceptual ethics', 'revision', and 'amelioration', in different ways, often explicitly so. That's how it should be given that this is currently a fast moving literature involving philosophers from many different background and sub-fields. That said, for the purposes of this Introduction, we will offer our own characterizations of conceptual engineering and conceptual ethics, with one of us (Cappelen) offering a characterization of conceptual engineering, and the other of us (Plunkett) offering one of conceptual ethics.[3] The basic reason we split up our discussion in this way is that one of us (Cappelen) likes to use the expression "conceptual engineering", whereas the other one (Plunkett) thinks

[2] In this chapter, we use small caps (e.g., CAT) to pick out concepts, single quotation marks (e.g., 'cat') strictly to mention linguistic items, and double quotation marks (e.g., "cat") for a variety of tasks, including quoting others' words, scare quotes, and mixes of use and mention.

[3] Our use of the term 'conceptual engineering' stems from Cappelen (2018), drawing chiefly on Scharp (2013) and Eklund (2015). Simon Blackburn also briefly uses 'conceptual engineering' in the opening pages of Blackburn (2001) in a related way, as does Brandom (2001). Our use of the term 'conceptual ethics' draws from Burgess and Plunkett (2013a,b).

that a number of the issues involved are best described as issues in "conceptual ethics" rather than "conceptual engineering".

Conceptual Engineering

According to Cappelen (2018), *conceptual engineering is concerned with the assessment and improvement of concepts*. However, since it's unclear and controversial what concepts are (and whether there are any), it's better to broaden the scope along the following lines:

> **Conceptual engineering** = (i) The assessment of representational devices, (ii) reflections on and proposal for how to improve representational devices, and (iii) efforts to implement the proposed improvements.[4]

Here are some issues that are central for those working on conceptual engineering:

- What are the relevant representational devices? Possible answers include: concepts (as they are construed in some part of psychology or philosophy), lexical items, and the semantic values of lexical items.[5] A closely connected cluster of questions concerns whether they are in language or thought or both. Different conceptual engineers will give different answers and that will have enormous implications for how the field is understood and practiced.
- Given an answer to the first cluster of questions, we can ask: What kinds of defects can representational devices have? Throughout the history of philosophy, a variety of defects have been proposed: cognitive defects (that undermine our ability to reason properly), moral or political defects (that undermine moral or political values of various sorts), theoretical defects (that undermine progress within some theoretical field), or semantic defects (where the semantic value is incoherent, incomplete, or missing). For illustrations of all of these and a more detailed taxonomy of defects, see Cappelen (2018: chapter 2).
- Once you have detected a defect in a representational device you care about, it's natural to think about how to improve it. What are the ameliorative strategies? There are four basic options once you've identified a defect in C: (i) Do nothing—just live with it (can't improve it, can't get rid of it), (ii) Abandonment of C (it's so defective it can't be improved), (iii) Improvement of C, (iv) Replacement of C (for certain purposes, in certain contexts).
- Once you have settled on an ameliorative strategy, you might want to do some work to implement it, that is, you might want to engage in a bit of activism on behalf of your ameliorative strategy. If that's something you want to do, it raises an 'implementation challenge': how are ameliorative strategies best implemented?

[4] Why call it 'conceptual' engineering when it's about representational devices more generally? Purely for aesthetic reason: 'representational devices engineering' doesn't roll off the tongue in the way 'conceptual engineering' does.

[5] Cappelen (2018) suggests we think of the basic case as that of revising the extensions and intensions of expressions.

If you're interested in conceptual engineering, you don't need to focus on all of these issues. Some will focus on discovering defects, some on ameliorative strategies, others on conceptual activism, and yet others want to do the whole shebang.

Conceptual engineering is not usefully construed as a branch of any other part of philosophy. It will draw on insights from philosophy of language, philosophy of mind, epistemology, political philosophy, philosophy of science, ethics, and other fields. That, of course, is also true about these other fields (i.e., they will draw on insights from each other). A case can, however, be made that conceptual engineering is prior to or more fundamental than all other philosophical disciplines. The argument is simple and obvious: reflection and argumentation in any part of philosophy must rely on concepts (epistemology relies heavily on, e.g., KNOWLEDGE and JUSTIFICATION; ethics on, e.g., OUGHT and BAD; and so on for each branch of philosophy.) As Nietzsche correctly observes in the passage at the beginning of this Introduction: it's intellectually irresponsible to throw yourself headlong into an inquiry before questioning the concepts you're using in that inquiry. So conceptual engineering, as construed here, comes first.[6]

Conceptual Ethics

Now that we have a rough characterization of conceptual engineering from one of us (Cappelen), here is a rough one of conceptual ethics, endorsed by the other one of us (Plunkett), drawing on previous co-authored work with Alexis Burgess.[7] Broadly, conceptual ethics concerns a range of normative and evaluative issues about thought, talk, and representation. Those include issues about which concepts we should use, ways in which concepts can be defective, what we should mean by our words, and when we should refrain from using certain words. (Which issues one thinks belong on this list, as well as how these issues are related to each other, will obviously depend on one's further philosophical commitments.) As the label suggests, some of the core issues in conceptual ethics concern *concepts* (assuming, for now, that there are such things). These include, centrally, normative issues about which concepts one should use (and why) and evaluative issues about which concepts are better than others (and why). Concepts can here be understood in rough terms as *constituent components of thoughts*, leaving it for different theorists to fill out that schematic characterization in different ways. As with conceptual engineering, parallel questions can of course arise for other representational devices beyond concepts (e.g., words).

The use of the term 'ethics' here in 'conceptual ethics' is meant very broadly, to cover "both the study of what one should or ought to do (dually, what can permissibly be done) as well as the study of which actions and outcomes are good or bad, better or worse".[8] Thus, this use of 'ethics' is *not* meant to privilege moral/political norms in particular (vs., e.g., those that find their central home in epistemology, metaphysics, aesthetics, etc.).[9]

[6] This raises tricky questions about the nature of the concepts used to think about conceptual engineering—for more on that, see Cappelen (2018: chapter 1).

[7] Burgess and Plunkett (2013a,b).

[8] Burgess and Plunkett (2013a: 1094).

[9] 'Conceptual ethics' is obviously not an ideal label. Many will still hear 'ethics' more narrowly—for example, as tied to distinctively *practical* norms of the sort that have their home in moral and political

Conceptual ethics is a branch of normative and evaluative inquiry, just as (at least certain parts of) epistemology, aesthetics, ethics, and political philosophy can be understood as branches of it. Thus, just as with other branches of normative and evaluative inquiry, people can approach conceptual ethics with very different philosophical commitments, very different views about how to make progress in it, and very different substantive views within it. Moreover, they can also approach it for very different reasons. For example, someone might be interested in conceptual ethics purely as an interesting part of philosophical theorizing. Or one might be interested in it because one is trying to actually change existing thought and talk.

This second point brings out an important aspect of the relation between conceptual ethics and those components of conceptual engineering that involve trying to actually change conceptual or linguistic practices. If one has such practical aims, then studying conceptual ethics might help. This is parallel to how studying normative political philosophy might help those interested in changing actual existing political institutions, or how studying normative aesthetics might help those creating art. But just as it would be a mistake to think of political philosophy *solely* in terms of the role it might play for the project of creating better political institutions, so too would it be a mistake to think of conceptual ethics solely in terms of the role it might play in practical projects of changing actual conceptual or linguistic practices. Conceptual ethics is a free-standing area of normative and evaluative inquiry, and some questions in it that might turn out to be of little use to those involved in actually trying to change conceptual or linguistic practices, or to those trying to engineer concepts.

In what follows, we will use the term 'conceptual engineering' and 'conceptual ethics' in roughly the ways introduced here. How exactly conceptual engineering and conceptual ethics relate to each other is something which there is live debate about in this area, and which the editors of this volume themselves have spent considerable time arguing about. Many of those engaged in these debates use the term 'conceptual engineering' and 'conceptual ethics' in ways that differ from our own. With that in mind, it should be emphasized that all of the issues we discuss below can be translated into other ways of using the labels 'conceptual engineering' and 'conceptual ethics' preferred by different philosophers (including some in this volume).

2. A Bottom-Up Characterization of Conceptual Engineering and Conceptual Ethics

We just gave what can be thought of as a top-down characterization of the topic of this volume. Another way to approach this is through examples. We could point to a range of paradigmatic cases as exemplifying conceptual engineering and conceptual ethics. We then hope that audiences will be able to find some relevant cluster of similarities between the examples.[10] Here are four paradigmatic cases:

philosophy. But other possible labels have their own drawbacks. For example: 'conceptual assessment', another possible label here, makes it sound as if this area solely concerns *evaluative* claims about concepts (e.g., which concepts are better than others), leaving out *normative* claims (e.g., about which concepts an agent should use).

[10] See Burgess and Plunkett (2013a) for a similar approach.

- Carnapian explication is an effort to improve on defective meanings. For Carnap the central defects have to do with vagueness and indeterminacy. Improvements—what Carnap calls "explications"—produce better meanings. Improvements for Carnap focus primarily on theoretical virtues.[11]
- Much of Sally Haslanger's work on race and gender has two components: it criticizes our gender and race concepts, and then suggests ameliorative strategies to improve those concepts. The defects she finds have to do with social and political effects of the meanings, and the ameliorations are also measured along those dimensions (e.g., by how much they can help us in the pursuit of social justice).[12]
- Peter Railton argues that moral philosophy should involve a methodology that is largely continuous with that of the natural and social sciences (this is the core of his methodological naturalism). Based on what he sees as the best practices within scientific inquiry, he argues that, in doing moral philosophy, we shouldn't just rely on our folk concepts. Instead, we should reform the meanings of our words to zero in on the topics that really matter, and in providing explanations of the phenomena at hand. Railton then offers improved moral language that can be used in this way, including, for example, reforming definitions of key terms such as 'moral goodness' and 'morality'.[13]
- Matti Eklund and Kevin Scharp explore the idea that TRUTH is inconsistent. If it is, that might be a serious defect (at least in some contexts). What might be needed is an improved, consistent, truth concept (or, in Scharp's case, multiple concepts). We might, they claim, be okay with using an inconsistent concept in certain areas of our life. But, for the purposes of doing advanced theoretical work in such areas as linguistics and logic, it would be better to avoid doing that, if possible.[14]

Here's the bottom-up way to introduce the topics of this volume: It's about *that kind of activity* or *these kinds of issues*. The kind of thing they are doing, or the kinds of things they are discussing.[15] The assumption then is that those activities form an interesting kind—a kind of activity or method or subject matter (or maybe all three at once). The idea of there being *some* sort of interesting kind here is a working hypothesis. One challenge for those working in the field is to try to substantiate it. Some chapters in this volume support that assumption, and some argue against it.

[11] See Carnap (1947/1956).

[12] See Haslanger (2000). It should be noted that Haslanger's views on conceptual engineering (and associated issues about the nature of concepts) have changed over time. So what we present here is only one strand of her thinking about the topic. See the collected papers in Haslanger (2012) for a fuller view of her thinking on the topic.

[13] See Railton (1986a,b), as well as many of the other papers collected in Railton (2003).

[14] See Eklund (2002) and Scharp (2013).

[15] To keep this Introduction at manageable length, we have chosen to not go into great detail of specific cases. But to get a real sense of how the bottom-up approach would work, more details would obviously be needed. For more discussion of examples, see Burgess and Plunkett (2013a,b) and Cappelen (2018).

3. Central Challenges

We turn now to what we take to be some of the central challenges in the areas of conceptual engineering and conceptual ethics. In what follows, we will often put the issues involved in terms of "conceptual engineering" rather than "conceptual ethics". We do so both for ease of exposition, and because of our belief that engaging with issues in conceptual ethics is an important part of conceptual engineering. However, it should be kept in mind that many of the issues we discuss below apply equally to conceptual engineering and conceptual ethics.

As a heuristic, it's useful to divide these into two categories:

- Category 1: Domain-general issues.
- Category 2: Issues that arise in thinking about the evaluation and engineering of specific concepts or groups of concepts (e.g., race and gender concepts).

As we'll emphasize below, there's arguably no sharp distinction here, and (at the very least) there are many important connections between category 1 and category 2. It's a rough division. Our focus here (and in the volume as a whole) is on category 1, and how it interacts with category 2. Category 2 is huge, and the details are too diverse for us to even begin to cover them in this brief introductory essay.

In what follows, we discuss six clusters of interconnected issues that arise in work on conceptual engineering and conceptual ethics. Many of these overlap and all of them interact in various ways. It goes beyond the scope of this Introduction to explore all those important connections, so we simply list a range of issues that strike us as both interesting and important, and that we think will be central to debates in conceptual engineering and conceptual ethics in the years ahead.[16]

Cluster 1

What are the objects being assessed and improved (and do they exist)? If the aim of conceptual engineering is to assess and improve concepts—or other representational devices—then we are ultimately on the hook for an account of what these objects are. There are very many theories of concepts.[17] Moreover, the term 'concept' is used in a variety of ways in philosophy, psychology, and ordinary speech. It is clear that not everyone in the debate means the same thing by the word 'concept', or is offering a theory of the same thing. If the objects of assessment are concepts, then one needs a theory of these things that makes it possible for them to be assessed and improved. Furthermore, one needs such a theory on which either their identity conditions are compatible with the idea of amelioration of a single concept, or where it makes sense to think of moving from one concept to another as a form of improvement. If the objects are not concepts, but something else, then that "something else" needs to be characterized (and, again, in a way that meets the above constraint that makes sense of the possibility of some kind of amelioration). For example, if the relevant

[16] Our list of central issues here draws heavily on our previous work, especially Burgess and Plunkett (2013a,b) and Cappelen (2018).

[17] See Margolis and Laurence (1999) for a collection of papers advocating different approaches.

objects are words or meanings, then those things also need to be characterized. Moreover, *whatever* the objects are that are being criticized and improved, there is the threat of skepticism that they actually exist.[18] (Think here, e.g., about Quinean skepticism about the analytic/synthetic distinction as the basis for a certain kind of skepticism about the existence of word "meanings".[19])

Concepts vs. conceptions vs. beliefs: Many people involved in conceptual engineering distinguish concepts from beliefs—for example, understanding concepts as "constituent components" of beliefs (or other attitudes) in one way or another.[20] Many others also include "conceptions" as part of the picture here. How exactly are these (and related) things distinguished? How *should* they be distinguished? If the answers here turn out certain ways, then perhaps many instances of conceptual engineering turn out not to be about engineering "concepts" at all, but rather something else.

Metasemantic foundations: Some theorists think of "metasemantics" as the study of the metaphysical foundations of meaning: it provides various accounts of what makes it the case (or grounds) that our words have the meanings they have.[21] So understood, metasemantics is very important for conceptual engineering (and for conceptual activists in particular): in order to change a meaning, you need to act on (or change in some way) that metaphysical foundation. So different views of metasemantics will generate different views of what conceptual engineering consists in. The parallel point applies about the import of work on the metaphysical foundations of other kinds of content, if one is seeking to assess and improve those kinds of content.

Internalism vs. externalism: An important issue in metasemantics is the distinction between internalism and externalism. In rough terms, according to internalists, meaning depends on facts about the individual (e.g, facts about "what's in her head"). Externalists deny this: they think meaning is determined at least in part by facts having to do with the history of linguistic usage, or complex use patterns over time, or the judgments of experts, or other things that are not individualistic. This divide might be put in terms of grounding (roughly, asymmetric metaphysical dependence of a certain kind) or supervenience (understood as a purely modal notion). There are important theoretical choice points here in the internalism vs. externalism debate, and the answers will profoundly affect how one thinks about and practices conceptual engineering.[22] Some of the key issues here extend beyond just whether one is an internalist or externalist (about either mental or linguistic content). They also include intramural debates among each camp. For example, does the correct

[18] An important question here is whether it is easier to meet this challenge when moving to one of these other objects instead of concepts. One of us (Cappelen) thinks it is. He argues that conceptual engineering concerns extensions/intensions of lexical items, rather than concepts. See Cappelen (2018) for discussion.

[19] See, for example, Quine (1951).

[20] For example, see the gloss of concepts in Burgess and Plunkett (2013a).

[21] See Burgess and Sherman (2014).

[22] The literature here is vast. An overview and helpful bibliography is Lau and Deutsch (2016).

externalism involve facts about the future usage of speakers,[23] or is it perhaps tied to "correct" theories of the relevant subject matter in some (perhaps indirect) way.[24]

Normativity of meaning: How sharp is the distinction between doing descriptive vs. normative work when engaged in the study of concepts, or of other representational devices? On some views, they will necessarily be deeply intertwined. On others, they are pretty far apart. And there is obviously room for a lot of middle ground here. Different theories in metasemantics (either about words or concepts) help inform these different views, as do different views about our thought and talk about semantics. For example, consider discussions about the "normativity of meaning".[25] If claims about meaning (or about mental content) are themselves always normative in some sense, how does this affect our understanding of the relationship of descriptive and normative work on concepts (or other representational devices)? For example: does it collapse the distinction between normative theorizing about concepts ("conceptual ethics") and descriptive theorizing about concepts? Or are the kinds of normative issues here importantly distinct in some sense?

Within our control vs. outside of our control: Many of those working on conceptual engineering are interested not just in general theories of conceptual engineering, but in actually bringing about conceptual or linguistic change. If that is part of one's motivation, it's important to get clear on whether conceptual change (or meaning change) is something that is within our control. And, if so, how *much* of it can we control, and how well can we control it? We need to get clear on the extent to which these kinds of changes are governed by our decisions, intentions, agreements, and preferences. Maybe, instead, they are governed by mechanisms that are difficult to understand and outside of our control.[26] One way these issues about control matter is for evaluating strategies for conceptual activism.

Cluster 2

What are the norms, goods, values, etc., that determine the normative/evaluative facts in conceptual ethics? Philosophers appeal to a wide variety of good, norms, values, etc. in making normative/evaluative claims in conceptual ethics. For example, some appeal to facts about what fundamental reality is like, independent of our thinking about it.[27] Others appeal to practical considerations of what would aid us in theoretical inquiry on a given topic, for example, by helping us zero in on an important set of issues, or helping us avoid false beliefs, or helping us smoothly communicate with other inquirers.[28] Others still appeal to practical considerations about the ethical/political effects of the use of certain concepts—for example, the way that the use of certain concepts might help promote social justice, freedom, or happiness.[29] These different

[23] See, for example, the sort of view advocated for by Derek Ball in his chapter, "Revisionary Analysis without Meaning Change (Or, Could Women Be Analytically Oppressed)?" (Chapter 2, this volume).

[24] See, for example, the sort of view advocated for in Schroeter and Schroeter (2014).

[25] See Kripke (1982) and Gibbard (2012) for defenses of the idea that "meaning is normative", and for criticisms of the idea, see, for example, Boghossian (1989) and Hattiangadi (2017).

[26] For an argument that much of the relevant changes are not in our control, see Cappelen (2018).

[27] See, for example, Sider (2012).

[28] See, for example, Eklund (2002) and Scharp (2013).

[29] See, for example, Haslanger (2000).

values/norms can interact in any number of ways. For example: it might be that using concept C helps promote social justice because it helps us keep track of important features of social reality we should be studying. This diversity of goods/values/norms raises the question of which of them ultimately matter in conceptual ethics. And how much is there is a general answer here, anyway, as opposed to answers for specific kinds of concepts (e.g., for race and gender concepts, or for truth concepts, etc.)?

What are potential defects and virtues of concepts? What counts as a defect of a concept and what are the potential dimensions of improvement? In the history of this topic, there have been a broad range of proposals. We can classify these into four rough categories: moral-political (e.g., hindering or promoting social justice), epistemological (e.g., hindering or promoting the acquisition of knowledge), cognitive (e.g., hindering or promoting good cognitive functioning), metaphysical defects (e.g., corresponding or not to joints in reality). Which of these proposals help us locate (and explain) genuine defects, and how are those defects related to each other?[30]

How much do aims matter? We use concepts in very different contexts, in cases where we aim to accomplish very different things—for example, making progress in mathematical inquiry vs. winning an election vs. trying to help build a better society. Many claims in conceptual engineering appeal to the aims an agent has in using a given concept (or set of concepts).[31] But how much do the aims an agent have matter here? What if (e.g.) she has immoral aims, or aims that won't be productive to furthering inquiry? Maybe what matters here are the aims an agent *should* have? Or maybe aims of any sort (either the ones an agent has, or the ones an agent should have) don't play any sort of fundamental explanatory role here. Perhaps reference to aims is just an unhelpful indirect way of talking about other factors that do the real explanatory work, such as facts about where the joints in reality are, or facts about how people should live, or facts about how our social/political institutions should be organized.

Scope of claims in conceptual engineering: Many of the claims involved in conceptual engineering are aimed at particular people, in particular circumstances. For example, someone might make a claim in conceptual ethics that "we should use concept A, instead of concept B". Who is the "we" here? It might be every rational agent, or every human being. But, much of the time, it will be a more limited group of people: for example, philosophers involved in the study of language, or people involved in a certain sort of social activism. In many cases (as noted above), the claims will often also be tied to *aims* those people have, or the purposes they have in those contexts. Thus, many claims in conceptual ethics take the form of something along the lines of: "a group of agents should use concept A, instead of B, in circumstances C, for purposes P".

Cluster 3

What are the plausible ameliorative strategies? And which ones are better than others? Once one has found a defect in a concept, there are many types of

[30] For further discussion of this topic, see Cappelen (2018).

[31] See, for example, Haslanger (2000) and Anderson (2001). For discussion, see Burgess and Plunkett (2013b).

ameliorative strategies that, at least prima facie, seem available. Which ameliorative strategy you endorse will be important for thinking about and engaging in conceptual activism. Here are some of those strategies:

(i) Improve/reform the concept and then use that one.
(ii) Replace uses of the concept with uses of an "explicated" concept which bears important similarities to the original one.
(iii) Replace uses of the concept with uses of a very different concept.

So far we have talked about improving/reforming concepts or uses of them, without being explicit about how this is reflected in language. It's difficult to engage in conceptual engineering without that having linguistic connections. How one thinks about this will in large part depend on how one thinks of the connections between thought and talk. Here are some views about the linguistic implementations of (i)–(iii):

(iv) Keep the lexical item and associate it with an improved/reformed concept.
(v) Introduce a new lexical item with associated improved concept(s).
(vi) Complete rejection: don't use that expression or the associated concept again.

This picture we just sketched presupposes that conceptual engineering operates on concepts and that these are then associated with (or expressed by) lexical items. As mentioned above, there are alternative views, according to which conceptual engineering operates directly on expressions and their intensions/extensions, so bypasses concepts entirely.[32] On this alternative view, here are some strategies that at least prima facie seem available:

(vii) Keep the lexical item and revise the intension/extension.
(viii) Introduce a new lexical item with a new intension/extension and then let this new lexical item replace uses of the old one.
(ix) Complete rejection: abandon the lexical item and its associated intension/extension.

As this last paragraph makes clear, how one thinks of the range of ameliorative strategies will depend in large part on what one takes the objects of conceptual engineering to be (e.g., concepts vs. conceptions vs. words), and one's views about their nature.

Obviously some of these strategies might work well in some contexts and not in others. It is far from obvious that there is a context-invariant way to assess ameliorative strategies.

What is the difference between improving concept C and improving one's beliefs about objects in the extension of C? Why choose one strategy over the other? An issue that keeps coming up in discussions of conceptual engineering is the difference between improving a concept, C, on the one hand, and improving (or revising) people's beliefs about objects in the extension of C, on the other. Why and when is one strategy superior to the other? How do they interact? The answers to these questions will depend heavily on the answer to the three previous questions. It will

[32] See Cappelen (2018).

depend on how one sees the connections between beliefs about C-objects and possession of the concept C—for some there's a constitutive connection, for others there isn't a connection at all.

Connections to ontology: Closely connected to the above point—to what extent can conceptual engineering change non-linguistic and non-conceptual aspects of the world? An obvious connection is this: if conceptual engineering succeeds in a particular case, it will change how people think, talk, and act on the (non-conceptual and non-linguistic) world. However, some conceptual engineers go further. For theorists who think conceptual engineering operates directly on extensions and intensions, the way to describe the effects of conceptual engineering (if the activism succeeds) is as a direct improvement of the non-linguistic and non-conceptual world.[33] On this view, we should not describe the effects of conceptual engineering as an amelioration of the concept of FREEDOM, but instead it is freedom itself that has been ameliorated. A final connection is this: for those who think conceptual engineering operates on concepts and also think that some concepts (or our use of them) can be constitutive of some element of non-conceptual reality (e.g, parts of social reality), there's an interesting connection: amelioration of an important social concept can change the nature of the relevant part of social reality (since the concept is partly constitutive of some element of social reality).[34]

Practical effects of conceptual engineering: How much practical import does conceptual engineering have? In what ways does it shape our thoughts, actions, and selves? And what about other parts of reality—for example, those that might depend on our conceptual practices in some way? In some cases, philosophers think the practical effects of conceptual engineering can be quite extensive and profound. But others are skeptical that engaging in conceptual engineering will have that big an impact at all (regardless of whether we are in control of that impact or not).

Cluster 4

Limits of revision and change of topic: How much revision is too much? When is a revision a complete change of topic? When would it be okay to change the topic, including perhaps completely abandoning the old topic in doing so? If you revise the concept COW to the concept FOX (or replace 'cow' with 'fox'), there's no sense in which you have improved on COW or 'cow'. You've just started talking about foxes. You've changed the topic. So the general issue then is how far we can go in amelioration without a complete change in topic? What degree of change is acceptable?[35]

Continuity of inquiry: Consider the debate over whether free will is compatible with determinism. This is a debate that has taken place over time involving many participants. Suppose that at some point during this debate that the concept FREE WILL is ameliorated. We now have, in some sense, a better "free will" concept. Does

[33] See Cappelen (2018).

[34] For example, this is true on the "dynamic nominalist" view that Hacking (2002) argues for concerning the relationship between naming practices and kinds of people (or, relatedly, social identities).

[35] This issue is central to Strawson's critique of Carnap on explication in Strawson (1963). For some of the more recent discussion on the topic, see Railton (2003); Haslanger (2012); Eklund (2017); and Cappelen (2018).

this constitute a discontinuity of inquiry: are we now engaged in an investigation of a new question and have we left the old question behind? Or is there some sense in which we can still say that we are discussing the same question as before, that is, whether free will is compatible with determinism? If there has been a discontinuity, what does that mean for the ability of speakers to meaningfully disagree with each other from different sides of the divide? What does it mean for our views about intellectual progress within that inquiry, and the ability of inquiry to build toward something that is "objectively" better in some sense?[36]

Conceptual engineering and verbal disputes: Consider again the debate over whether free will is compatible with determinism. Suppose as before that at some point the concept FREE WILL is ameliorated. Here is an issue that's closely related to the issue of continuity of inquiry: Isn't there now a significant risk of people engaging in verbal disputes? Those using the pre-ameliorated meaning for 'free will' say free will is compatible with determinism", and then those using the ameliorated concept say free will is not compatible with determinism". It will look like they are disagreeing, but since they mean different things, they might be engaged in a verbal dispute. Isn't the entire project of engaging in conceptual engineering at risk of generating an endless amount of verbal disputes?[37]

Cluster 5

Conceptual fixed points: Are some concepts or terms so basic that they cannot be engineered (or at least not in a way that is rational, or well-supported by reasons)? Are some concepts or terms so fundamental that we are stuck with them, meaning that evolution, revision, and amelioration are impossible? David Chalmers and Matti Eklund have defended the idea that there are conceptual fixed points (or "bedrock concepts", as Chalmers call them).[38] A central challenge for such views is to identify in a principled way the bedrock concepts and explain what makes them more fixed than those that can be engineered. A separate question is whether bedrock concepts should in some sense be normatively privileged, or whether the fact that they are a bedrock is just a descriptive fact about which concepts we happen to be stuck with in some sense.

The self-reflectiveness of conceptual engineering: Many philosophers engaged in conceptual engineering are, unsurprisingly, interested in the concepts used to articulate and describe conceptual engineering itself. For example, CONCEPT is itself an excellent candidate for conceptual engineering. So conceptual engineering can become (and perhaps should become) self-reflective. That self-reflection may change the nature of the activity.

Hypocrisy: How much of a problem is it if one uses (and not just mentions) concept C to make an argument that concept C should be replaced or revised? Is this an

[36] For some of our own take on these questions, see Plunkett (2015) and Cappelen (2018). Note that the issues here about continuity of inquiry have long been at the heart of debates in the history and philosophy of science, especially in the wake of Kuhn (1962/2012).

[37] For more on verbal disputes, see Chalmers (2011) and Jenkins (2014).

[38] See Chalmers (2011) and Eklund (2015).

objectionable form of hypocrisy? Or is it better described, at least in some cases, as a form of internal critique, or addressing one's opponents "on their own terms"? If so, then perhaps it is sometimes a virtue. The issues here will be particularly important for arguments in conceptual ethics involving foundational normative concepts, such as OUGHT and VALUE, that are hard (perhaps impossible) to avoid using in conceptual ethics. Perhaps it is impossible to avoid either some kind of "vindicatory circularity" or "hypocrisy" in such cases.[39] Importantly, these issues of hypocrisy and ineffability not only matter for those engaged in theoretical reflection in conceptual ethics. They also matter for those engaged in conceptual activism. For example: perhaps certain ways of advocating for conceptual change inevitably involve using the very concepts one aims to criticize. If so, then there are potentially issues involving not only hypocrisy here, but also issues of misleading or lying.[40]

Cluster 6

How often are we already engaged in conceptual engineering? And do we need to be aware of doing conceptual engineering in order to do it well? As we will discuss in a bit more detail later on in this Introduction, some philosophers think that much of existing philosophical inquiry involves conceptual engineering to some degree. Many hold that this engagement with conceptual engineering is going on implicitly, perhaps even without the philosophers themselves being aware that is what they are up to.[41] But many are skeptical of such claims. This raises a question: how much does one need to be aware of doing conceptual engineering to count as doing it? It also raises the question: how much does one need to be aware of doing it in order to do it *well*? Is explicit engagement with conceptual engineering always better than implicit engagement with it? For example, perhaps the best methodology for pursuing conceptual engineering (at least in certain kinds of inquiry) is not to focus on conceptual engineering as such, but rather just engage in ongoing inquiry into the intuitively relevant subject matter and then just let conceptual evolution happen naturally as part of the process. For example: for physicists to improve on and introduce new concepts along the way while trying to study physical reality, but without self-consciously ever thinking about part of their activity as an exercise in conceptual engineering.

How important is it to have a correct description of our representational devices before we do conceptual engineering? We've described the aim of conceptual engineering as that of assessing and ameliorating concepts and other representational devices. How important is it to have a correct descriptive account of those devices in order to do the engineering project well? Some analogies spring to mind: to think about how to improve a particular bridge, you need to know about that bridge—the ameliorative work can't be done in isolation from the descriptive work. How helpful are such analogies for understanding the connection between the descriptive and the

[39] See Alexis Burgess's chapter "Never Say 'Never say "Never"'?" (Chapter 6, this volume) for further discussion, as well as Eklund (2017).

[40] See Rachel Katharine Sterken's "Linguistic Intervention and Transformative Communicative Disruptions" (Chapter 20, this volume) for further discussion.

[41] For example, see Plunkett and Sundell (2013); Ludlow (2014); Plunkett (2015); and Thomasson (2016).

normative in the conceptual domain? How sharp is the distinction between the descriptive and the normative when doing conceptual engineering, or when thinking about concepts (or other representational devices) in general?

Conceptual engineering, the method of cases, and the role of intuitions in philosophy: Many philosophers think that the so-called "method of cases" is central to philosophical methodology and that intuitions about cases provide the most important kind of evidence for philosophical theories. One (controversial) way to spell out that view goes as follows: We have a concept, C, and our possession of that concept guides the intuitions about C-related thought experiments. So we can use intuitions about whether someone knows in, say, a Gettier case, as evidence of whether KNOWLEDGE applies in that case (because the concept somehow guides those intuitions). On this view, intuitions about cases reveal or illuminate core philosophical concepts and that is the reason why the method of cases is central to philosophy.[42] If, however, your goal is no longer to describe the concepts we have but to improve them—to think of how our concepts should be, then it is much less clear that asking questions of the form 'Is this a c?' about an imagined case can serve our purposes. There is a significant worry that this method, at best, reveals something about the concept C we as a matter of fact have, but our goal now is to think of what the concept *should* be.

Role of conceptual history/genealogy: Some people (notably Nietzsche in the passage quoted at the beginning of this Introduction) support claims in conceptual ethics by appeal to facts about the history/genealogy of concepts. Or, put in a way that will be more accurate on some theories of concepts: they appeal to historical facts about our (or other people's) engagement and use of those concepts. On the one hand, there is an obvious worry here that appeals to conceptual history/genealogy might fall prey to versions of the genetic fallacy. On the other hand, there are cases where such historical facts seem at least prima relevant to our assessment of our current conceptual practices—for example, if we had acquired our concepts by being brainwashed by an evil scientist, that fact should presumably play a role in our assessment and improvement of those concepts. What role should conceptual history/genealogy have in conceptual ethics?[43]

Let's take stock of where we are. In this section, we have presented six clusters of domain-independent issues in conceptual engineering. The issues interact and overlap in various interesting ways and should, we hope, make clear that theorizing about conceptual engineering is fertile ground for philosophical exploration. Many of these issues have not yet been systematically explored. There are as of yet few efforts to give unified theories of conceptual engineering. It should also be clear from the outline above that conceptual engineering interacts in various intriguing ways with topics in ethics, philosophy of language, philosophy of mind, philosophical methodology, metaphysics, philosophy of mind, and epistemology. It also interacts with issues in

[42] For a criticism of this way to thinking about the method of cases, see Cappelen (2012) and Deutsch (2015).

[43] For discussion of this issue, see Plunkett (2016).

linguistics, cognitive science, psychology, history, and sociology. Conceptual engineering has implications for those fields, but the theory of it will also draw on results from those fields. Towards the end of this Introduction we return to the issue of how one can see the position of conceptual engineering in philosophy overall.

4. Interaction between Specific Cases of Conceptual Engineering and General Theorizing

The theoretical foundations of conceptual engineering can be interesting in its own right, for much the same reasons any topic in philosophy can be. But many of those interested in conceptual engineering (and many conceptual activists in particular) are primarily motivated by an interest *not* in the general theoretical questions of the kind we just raised above for their own sake, but rather by a concern with specific concepts or words. For example: the concepts PERSON, FREEDOM, TRUTH, race and gender concepts, or concepts used for classifying mental illness in psychology. And in most cases they will also be interested in the lexical items used to express these concepts. Much of the contemporary discussion in conceptual engineering has been driven by concern with specific concepts or words.

In what follows we say a bit about some of the core issues we see involved in the *interaction* between the general and the specific here: how general theoretical issues (of the sort we canvassed in the last section) interact with more "applied" parts of conceptual engineering, focused on a specific concept or set of concepts.

Here is an analogy: A question that arises in *many* kinds of normative theorizing—including in ethics, political philosophy, and epistemology—is in what way (and to what extent) progress on specific cases (or more "applied" issues) is tethered to more general theory. For example: if we want to make progress on issues about debates in moral philosophy about climate change or abortion, how much (and when) should we appeal to a general normative ethical theory (e.g., act-utilitarianism)? What's the best way to proceed with this? Is there a general, informative theory about this methodological question? Or can we only answer it when we have a specific set of ethical questions on the table, in particular social-historical contexts? Parallel questions arise here in the context of thinking about conceptual ethics.

In a connected vein, we can also ask about the extent to which normative theorizing about a domain (whether about systematic/general issues, or more applied issues) is tied to *meta*-level theorizing about the domain. For example: we can ask about the extent to which work in normative ethics (e.g., about whether act-utilitarianism is correct) and applied ethics (e.g., about abortion) should be informed by work in metaethics. Even framing this question (as well as our questions in the previous paragraph) inevitably raises thorny issues about how (if at all) one should distinguish between these different topics, or projects.

Our goal is not to settle these debates here. Rather it is to flag them, and note their importance for work on conceptual engineering. We also want to make a few general remarks about some important points that should be kept in mind when thinking about these issues.

First, the basic issues on the table here aren't idiosyncratic to specifically normative domains. We can wonder—for example—about how much our theorizing about a particular topic in linguistics or biology should be informed by our more general theorizing in those domains, as well as our theorizing about our theorizing in those domains.

Second, it should be clear that in order to engage in conceptual engineering you do not need to have worked-out, explicit views on all these issues mentioned above. If that was required, conceptual engineering would never have happened. The questions here are about the interaction between more general theoretical reflection and more applied issues in conceptual engineering, rather than the issue of whether the former is a prerequisite for the latter.

Third, we have many good examples of someone being good at X without being a good theorist about X (or of having a theory about what being-good-at-X consists in, or possessing views about the best methodology for engaging in X, etc.). For example, many scientists make massively important contributions to science while having bad views in the philosophy of science. And many good tennis players don't have good theories of their own activity. Moreover, there might well be costs to theoretical reflection as well. Perhaps the time it takes to engage in that reflection could have been spent better doing something else. Or, more dramatically, perhaps theoretical reflection will make someone worse at what she does; think of a tennis player who can't serve as well after thinking too much about her serving technique. None of this means that theoretical reflection can't aid people in many cases. But it does suggest we should proceed with caution in assuming theoretical reflection here will be crucial to success.

Fourth, even if the thoughts in the last paragraph suggest some amount of modesty and caution here, we don't want to be overly pessimistic about the contributions of general theory (or meta-theory) about a domain to more "applied" or more specific issues. Such theoretical reflection can, and we think often does, help make contributions. This is especially so when our theoretical reflection is in relatively good epistemic standing, compared to the standing of our theorizing about the more "applied" issues.

Fifth, many find the following view attractive here: the interaction (in terms of evidential import, methodology, etc.) between more applied issues and more general issues in conceptual engineering will go in *both* directions. If some version of that idea is on the right track, then the following becomes important: there will be many interesting questions about *how* the general informs the specific and vice versa. This will, we predict, pattern (at least to a certain degree) in sync with other theoretical domains. General theories tend to take the form of models that abstract from the messiness of particular cases and that is in part what makes systematic theorizing possible. On the other hand, such models will then include idealizing assumptions that often (or sometimes) will make it hard to see how to apply it to particular cases.

Finally, it is worth highlighting that some concepts are tied up with the general theory in a particularly direct way: As we mentioned above, the efforts to conceptually engineer concepts such as CONCEPT, TRUTH, OUGHT, or CONCEPTUAL ENGINEERING will have a direct and immediate impact on the general theory. These are points where the general theory of conceptual engineering and the engineering of specific concepts will be deeply connected.

5. Role and Scope of Conceptual Engineering in Philosophy: Descriptive

One sense we sometimes get when talking to people about conceptual engineering and conceptual ethics are that they are *'hot' new topics*—a *trendy new field*. While we hope it is true that conceptual engineering and conceptual ethics are things that many philosophers will work on and think about, it would be misleading in the extreme if we gave the impression that these are topics/activities that haven't been important throughout the history of philosophy. Many philosophers, working in many different theoretical traditions, across many centuries, have thought of their work as involving some kind of conceptual engineering or conceptual ethics, and/or conceived of the work of other philosophers along such lines (even if they didn't use the terminology we use here).

For example, consider the founding work of analytic philosophy in the early twentieth century. A case can be made that much of this work centrally involved conceptual engineering. For example, Frege's *Begriffsschrift* is a paradigm of conceptual engineering: he aimed to improve language for certain purposes. As he puts it: "If the task of philosophy is to break the domination of words over the human mind . . . then my concept notation, being developed for these purposes, can be a useful instrument for philosophers".[44] Or take Wittgenstein. In *Tractatus Logico-Philosophicus*, he aimed to draw a line between what could be said and what could only be shown. You *shouldn't*, according to Wittgenstein, try to say what can only be shown. The aim of telling philosophers (and others) about the legitimate and illegitimate uses of language is a *normative* aim.

Next, consider Carnap. His work on explication and language choice are paradigms of conceptual engineering. He writes:

> The task of making more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather of replacing it by a newly constructed, more exact concept, belongs among the most important tasks of logical analysis and logical construction. We call this the task of explicating, or of giving an explication for, the earlier concept. (Carnap, 1947/1956: 8–9)

Carnap's interest in conceptual engineering expands beyond the idea of the explication of concepts. For example, his criticism of metaphysics as lacking a semantic foundation (as being nonsensical), and his proposal for an improved language (inspired by his verificationism), are also deeply bound up with issues in conceptual engineering. Importantly, much of the conceptual engineering in early twentieth-century analytic philosophy wasn't just concerned about purely epistemological or scientific goals. For example, consider Carnap's aim of modifying language to allow multiple people, from multiple places, to engage in collective, rational inquiry. Making that possible was in part a *political* aim, tied to a democratic, enlightenment view of politics that ran through Carnap's work.[45] This political side of things is also pronounced in Susan Stebbing's worries about how certain ways of using

[44] Frege (1879/1967: 7).

[45] For connected discussion, see Galison (1990).

key terminology in politics (e.g., 'democracy' or 'freedom') hindered clear thinking about social and political issues, which in turn made it difficult to effectively critique the rise of fascism. Stebbing saw analytic philosophy as helping provide tools to combat the relevant problematic sorts of thinking; both in diagnosing what was going wrong with it and in helping us make it better.[46]

As this brings out, Frege, Wittgenstein, Carnap, Stebbing, and other founders of analytic philosophy were extensively engaged in conceptual engineering. So rather than describe conceptual engineering as a 'hot' new topic in analytic philosophy, we could instead think of it as simply paying more attention to a key aspect of analytic philosophy that has been with us since it origins.

Moreover, the idea that key parts of philosophy involve conceptual engineering is hardly parochial to self-consciously "analytic" philosophy. Consider here the quote from Nietzsche we introduced at the start of this chapter, in which he claims that "what is needed above all is an absolute skepticism toward all inherited concepts" and that philosophers "must no longer accept concepts as a gift, nor merely purify and polish them, but first make and create them, present them and make them convincing".[47] Nietzsche is here advocating a radical skeptical stance with respect to inherited concepts, as well chastising other philosophers for failing to (at the very least) seriously engage that position.[48] In Nietzsche's view, it seems, philosophy should involve more conceptual engineering than it in fact has. This attitude is, in turn, reflected in his own work. For instance, a good part of *On the Genealogy of Morality* can be read as a critique of the distinctively *moral* concepts that have shaped much of modern life.[49] To put it in Nietzsche's own terms (from *Twilight of The Idols*), we have become "stuck in a cage, imprisoned among all sorts of terrible concepts", and part of his goal is to help us (or at least *some* of us) find a way out of that cage.[50]

In the passage from *The Will to Power*, Nietzsche (in characteristic fashion) positions his views as a radical break from much of the history of philosophy. But while Nietzsche might well be advocating for a more radical view of conceptual engineering than many have, there is a strong case that some amount of conceptual engineering, involving some amount of skepticism toward our inherited concepts, has played an important role in philosophy throughout its history. On this front, consider here what Strawson says in the introduction to *Individuals* about the difference between revisionary and descriptive metaphysics. He writes that "descriptive metaphysics is content to describe the actual structure of our thought about the world, revisionary metaphysics is concerned to produce a better structure".[51]

[46] See Stebbing (1939/1941) and Stebbing (1941/1948). Thanks to Bryan Pickel for helpful discussion of these parts of Stebbing's work.

[47] Nietzsche (1968: 220–1, section 409).

[48] Note that Nietzsche, later on in the passage we quoted, gives a nod to Plato for (at least possibly) seriously engaging the kind of radical skepticism about our current concepts that Nietzsche advocates. Plato's defense of our concepts, thinks Nietzsche, *perhaps* results from him taking the skeptical challenge seriously. He writes: "What is needed above all is an absolute skepticism toward all inherited concepts (of the kind that one philosopher *perhaps* possessed—Plato, of course—for he taught the reverse)" (Nietzsche 1968: 221, section 409).

[49] Nietzsche (1887/1994).

[50] Nietzsche (1889/1954: 502).

[51] Strawson (1959: 9).

Strawson describes the revisionist as insisting that metaphysics is "essentially an instrument of conceptual change, a means of furthering or registering new directions or styles of thought".[52] Importantly, Strawson gestures at a way of reading the history of philosophy as a division between revisionists and descriptivists. He says: "Perhaps no actual metaphysician has ever been, both in intention and effect, wholly the one thing or the other. But we can distinguish broadly: Descartes, Leibniz, Berkeley are revisionary, Aristotle and Kant descriptive. Hume, the ironist of philosophy, is more difficult to place. He appears now under one aspect, now under another".[53] If Strawson is correct, then conceptual engineering has been a key part of the history of philosophy stretching back centuries.

Moreover, the picture we get from Strawson above might well be understating the import of conceptual engineering to philosophy. If certain views in the philosophy of language are right, then philosophers might be *tacitly* engaging in conceptual engineering much more than they realize. For example, consider Peter Ludlow's recent work on what he calls the "dynamic lexicon". His idea, in rough outline, is that speakers regularly adjust and create new meanings for words on the fly in conversation, such that those meanings are "dynamic" and in flux. If Ludlow is correct, then much of conversation (including much of conversation *in philosophy*) involves navigating issues in conceptual engineering; including issues about what the best meaning of a word is for the context at hand, or going forward into further contexts.[54] A similar view of philosophy—in which it *in fact* involves extensive conceptual engineering (perhaps without the awareness of philosophers that it does so)—emerges in the work of other contemporary philosophers, including, for example, Amie Thomasson and Kevin Scharp.[55]

What we have just said is of course not even the beginning of a sketch of the role of conceptual engineering and conceptual ethics in the history of philosophy. There is rich terrain here to explore in many places, including, for example, its role in the Platonic dialogues, Hume, Kant, pragmatism, Heidegger, Foucault, Deleuze, Tarski, and ordinary language philosophy. We predict that there are rich new perspectives available for readings of the history of philosophy that more actively pay attention to the idea of conceptual engineering. We hope that scholars with better historical knowledge than us will write histories of philosophy from this perspective.

Of course, it might turn out that that philosophy has *in fact* involved extensive conceptual engineering, but that it is a deep mistake for it to continue in this vein. Or perhaps the reverse is true: perhaps philosophy has *not* in fact involved much conceptual engineering and should involve much more. This raises further normative issues. These issues are not about the descriptive question about the role that conceptual engineering (or conceptual ethics) has played in the history of philosophy (or currently plays in contemporary philosophy). Rather, they are about the question of what role conceptual engineering (or conceptual ethics) *should* play. We turn to this set of questions in the next section.

[52] Strawson (1959: 10). [53] Strawson (1959: 9). [54] Ludlow (2014).

[55] See Thomasson (2016) and Kevin Scharp's chapter "Philosophy as the Study of Defective Concepts", (Chapter 19, this volume). See also Plunkett and Sundell (2013) and Plunkett (2015) for sympathetic discussion of this possibility.

6. Role and Scope of Conceptual Engineering in Philosophy: Normative

So what role *should* conceptual engineering and conceptual ethics play in philosophy? For ease of exposition in what follows, let's take this question in terms of conceptual engineering. Here's a way to break up that normative question into two components:

1. How many parts of (or sub-fields of, or issues in) philosophy should conceptual engineering play a role in?
2. For each parts it should be involved in, how important should it be?

Focusing on 1 and 2 makes the answer to the question 'What role should conceptual engineering play in philosophy?' a matter of degree. Below we briefly sketch possible motivations for four kinds of answers:

1. All of All
2. All of Some
3. Some of All
4. Some of Some
5. Nothing

1. All of All of Philosophy

On this view, conceptual engineering should be seen as relevant to every issue in philosophy and it's the only thing that's relevant. So philosophy should consist entirely of conceptual engineering.

Kevin Scharp endorses All of All of philosophy. He argues that all philosophical concepts are defective, and that philosophy's task should be to discover those defects and then create replacement concepts. Once that is done, we ship the questions off to the sciences, and that's the end of philosophy.[56]

2. All of Some

This is the view you would hold if you think, for example, Scharp's view is correct for some parts of philosophy: in those parts, conceptual engineering is all there is to do. Then there are some other parts of philosophy where there's more to do than conceptual engineering.

3. Some of All of Philosophy

According to this view, conceptual engineering is relevant to all philosophical questions, but answering those questions requires more than doing conceptual engineering (so it is relevant to all of philosophy, but isn't all of philosophy).

One important central argument for Some of All is that we have no particular good reason to think that the concepts that we have inherited are *ideal* for philosophical theorizing. The default view should be that they could be improved. That thought is natural even for those who think we have good reason to preserve many of the

[56] See Scharp's "Philosophy as the Study of Defective Concepts" (Chapter 19, this volume).

distinctions found in natural language. For example, in "A Plea for Excuses", Austin says that

> ordinary language . . . embodies . . . the inherited experience and acumen of many generations of men. . . . If a distinction works well for practical purposes in ordinary life (no mean feat, for even ordinary life is full of hard cases), then there is sure to be something in it, it will not mark nothing.[57]

However, in a point that is often neglected by those who quote the above passage, Austin then goes on to note that "ordinary language is not the last word: in principle it can everywhere be supplemented and improved upon and superseded".[58] The challenge here is to recognize when ordinary language is good enough and when it can be improved upon. Philosophy might play a helpful role in that.

One view one could hold in support of Some of All is that the default assumption should be that the ordinary distinctions were not made for or developed to be ideal for philosophical theorizing. As a corollary, part of what philosophers should do is reflect critically on the usefulness of ordinary concepts for philosophical theorizing. A stronger view would hold that the concepts we currently have embody ideologies and power structures that can be repressive both on a political and personal level, so we should always critically examine inherited concepts. On those kinds of views, conceptual engineering isn't All of All of philosophy, but is a part of all of philosophy.

4. Some of Some of Philosophy

On this view conceptual engineering should be seen as relevant to some aspects of some philosophical questions. This is simply a more restricted version of Some of All. If you think the arguments we just sketched for Some of All don't apply to all philosophical concepts, but to some, this would be your favoured normative view of conceptual engineering.

One motivation for excluding conceptual engineering from some domains has been mentioned above: maybe some concepts are so basic that they are irreplaceable and immutable. Maybe TRUTH is like that. Maybe basic normative concepts, like OUGHT, are like that. If so, these are concepts you can use to engage in conceptual engineering, but are not themselves engineerable. If so, conceptual engineering is relevant in some, but not all philosophical domains.

5. Nothing

According to Nothing, conceptual engineering plays no part in philosophy *at all.* This is a view the editors of this volume bet against, but some contributors to this volume (e.g., Ball, Chapter 2, and Greenough, Chapter 11) flirt with or endorse it. Here are three possible motivations for Nothing:

[57] Austin (1956: 11).

[58] Austin (1956: 11).

(i) Conceptual engineering requires the existence of concepts, there are no concepts, so we can't do conceptual engineering. Since ought implies can, we shouldn't do conceptual engineering.[59]

(ii) Conceptual engineering involves fiddling with concepts, but fiddling with concepts is impossible.[60] Maybe that is because they are abstract entities (or entities of some other kind you can't fiddle with). Since ought implies can, we shouldn't do conceptual engineering.

(iii) It simply isn't within the proper domain of philosophy to fiddle with concepts. Philosophy is about reality: knowledge, freedom, meaning, belief, etc. Those phenomena are within the proper domain of philosophy, but the concepts are not.[61]

Most of the chapters in this volume are opposed to Nothing. This is natural since the aim, in part, is to provide readers with a broad range of frameworks for theorizing about conceptual engineering. In our view, it is only when many of those frameworks are on the table that proponents of Nothing will have a clear enough target.

7. Conceptual Engineering beyond Philosophy

Almost everything we have said about the potential significance of conceptual engineering (and conceptual ethics) apply beyond philosophy to inquiry more generally (as well as to speech and thought more generally).

Many of the arguments for the importance of conceptual engineering in philosophy are also arguments for its importance in other fields, such as biology, mathematics, physics, psychiatry, law, and politics. We use concepts and words in all areas of inquiry, and they can be better or worse relative to the goals and standards of a given part of inquiry. Thus, conceptual engineering matters for all domains of inquiry.

The point we just made applies beyond inquiry: many activities involve concepts and lexical items in various ways. The activity of cooking a dinner depends in part on the chefs' conceptual repertoire. So does hiking, going to war, and making friends. Some activities, like breathing and blinking one's eyes, might be independent of our concepts, but the range and significance of concept-involving activities is very broad. If that line of thought is correct, conceptual engineering is important not just for linguistic and cognitive activities, but also for many other core elements of human life.

These facts about the import of conceptual engineering beyond philosophy help underscore the potential significance of philosophical work done on conceptual engineering (and conceptual ethics). At the same time, it helps open up potentially fruitful avenues for future research. For once we see the import of conceptual engineering to other areas of inquiry—and to a wide range of activities that people engage in—we can look to see how conceptual engineering works in these other areas

[59] Note that this presupposes that conceptual engineering requires the existence of concepts in the relevant sense, which one of us (Cappelen) rejects. See Cappelen (2018).

[60] We here use 'fiddling' to mean roughly the following: do something with or to.

[61] There is, for example, a way of reading Williamson (2007) as advocating a version of this view. In fact, we think his view is more subtle than this, but we won't go into the details of it here.

as input to our philosophical theorizing about the topic. Such investigation might help us not only better understand how conceptual engineering in fact works, but also provide us with helpful material to think about in our theorizing about how it should work.

Conclusion

So far we have tried to remain fairly neutral in our presentation. In conclusion, it's time to put some of our cards on the table: we think conceptual engineering and conceptual ethics both *are* and *should be* central to philosophy. The role of conceptual engineering and conceptual ethics in philosophy and other areas of inquiry has been underexplored, often overlooked, and typically underappreciated. Philosophers have engaged in conceptual engineering and conceptual ethics in various sub-fields in various time periods, and there has been some scattered theorizing about them. But, especially in comparison to a range of descriptive questions about concepts and words—including issues about what they are, and how we use them—there has been relatively little sustained engagement with the normative and evaluative questions about concepts and words at the heart of conceptual engineering and conceptual ethics. We don't want to hang too much on the question of just how central conceptual engineering and conceptual ethics are to philosophy, but whatever degree of centrality one assigns to them, we hope this Introduction and the contributions in the volume helps highlight a broad range of interesting, important, and underexplored questions.

Acknowledgments

This chapter has benefited enormously from many discussions with other philosophers. We owe a particular debt to Alexis Burgess, our co-editor of this volume. We would also especially like to thank Derek Ball, Josh Dever, Matti Eklund, Øystein Linnebo, Tristram McPherson, Tim Sundell, Amie Thomasson, David Braddon-Mitchell, Peter Galison, Sally Haslanger, Peter Railton, Rachel Sterken, and Kevin Scharp. Note that the views in this Introduction are not the considered views of either Plunkett or of Cappelen. Their own considered views can be found in Plunkett (2015) and Cappelen (2018). The aim of this Introduction is to provide an easily accessible entry point for those not already invested in the topic.

References

Anderson, Elizabeth. 2001. Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics. *Economics and Philosophy* 17 (1):21–38.

Austin, John. 1956. A Plea for Excuses. *Proceedings of the Aristotelian Society* 57:1–30.

Ball, Derek. Chapter 2, this volume. Revisionary Analysis without Meaning Change (Or, Could Women Be Analytically Oppressed)? In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Blackburn, Simon. 2001. *Think: A Compelling Introduction to Philosophy*. Oxford: Oxford University Press.

Boghossian, P. 1989. The Rule-Following Considerations. *Mind*, 98, 507–49.

Brandom, Robert B. 2001. Modality, Normativity, and Intentionality. *Philosophy and Phenomenological Research* 63 (3):611–23.

Burgess, Alexis. Never Say "Never say 'Never'"? Chapter 6, this volume. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Burgess, Alexis, and Plunkett, David. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, Alexis, and Plunkett, David. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Burgess, Alexis, and Sherman, Brett. 2014. *Metasemantics: New Essays on the Foundations of Meaning*. Oxford: Oxford University Press.

Cappelen, Herman. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Carnap, Rudolf. 1947/1956. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

Chalmers, David J. 2011. Verbal Disputes. *Philosophical Review* 120 (4):515–66.

Deutsch, Max. 2015. *The Myth of the Intuitive*. Cambridge, MA: MIT Press.

Eklund, Matti. 2002. Inconsistent Languages. *Philosophy and Phenomenological Research* 64 (2):251–75.

Eklund, Matti. 2015. Intuitions, Conceptual Engineering, and Conceptual Fixed Points. *The Palgrave Handbook of Philosophical Methods*, ed. C. Daly. London: Palgrave.

Eklund, Matti. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.

Frege, Gottlob. 1879/1967. *Begriffsschrift: Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Trans. S. B.-M. in Jean van Heijenoort (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA: Harvard University Press.

Galison, Peter. 1990. Aufbau/Bauhaus: Logical Positivism and Architectural Modernism. *Critical Inquiry* 16 (4):709–52.

Gibbard, Allan. 2012. *Meaning and Normativity*. Oxford: Oxford University Press.

Hacking, Ian. 2002. Making Up People. *Historical Ontology*. Cambridge, MA: Harvard University Press.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Nous* 34 (1):31–55.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Hattiangadi, Anandi. 2017. The Normativity of Meaning. In B. Hale, C. Wright and A. Miller (eds.), *A Companion to the Philosophy of Language*. Oxford: Wiley Blackwell.

Jenkins, C. S. I. 2014. Merely Verbal Disputes. *Erkenntnis* 79 (1):11–30.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Kuhn, Thomas. 1962/2012. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lau, Joe, and Deutsch, Max. 2016. Externalism about Mental Content. *The Stanford Encyclopedia of Philosophy* (winter edn), ed. Edward N. Zalta. <https://plato.stanford.edu/archives/win2016/entries/content-externalism/>.

Ludlow, Peter. 2014. *Living Words: Meaning Undetermination and the Dynamic Lexicon*. Oxford: Oxford University Press.

Margolis, Eric, and Laurence, Stephen. 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Nietzsche, Friedrich Wilhelm. 1887/1994. *On the Genealogy of Morality*. New York: Cambridge University Press.

Nietzsche, Friedrich Wilhelm. 1889/1954. Twilight of the Idols. In W. Kaufmann (ed.), *The Portable Nietzsche*. New York: Viking.

Nietzsche, Friedrich Wilhelm. 1901/1968. *The Will to Power*. Trans. W. Kaufmann. New York: Random House.

Plunkett, David. 2015. Which Concepts Should We Use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 58 (7–8):828–74.

Plunkett, David. 2016. Conceptual History, Conceptual Ethics, and the Aims of Inquiry: A Framework for Thinking about the Relevance of the History/Genealogy of Concepts to Normative Inquiry. *Ergo* 3 (2):27–64.

Plunkett, David, and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Quine, Willard V.O. 1951. Two Dogmas of Empiricism. *Philosophical Review* 60 (1):20–43.

Railton, Peter. 1986a. Facts and Values. *Philosophical Topics* 14:5–31.

Railton, Peter. 1986b. Moral Realism. *The Philosophical Review* 95:163–207.

Railton, Peter. 2003. *Facts, Values, and Norms: Essays toward a Morality of Consequence*. New York: Cambridge University Press.

Scharp, Kevin. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Scharp, Kevin. Chapter 19, this volume. Philosophy as the Study of Defective Concepts. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Schroeter, Laura, and Schroeter, Francois. 2014. Normative Concepts: A Connectedness Model. *Philosophers' Imprint* 14 (25):1–26.

Sider, Theodore. 2012. *Writing the Book of the World*. Oxford: Oxford University Press.

Stebbing, L. Susan. 1939/1941. *Thinking to Some Purpose*. New York: Allen Lane: Penguin Books.

Stebbing, L. Susan. 1941/1948. *Ideals and Illusions*. London: Watts and Co.

Sterken, Rachel Katharine. Chapter 20, this volume. Linguistic Intervention and Transformative Communicative Disruptions. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Strawson, P. F. 1959. *Individuals: An Essay in Descriptive Metaphysics*. London: Routledge.

Strawson, P. F. 1963. Carnap's Views on Conceptual Systems versus Natural Languages in Analytic Philosophy. In P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*. London: Open Court.

Thomasson, Amie L. 2016. Metaphysical Disputes and Metalinguistic Negotiation. *Analytic Philosophy* 57 (4) :1–28.

Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Abstracts of Chapters

1 Introduction

In this Introduction, we aim to introduce the reader to the basic topic of this book. As part of this, we explain why we are using two different expressions ('conceptual engineering' and 'conceptual ethics') to describe the topics in the book. We then turn to some of the central foundational issues that arise for conceptual engineering and conceptual ethics, and finally we outline various views one might have about their role in philosophy and inquiry more generally.

2 Revisionary Analysis without Meaning Change (Or, Could Women Be Analytically Oppressed?)

This chapter develops a conception of philosophical analysis which makes sense of the idea that a correct analysis can be revisionary (in that it departs from ordinary or expert belief and linguistic usage). The view is superior to the alternatives defended by most proponents of 'conceptual ethics' and 'conceptual engineering' (according to which revisionary theorizing involves replacing words or concepts) because it better explains the arguments we advance when we engage with proposed revisionary analyses. A key idea is that analytic claims can emerge in the course of debate without change of meaning, so that our acceptance (perhaps late in the debate) of some analyticity can fix the meaning of a word as we used it all along. The discussion focuses on Haslanger's revisionary analysis of gender.

3 Minimal Substantivity

This chapter defends a notion of "minimal substantivity" for ontological disputes, focusing on persistence and composition as case-studies. Deflationists argue that these disputes are defective: they are merely verbal, or epistemically underdetermined. Anti-Deflationists emphasize the substantivity of these controversies, either in Quinean terms, or in terms of fundamentality. This chapter aims at striking a mid-way between these approaches, thus outlining a Minimal Anti-Deflationism. First, it is shown that even if these ontological disputes are merely verbal, they can be recast as non-verbal metalinguistic disputes. These exchanges have in turn an ontological substantivity if we understand them as disputes about which ontological commitments we should undertake. This substantivity is however minimal, because it need

not be articulated either naturalistically or in terms of fundamentality. In closing, an illustration is provided of the fruitfulness of the minimal substantivity notion for cases of conceptual engineering.

4 Reactive Concepts

This chapter recommends that we consider a kind of concept which bears a relation like the one traditional accounts of concepts bear to beliefs, but instead bears it to states individuated not only by their causal inputs, but also by their direct causal outputs. They will be called reactive representations, RRs for short. They are partially representational states which are reactive inasmuch as they bypass interaction with distinct desires to directly motivate behaviour. Associated with these representations are abilities that will be called Reactive Concepts. The chapter argues that taxonomizing mental states this way casts light on the nature of a range of phenomena, including hate speech, crypto-evaluative terms, and phenomenal concepts.

5 Strategic Conceptual Engineering for Epistemic and Social Aims

This chapter advocates strategic conceptual engineering, that is, the employment of a (possibly novel) concept for specific epistemic or social aims, concomitant with the openness to use a different concept for other contexts. We illustrate this approach by sketching three distinct concepts of gender and arguing that all of them are needed, as they answer to different social aims. The first concept serves the aim of identifying and explaining gender-based discrimination. It is similar to Haslanger's account, except that rather than offering a definition of 'woman' we focus on 'gender' as one among several axes of discrimination. The second concept of gender is to assign legal rights and social recognitions, and is to be trans-inclusive. We argue that this cannot be achieved by previously suggested concepts (including Jenkins's) that include substantial gender-related psychological features or awareness of social expectations. The third concept of gender serves the aim of personal empowerment through gender identity. This chapter points to contexts where a concept's role in explanation and moral reasoning can be more important than determining the extensions of concepts.

6 Never Say 'Never Say "Never"'?

Is there anything wrong with using a concept C in the course of arguing against the use of C? You might think not, so long as the argumentation is framed as a kind of reductio; or climbing a ladder to be kicked away. On the other hand, you might find such "hypocritical" arguments self-undermining: if they're sound, then we shouldn't accept any of their premises that make use of C. Understanding the status of hypocrisy is an urgent question for conceptual ethics (and therefore engineering),

insofar as concepts of philosophical interest tend to be central to our conceptual schemes, in the sense that they tend to turn up in the analyses of lots of other concepts. Developing these points (of view), the present chapter exposes a worrisome feature of the anti-hypocrisy argument just sketched: it seems to be indeterminate whether or not the argument is itself hypocritical.

7 Conceptual Engineering

This chapter develops and defends the Master Argument for Conceptual Engineering: (1) If W is a word that has a meaning M, then there are many similar meanings, M1,M2,...,Mn, W could have. (2) We have no good reason to think that the meaning that W ended up with is the best meaning W could have: there will typically be indefinitely many alternative meanings that would be better meanings for W. (3) When we speak, think, and theorize it's important to make sure our words have as good meanings as possible. (4) As a corollary: when doing philosophy, we should try to find good meanings for core philosophical terms and they will typically not be the meanings those words as a matter of fact have. (5) So no matter what topic a philosopher is concerned with, she should assess and ameliorate the meanings of central terms. I respond to seven objections to this argument.

8 Preliminary Scouting Reports from the Outer Limits of Conceptual Engineering

If we distinguish between the conceptual engineering task of determining what concepts can be built and the conceptual ethics task of determining what concepts ought to be used once built, we make clear the possibility of a conceptual maximalist position that avoids conceptual ethics by holding that the norms of theorizing require a big theory including all truths expressible using any concepts. But the conceptual maximalist then assumes the burden of saying what the range of possible concepts (equivalently, possible languages) is. Despite the attractions of conceptual maximalism, we lack even the beginnings of an answer to the question of what concepts there could be.

9 Descriptive vs. Ameliorative Projects: The Role of Normative Considerations

Feminist philosophers have famously argued that contextual factors play a role in the justification and assessment of scientific theories. A similar question might arise regarding inquiries in metaphysics. Contemporary metaphysicians often assume that contextual factors do not belong in metaphysics. The main aim of this chapter is to argue that normative considerations such as moral and political considerations are relevant in metaphysics. In particular, this chapter explores the nature of descriptive projects seeking to reveal the operative concept associated with a term (as opposed to ameliorative projects seeking to reveal the target concept that we should associate with the term), and argues that moral and political considerations are relevant not only with respect to ameliorative projects but also with respect to descriptive projects.

10 Variance Theses in Ontology and Metaethics

The chapter illustrates conceptual engineering by bringing up a number of issues in metaontology and metaethics. A prominent debate in metaontology relates to whether some existence concept is metaphysically privileged. On the one hand, ontological realists say yes, and, on the other hand, friends of quantifier variance say no. The chapter brings up the corresponding question in metaethics by asking, is some rightness concept normatively privileged? It investigates this question, and compares the metaethics case and the metaontology case. One aim is to arrive at conclusions regarding possible limits to the project of conceptual engineering.

11 Neutralism and Conceptual Engineering

Conceptual engineers often invoke a distinction between happy-face and unhappy-face solutions to alethic paradoxes. Happy-face solutions are thoroughly specific: they isolate a single, basic principle ("the culprit"). Unhappy-face solutions, meanwhile, are thoroughly non-specific: they merely establish the collective guilt of a group of principles which together produce the paradox. According to this taxonomy, conceptual engineering can only take place via unhappy-face solutions. In this chapter, I: (1) give an expanded taxonomy which allows for both Happy-Face and two forms of Unhappy-Face Conceptual Engineering and show that (2) happy-face treatments represent a limit case. (3) Unhappy-face treatments also represent a kind of limit case. (4) Between these limit cases are treatments which are neither maximally specific nor maximally unspecific but nonetheless specific enough to treat a paradox. (5) Such treatments become thoroughly neutralist when they reject some principle at work in a paradox from a theory-neutral perspective. The upshot is Neutralism—the view that philosophical progress can take place when (and sometimes only when) a thoroughly neutral, non-specific theory is adopted.

12 Going On, Not in the Same Way

This chapter considers, within an externalist semantics, several ways we might understand the project of improving our concepts to promote greater justice. The tools that culture provides us with—such as language, concepts, and inferential patterns—provide frames for coordination and shape our interaction. There are multiple ways these tools can fail us, for example by the limited structure of options they make intelligible. However, we can sometimes reconfigure the resources so that our practical orientations are more responsive to what is good and coordinate in ways that are just. Such reconfiguration often happens in law; it also occurs in social movements, counter-publics, subaltern communities, and in fascist propaganda. Contestation over meaning is not "mere semantics" for—together with political and material change—it can shape our agency and our lives together.

13 The Theory–Theory Approach to Ethics

One way to approach the theory of reference for proper names is by asking what proper names are good for in the sense of the valuable purposes they serve. Suppose we approach ethical terms and concepts in the same spirit, asking questions like: What purposes do they serve? How could we do something similar but do it better? This chapter explores the implications of this way of thinking about ethical terms and concepts, and explains why a theory–theory or moral functionalist account of them is so attractive when we approach matters from this perspective. The discussion is set inside an avowedly cognitivist, naturalist framework, and touches on the implications of this framework for how to adjudicate debates between rival views in ethics, and the relevance of evolutionary considerations.

14 Conceptual Ethics and the Methodology of Normative Inquiry

This chapter explores two central questions in the conceptual ethics of normative inquiry. The first is whether to orient one's normative inquiry around folk normative concepts (like KNOWLEDGE or IMMORAL) or around theoretical normative concepts (like ADEQUATE EPISTEMIC JUSTIFICATION or PRO TANTO PRACTICAL REASON). The second is whether to orient one's normative inquiry around concepts whose normative authority is especially accessible to us (such as OUGHT ALL THINGS CONSIDERED), or around concepts whose extension is especially accessible to us (such as BETRAYAL). The chapter aims to make vivid and plausible a range of possible answers to these questions, and important forms of argument that can be used to favor certain answers over others.

15 Conceptual Evaluation: Epistemic

On a view implicitly endorsed by many, a concept is epistemically better than another if and because it does a better job at 'carving at the joints', or if the property corresponding to it is 'more natural' than the one corresponding to another. This chapter offers an argument against this seemingly plausible thought, starting from three key observations about the way we use and evaluate concepts from an epistemic perspective: that we look for concepts that play a role in explanations of things that cry out for explanation; that we evaluate not only 'empirical' concepts, but also mathematical and perhaps moral concepts from an epistemic perspective; and that there is much more complexity to the concept/property relation than the natural thought seems to presuppose. These observations, it is argued, rule out giving a theory of conceptual evaluation that is a corollary of a metaphysical ranking of the relevant properties.

16 Analyzing Concepts and Allocating Referents

The analysis of concepts is an important part of the philosophical exercise but has to be complemented by the allocation of a referent to any world-tracking concept: this will give an account of what constitutes the phenomenon tracked. But analysis generally leaves room for negotiation about the referent to be allocated, for two reasons. First, because it may underdetermine the specification of conditions that anything must satisfy if it is to count as a referent. And, second, because the specification chosen may itself underdetermine the property that constitutes the referent. The license that philosophy consequently enjoys has to be exercised on the basis of analytically independent criteria linked with the theory within which any realistic concept has to take its place. The lesson is illustrated schematically, it comments on concepts like those of causation, freewill, and value and in a case study of the concept of freedom. It connects with the line argued by a range of recent, more revisionary approaches to philosophy.

17 The A-project and the B-project

Ameliorative philosophical analysis elucidates a concept by looking at the purposes it has for its users. It evaluates them, and gives an analysis on the basis of the purposes the analyst thinks ought to control the concept's use. As such it can look wildly revisionary, an attempt not to tell us what we mean but to change the subject. This chapter sketches a view of meaning on which ameliorative analysis can be understood as not at all subject changing: a word's meaning is the evolving collection of presuppositions its users make, and expect to be recognized as made, when the word is used. Focusing on Sally Haslanger's ameliorative accounts of race and gender, this chapter argues that ameliorative analysis is best understood not as a project that is successful only if it results in the concept's reference reflecting the ameliorative analysis. Conceptual amelioration and engineering are attempts to foster a kind of evolution within a population, a change in the presuppositions that (it is common knowledge) accompany the concept's use. This can be achieved and have worthwhile effects even while no shift in reference occurs.

18 Talk and Thought

This chapter provides an externalist account of talk and thought that clearly distinguishes the two. It is argued that linguistic meanings and concepts track different phenomena and have different explanatory roles. The distinction, understood along the lines proposed, brings theoretical gains in a cluster of related areas. It provides an account of meaning change which accommodates the phenomenon of contested meanings and the possibility of substantive disagreement across theoretical divides, and it explains the nature and value of conceptual engineering in a way that addresses recent prominent concerns.

19 Philosophy as the Study of Defective Concepts

From familiar concepts like tall and table to exotic ones like gravity and genocide, they guide our lives and are the basis for how we represent the world. However, there is good reason to think that many of our most cherished concepts, like truth, freedom, knowledge, and rationality, are defective in the sense that the rules for using them are inconsistent. This defect leads those who possess these concepts into paradoxes and absurdities. Indeed, I argue that many of the central problems of contemporary philosophy should be thought of as having their source in philosophical concepts that are defective in this way. If that is right, then we should take a more active role in crafting and sculpting our conceptual repertoire. We need to explore various ways of replacing these defective concepts with ones that will still do the work we need them to do without leading us into contradictions.

20 Linguistic Intervention and Transformative Communicative Disruptions

What words we use, and what meanings they have, is important. We shouldn't use slurs; we should use 'rape' to include spousal rape (for centuries we didn't); we should have a word which picks out the sexual harassment suffered by people in the workplace and elsewhere (for centuries we didn't). Sometimes we need to change the word-meaning pairs in circulation, either by getting rid of the pair completely (slurs), changing the meaning (as we did with 'rape'), or adding brand new word-meaning pairs (as with 'sexual harassment'). A problem, though, is how to do this. One might worry that any attempt to change language in this way will lead to widespread miscommunication and confusion. I argue that this is indeed so, but that's a feature, not a bug, of attempting to change word-meaning pairs. The miscommunications and confusion such changes cause can lead us, via a process I call transformative communicative disruption, to reflect on our language and its use, and this can further, rather than hinder, our goal of improving language.

21 A Pragmatic Method for Normative Conceptual Work

How ought we to do work in conceptual ethics? Some have thought that conceptual choice should itself be guided by (heavyweight) metaphysics—for we should be sure that our concepts pick out things that exist or should aim to choose concepts that

really ‘carve the world at its joints’. An alternative is to take a pragmatic approach to conceptual ethics. But pragmatic approaches are often criticized as unable to account for intuitions that some conceptual choices are objectively better than others, and intuitions that the world is structured. As a result, the fear is that a pragmatic approach leaves conceptual choices arbitrary and insusceptible to critique. This chapter confronts such worries and develops a pragmatic method for conceptual ethics that clearly avoids these problems. As a result, we need not rely on heavyweight metaphysics and become entangled in its epistemological mysteries to do conceptual ethics.

2

Revisionary Analysis without Meaning Change (Or, Could Women Be Analytically Oppressed?)

Derek Ball

There are a number of conceptions of philosophical analysis, but a common thread in most of these conceptions is that the aim of analysis is descriptive: we take our words and concepts as they are, and try to draw some philosophical conclusion from them. Frank Jackson captures this picture of philosophy when he writes,

> What we are seeking to address is whether free action *according to our ordinary conception*, or something suitably close to our ordinary conception, exists and is compatible with determinism, and whether intentional states *according to our ordinary conception*, or something suitably close to it, will survive what cognitive science reveals about the operations of our brains.
>
> (1998: 31; original emphasis)

Analysis is important, on Jackson's view, precisely because it ensures that we do not stray too far from ordinary thought and linguistic usage. The upshot is much in the spirit of Wittgenstein: "Philosophy may in no way interfere with the actual use of language; it can in the end only describe it. [. . .] It leaves everything as it is" (1953: section 124).[1]

Other philosophers—across a range of subdisciplines—have advocated a critical and constructive philosophical project: we should not simply take our words and concepts as they come, but should aim to improve them if we can. Among the most provocative of such attempts is Sally Haslanger's social constructivist analysis of race and gender terms. Haslanger proposes the following analysis of 'woman':

[1] Jackson himself would not endorse Wittgenstein's claim in full generality; he thinks that it is sometimes appropriate to replace one concept with another that can better do the relevant "theoretical job" (1998: 44). The point nonetheless stands that on Jackson's view, analysis is important precisely because it ensures continuity with ordinary thought and linguistic usage.

Derek Ball, *Revisionary Analysis without Meaning Change (Or, Could Women Be Analytically Oppressed?)* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Derek Ball.
DOI: 10.1093/oso/9780198801856.003.0002

> S is a woman iff
>
> (i) S is regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction;
>
> (ii) that S has these features marks S within the dominant ideology of S's society as someone who ought to occupy certain kinds of social position that are in fact subordinate (and so motivates and justifies S's occupying such a position); and
>
> (iii) the fact that S satisfies (I) and (ii) plays a role in S's systematic subordination, that is, *along some dimension*, S's social position is oppressive, and S's satisfying (i) and (ii) plays a role in that dimension of subordination. (2012a: 234)

Call this claim (W) (and set aside the many interesting questions involved in working out the details). Haslanger concedes that (W), construed as an analysis, clashes with a number of aspects of ordinary and expert usage and belief (e.g., given Haslanger's view, feminists should attempt to eliminate women (Haslanger 2012a: 239)). But she nonetheless maintains that this is the analysis that we should accept—at least if we share her broadly feminist goals, such as identifying and explaining "persistent inequalities between females and males" and uncovering how "social forces [...] work to perpetuate such inequalities" (2012a: 226–7), with an eye toward ultimately undermining these forces for the sake of social justice.

Analyses like (W) are *revisionary* in that accepting such an analysis involves a departure from ordinary or expert usage and belief. But the idea that something could be both *revisionary* and an *analysis* seems to stand in considerable tension with the idea that analysis must be descriptive. This tension has led opponents of revisionary analyses to the accusation that such analyses are *changing the subject*; roughly, the idea is that if our analysis is not describing current usage of, say, 'woman', it must be describing an alternative usage of a homophonic word, a usage on which 'woman' would take on a novel meaning. To take an example from popular politics, proponents of same-sex marriage are often accused of trying to redefine 'marriage', where the subtext of the accusation is that opponents of same-sex marriage are right about *marriage*, and proponents (presumably knowing this) want to stop talking about marriage and start using the word 'marriage' to talk about something else. A similar worry motivates Jackson's emphasis on continuity with ordinary thought and usage; as Jackson sees it, the alternative to descriptive analysis is stipulated change of meaning, but such stipulation is of little philosophical interest:

> If I say that what I mean—never mind what others mean—by a free action is one such that the agent would have done otherwise if he or she had chosen to, then the existence of free actions so conceived will be secured, and so will the compatibility of free action with determinism [...] I have turned interesting philosophical debates into easy exercises in deductions from stipulative definitions together with accepted facts. (1998: 31)

Philosophical discussion of revisionary analysis, from Carnap (1956: 7) on, generally takes for granted the idea that revisionary analyses change the subject in this sense. The controversy is only over whether (and when and how) it makes sense to advocate for particular concepts, for particular changes of subject. Thus proponents of particular revisionary analyses often see themselves as defending the replacement of one concept with another (e.g., Scharp 2013), and proponents of "conceptual engineering" and "conceptual ethics" present themselves as defending the idea that we can study questions like "Should we use concept C (over alternative A)?" (Burgess and Plunkett 2013a,b), and that the answers can inform our understanding of a range of popular and philosophical debates (Plunkett and Sundell 2013; Thomasson 2016).

In my view, the idea that revisionary analysis requires the *replacement* of one concept with another, or *changing* the meaning of our words, or *developing new concepts*, is entirely misguided. There is a perfectly natural sense in which one can advance a revisionary analysis like (W) as an analysis of 'woman' as we have always used it, the word and the concept that we have been employing all along—without changing the subject or engineering a new concept or anything of the sort.[2] There is no conflict between the idea that such analyses are revisionary and the idea that they are descriptive: they are revisionary, in that they depart from present usage and beliefs, but are still descriptive in that they are making a claim about what we mean now and have meant all along. Further, there is a perfectly natural sense in which revisionary analyses might turn out to be correct—not that it might turn out to be correct to adopt a new term or concept, or that changing the subject might be the right thing to do, but that (W) might turn out to be the correct analysis of 'woman' as we have used it all along. (So it might turn out that women are subordinated not just as a matter of fact, but analytically, and have always been so—no matter how widespread belief to the contrary may be or have been.)

Of course, none of this is to deny the obvious fact that we sometimes introduce new terminology (or technical uses of extant terminology), and correspondingly new concepts, in the course of theorizing. The point is that this is neither the only nor the most plausible account of what proponents of same-sex marriage, or social constructivist analyses of gender, or most other related debates are doing. A better picture—better because (as I will show) it better explains our epistemic engagement with revisionary analyses—is that conceptual analysis can go beyond (and even overturn) extant belief and linguistic usage without changing of meaning, and without introducing new concepts. Our theoretical activity shapes what we mean, but it does so not by making us mean something new, but by shaping what we meant all along. So we are rarely if ever faced with two concepts and forced to choose which to use. We *are* often faced with new claims and new theories; we accept some and reject others, on good or bad grounds, and this plays a role in determining our meanings and our concepts. But this is just part of what it is to be a thinker. There are

[2] Though Haslanger (2012a) suggests that accepting (W) involves change of meaning, in more recent work Haslanger (2012b,d) agrees that revisionary theorizing need not be seen as meaning changing. In support of this claim, she appeals to the metasemantic views of Schroeter, Schroeter, and Bigelow (Bigelow and Schroeter 2009; Schroeter and Schroeter 2009), according to which (as Haslanger puts it), "interpretation of our own past linguistic practice with a term and the practice of those around us, together with empirical investigation, enables us to make judgments about how the term applies; the term refers to what a fully informed and rational judge in such circumstances would take herself to refer to. To the extent that we are informed and rational, we can know the correct application of the term" (2012b: 437). I am not unsympathetic to the spirit of this, and though I have significant doubts about the details, this is not the place to air them. In the context of this chapter, I would insist only that there is no reason to think that past linguistic practice (as opposed to the entire course of a linguistic practice, including its future) plays a distinctive role.

no interesting questions of conceptual engineering or conceptual ethics, over and above questions about when we accept (or should accept) certain claims or theories, of the familiar sort long studied by epistemologists and philosophers of science, as well as psychologists and philosophers of mind.

This chapter aims to clarify and defend these claims by developing an account of revisionary analysis that makes sense of our engagement with revisionary analyses, and in particular, makes sense of the way we argue for and against them—our epistemic engagement with them. I begin by examining these arguments, and showing that the view that revisionary analysis involves meaning change or concept replacement cannot make sense of them. In section 2, I turn to Haslanger's proposal, and suggest that her view is best formulated in terms of competing analyses (rather than in terms of competing concepts). Section 3 develops the relevant notion of analysis—a notion on which analytic claims fix the meanings of words. Sections 4 and 5 show how this notion can be developed to explain our engagement with revisionary analyses. A key idea is that analytic claims can emerge in the course of debate without change of meaning, so that our acceptance (perhaps late in the course of a discussion) of some analyticity can fix the meaning of a word as we used it all along.

1. Changing the Subject

Consider the view that accepting a sentence like 'Two men can marry each other' amounts to changing the meaning of the word 'marry'.[3] As Jackson might put it, the claim is that the proponent of same-sex "marriage" says, "this is what I will mean by 'marriage'—never mind what others mean"; or perhaps better, "this is what I will mean by 'marriage', and others should do the same—never mind what we have meant in the past". If this is right, the proponent of same-sex "marriage" might go on to speak truly by saying things like "Two men can marry each other". But if someone thinks that a man can only marry a woman, she should not regard the proponent of same-sex "marriage" as contradicting her. The proponent's utterance of 'Two men can marry each other' does not express the proposition that two men can marry each other; she has not changed her belief in the falsehood of this proposition, but only her beliefs about what sentences are best used to express it. Her choice to use the term 'marriage' in an idiosyncratic way does not change the facts about marriage, any more than choosing to call a tail a 'leg' makes it the case that horses have five legs.

[3] The case of same-sex marriage is complicated by the fact that it was until recently a *legal* impossibility everywhere for two men to marry each other. But opponents of same-sex marriage—especially the kind of opponent of same-sex marriage who claims that proponents are changing meaning—tend to think that same-sex marriage is impossible in a deeper sense (conceptually impossible), so that laws permitting same-sex "marriage" must be regulating some state other than marriage. The claim 'Two men can marry each other' is to be read as advocating the conceptual possibility of same-sex marriage; so (in the dispute I have in mind) the proponent of same-sex marriage is claiming that it would make sense to have laws that allowed same-sex marriage (and not merely same-sex "marriage"). This claim could be true even if same-sex marriage were legally impossible (though of course the proponent would presumably also advocate changing such a law).

(And of course this would be true even if the proponent of this way of speaking is able to convince everyone to adopt her usage.)

If we are talking about legs, and I try to stipulate that 'leg' is to be applied to tails, I am trying to change the subject.[4] If my stipulation succeeds and we continue to speak of "legs", then the subject has (at least partially) changed; we are no longer speaking (just) of legs, but of tails as well. (If words are individuated by their meanings, we are no longer using the same word; I have introduced, and gone on to use, a mere homonym of the ordinary English word 'leg'.) When we say "Horses have five legs", we should not take ourselves to be disagreeing with those who, unfamiliar with my stipulation, say, "Horses have four legs"—except, of course, about a matter of linguistic usage.

Let's say that an analysis that involves a change of meaning in this way *changes the subject*. Self-avowed proponents of conceptual engineering and conceptual ethics, and some proponents of revisionary analyses, tend to regard revisionary analyses as changing the subject. Call this claim the *subject-change view* of revisionary analysis.

Proponents of the subject-change view have made concerted attempts to explain some aspects of disputes about revisionary analyses. For example, they point to other cases in which we have a sense of genuine disagreement even where it is clear that the parties are not asserting contradictory contents (e.g., "I like Steve"; "No, you're wrong: I hate him"). And they claim that in many circumstances, a primary purpose of an assertion may be to communicate metalinguistic information about how words are to be used (Sundell 2011; Plunkett and Sundell 2013).

These observations are fair enough as far as they go. But there is a great deal more to explain. We do not simply express disagreement about revisionary analyses and leave it at that: we offer and respond to arguments for and against them. I claim that the subject-change view cannot explain our argumentative practice: it cannot make sense of the kinds of arguments we offer, and the way we respond to these arguments.

Consider, for example, disputes about same-sex marriage. One kind of argument, invoked especially by opponents of same-sex marriage, appeals explicitly to the proper use of a word; for example, "We have always used 'marriage' to mean *a union of one man and one woman*". (This is, of course, enthymematic; the suppressed premise, apparently, is that we should continue to use the word as we have in the past.) Call arguments of this kind—arguments that concern the use of a word, such as 'marriage'—*metalinguistic arguments.*

Although opponents of same-sex marriage do advance metalinguistic arguments, this is not their only argumentative tactic; for example, they sometimes claim that the purpose of marriage is to produce children. This is (or at least seems to be) an argument about marriage (rather than an argument about 'marriage'.) Call arguments of this kind—arguments that concern something other than the use of a word—*first-order arguments.*

[4] It might be possible to form a *theory* that tails are legs; for example, perhaps animals with tails evolved from animals with five functional legs (so that tails are a sort of vestigial leg). In that sort of case, there need be no change of subject. But that is not the case I have in mind; the case I have in mind is one in which I simply stipulate (without argument) that I will use 'leg' to apply to tails.

1.1. The Argument Argument

As a matter of methodology, we should look for an interpretation of these arguments and our responses to them that *makes sense of what we are doing*. Of course, there may be cases in which our arguments just don't make sense—in which the parties to the debate are just confused or irrational. But in general, when a debate seems unconfused, we should try to understand what the participants in the debate are doing in a way that makes them unconfused. In this kind of case, we want to be able to say that the parties to a debate are rational, that their contributions to the debate make sense and contributes in a recognizable way to their aims. This suggests two requirements on our interpretation of a debate about a revisionary analysis:

1. It must give parties to the debate reasonably good epistemic status with respect to the things they say. In many typical cases, this will mean that the parties to the debate are saying things that they know, or at least justifiably believe, to be true. In other cases, the relevant epistemic status may be different. For example, if one party to the debate is attempting a reductio of her opponent's position, she should at least know or justifiably believe that what she says is a consequence of her opponent's position.
2. It must make assertions relevant to the debate. When a party to the debate makes an assertion, that assertion should serve some purpose: for example, by giving evidence for her position or evidence against her opponent's position.

I claim that no such interpretation of first-order arguments for or against revisionary analyses is available on the hypothesis that such analyses change the subject. Let's call the opponent of same-sex marriage O and the proponent of same-sex marriage P; and consider O's assertion of (1):

(1) The purpose of marriage is to produce children.

Now on the hypothesis that O is changing the subject, there are two ways of interpreting this utterance, since the change of subject would mean that there are two meanings of 'marriage' at issue in the conversation. On one interpretation, O is using 'marriage' in her own preferred way, to pick out a status that it is by definition impossible for same-sex couples to enter into; call this status *marriage$_1$*. On the other, O is using 'marriage' in P's way, to pick out a status that same-sex couples can enjoy; call this status *marriage$_2$*.

I claim that neither interpretation can make sense of the debate. Suppose first that O is using 'marriage' in her own preferred way. Then (let's suppose) her assertion of (1) might well express a truth—a truth that O is in a position to know. So we are able to satisfy our first requirement on an interpretation.

But we are not in a position to satisfy the second requirement. For the proposition that the purpose of marriage$_1$ is to produce children simply does not bear on P's position. Indeed, P may accept this claim; it is, after all, entirely consistent with the idea that 'marriage' should be used to mean marriage$_2$. So on the hypothesis that O is using 'marriage' to mean marriage$_1$, her utterance of (1) just looks irrelevant to the debate.

Suppose instead that O is using ‘marriage’ in P’s way. This may seem like a surprising way for O to behave, but it could make sense for O to take on P’s way of speaking for the sake of argument, in order to illustrate its consequences. For example, if you try to stipulate that ‘tail’ be applied to all and only round squares, I might object “But then there are no tails!” In this kind of situation, my first-order objection makes sense: I argue the way I do because I take it that you will be able to see that what I am saying is a consequence of your view, and I think it likely that you will regard this consequence as unacceptable.

But this cannot be what is going on in O’s assertion of (1). Though (1), on this interpretation, is something that P will regard as unacceptable, it is no consequence of her view. On the contrary: on the hypothesis that P is using ‘marriage’ to mean *marriage*$_2$, she should regard (1) as an obvious falsehood; moreover, it is an obvious falsehood that is obviously not entailed by her position. So on this interpretation, O’s utterance of (1) fails to meet either of our criteria: it is an obvious falsehood that O could neither reasonably believe nor reasonably believe bears any relevance to P’s position, and it is hard to imagine any reason that it might make sense for O to attempt to use it in the debate.

So the most natural interpretations of the debate on the subject change view fail to make sense of O’s assertion. As the debate continues, we will find the same pattern: the subject-change view simply cannot make sense of many of the argumentative moves. So even if the subject-change view can explain the fact that disputes about revisionary analyses involve genuine disagreement, it cannot explain how these disputes are conducted: it cannot explain the way we argue. And—given that these arguments really do make sense—this constitutes a powerful argument against the subject-change view, which I call *the argument argument*.

1.2. Two Responses

I now want to consider two ways of responding to the argument argument. The first response has it that we are adopting too limited a view of the possible purposes of the debate. True, O’s utterance of (1) will not convince anyone of a first-order proposition. But there are other cases in which an utterance of an apparently first-order sentence primarily functions to communicate metalinguistic information. For example, a person newly arrived at an Antarctic research station may ask with a shiver, “Is this cold?”; we could answer the question by pointing at a salient thermometer and saying, “This isn’t cold”. The thermometer is clearly visible: there is no mystery about the temperature, and no point in trying to communicate information about what the temperature is. Our utterance—despite the fact that it is not overtly metalinguistic—functions primarily to communicate information about how the word ‘cold’ should be used around here (Sundell (2011); Plunkett and Sundell (2013); and see also Barker (2002, 2013)).

I agree that our utterances often communicate metalinguistic information in the way described; and I also agree that there is *something* metalinguistic going on in the dispute: if I say, “Two women can marry each other”, one of the purposes of my utterance is to communicate information about how I think the word ‘marriage’ ought to be used. But two points need to be made. First, even if our dispute is in part metalinguistic, this does not entail (or even suggest) that the subject-change view is correct. It is entirely possible (indeed, common) for someone who uses a word with a particular meaning to communicate metalinguistic information about how that word ought to be used to someone who uses it with the very same meaning. (Plausibly, this happens every time I use a sentence to make an assertion: I assert the proposition expressed by my use of the sentence, but I also communicate that that sentence is correctly used in these circumstances to express a truth.)

Second, it is not obvious how the claim that our dispute is metalinguistic helps us make sense of O's utterance of (1). This utterance does not bear in any obvious way on the claim that 'marriage' should be used to express marriage$_2$. Even if we concede that O's main intent in uttering (1) is to convince P of a metalinguistic proposition, it is very hard to imagine any way that uttering (1) could reasonably be thought to aid in this goal. So the observation that the debate is or could be in part metalinguistic simply does not help meet our second desideratum. Even if the debate is metalinguistic, given the subject-change view, advancing arguments of this kind just doesn't make sense.

So the subject-change view cannot make good sense of the role an assertion of (1) plays in the debate. The second response to the argument is to question whether making sense of the debate should really be one of our aims. Perhaps arguments of this kind are just confused; perhaps an utterance of (1) really would express either the obvious falsehood that the purpose of marriage$_2$ is to produce children, or (more plausibly) the possibly true but irrelevant claim that the purpose of marriage$_1$ is to produce children.

If this is what is going on, then we would expect P (to the extent that she is unconfused) to reply simply by pointing out the fact that O's assertion of (1) is irrelevant. This is, after all, how a comparable debate would proceed in a case where it is uncontroversial that the subject has changed. For example, suppose I stipulate that 'leg' is to apply to tails, and say "Horses have five legs"; and suppose that you object that horses have only four legs. To the extent that I am unconfused, I will not see your objection as a threat to my view and I will not try to refute it; the right response for me to make is something like, "As you are using the term, I agree; but that fact—admitted on all sides—does not bear on my claim about the preferred use of 'leg'."

I admit that it would be possible in many cases for P to respond in something like this way. (P need not actually agree that (1) is correct on anyone's view of 'marriage', but she could at least argue that its correctness is irrelevant to her view: "That may be so, as you are using the term; but true or no, it does not bear on my claim".) I also agree that the subject-change view is plausible *if this kind of response is made and accepted*. (In my view, responding in this way goes toward making it the case that we are using the word 'marriage' with different meanings; see section 5 below.) But this is not the way debates about same-sex marriage typically proceed, and it is certainly not the only way they can proceed. When confronted with (1), it is very natural (and apparently rational) for P to feel a tension between this claim and her view. Given this tension, a typical, apparently reasonable response is to dispute the claim (e.g., by pointing out that elderly opposite-sex couples can be married, even though they cannot produce children).

Now perhaps the subject-change view can make some kind of sense of this response; the most natural suggestion is probably that P is adopting O's way of speaking in order to show that her view has absurd consequences, as in the "But then there are no tails" case. But this does not fully explain the phenomenon, because it does not explain P's reaction to O's utterance, and therefore does not explain why P responds in the way she does. In a typical case, P will feel that O has launched an attack on her view that puts her view under pressure; if O's claim (1) is true, P thinks, her view has a real problem. So she feels obliged to argue against (1).

This reaction—the feeling that what has been asserted is incompatible with one's view, so that one must offer first-order resistance—is something that the subject-change view cannot explain. Of course, it is possible that this reaction is grounded in confusion—that in fact there is no such incompatibility, no such pressure on P's view. No one thinks that debates over same-sex marriage are conducted under conditions of ideal rationality, and it is certainly possible that we make mistakes. But, as I have emphasized, this reaction is utterly normal. Most participants in this kind of debate—even philosophers raised on a strict regimen of distinguishing use and mention—feel that assertions like (1) constitute a threat to their view, and a corresponding need to engage in first-order argument. The claim that we are all so confused about what is going on is to be resisted as long as some alternative interpretation is available (and, as I will go on to show, one is available once we reject the subject-change view).

I conclude that the replies to the argument argument fail: the subject-change view simply fails to make sense of the way we argue for and against revisionary analyses; and even in those cases where the subject-change view has a partial story to tell, it often fails to explain the whole phenomenon.

2. Concepts

Let me try to relate the way of thinking about the threat of subject change that we have been developing to some things Haslanger says. As Haslanger puts the point, the worry is that when we have changed the subject "The term [. . .] expresses a different concept than it did" (2012c: 394). The picture seems to be that concepts are or correspond to word meanings, so that distinct words (i.e., distinct word types) express different concepts. This way of putting the point suggests that terms always or at least typically express only one concept at a time (in the sense that for a given person at a given time, there is a single concept expressed by all typical uses of the term). But this would sit awkwardly with other things Haslanger says about concepts. Haslanger distinguishes between two analytic projects (in addition to her own "ameliorative project"). One involves collecting our "intuitive" reactions to hypothetical cases and principles and trying to reach a set of principles that best systematizes these reactions; Haslanger regards this as producing an analysis of our *manifest concept*—"the concept I thought I was guided by and saw myself as attempting to apply" (2012c: 388). The other that involves collecting and systematizing empirical information about our actual classificatory judgments; Haslanger regards this as producing an analysis of our *operative concept*—"the concept that best captures the type we are concerned with" (p. 387). One's manifest concept can be distinct from one's operative concept, as the following example from Jennifer Saul illustrates:

> If we are investigating our manifest concept of democracy, then, we will probably arrive at a standard that requires (at least) elections that are free of voter intimidation, the counting of all

> ballots, equal access to polling places for all voters, and so on. [...] If we are investigating our operative concept of democracy, [...] we may well find that we apply the term far more broadly than our manifest concept would suggest. It may turn out that our operative concept of democracy requires only regular elections and that all adult citizens be formally permitted to vote. This is compatible with substantial voter intimidation, great variation in access to polling places, and ballots going uncounted. We could find out that this is our operative concept by noticing that in fact we apply the term 'democracy' even when we know that voters were intimidated, ballots went uncounted, and so on. (2006: 124)

So on the one hand, Haslanger thinks that change of subject involves using a word to express a different concept; but on the other, she thinks that words are often associated with multiple concepts. And while this combination of views isn't incoherent, it seems awkward. Fortunately, the difficulty is not deep; resolving it just requires some terminological regimentation.

I take beliefs (desires, etc.) to be mental representations. There is a sense in which different thinkers never share mental representations: my representations are in my head, and yours are in your head. But there is another sense in which you and I might believe the same thing; we do not share a belief *token*, but we have a belief of the same *type*. To say that a thinker possesses a particular concept is to say that she has mental representations of a certain type.

There are many ways of grouping token representations into types—type-individuating them, in the jargon—and different ways may be appropriate for different purposes. For example, it may make sense to individuate mental representations in an extremely fine-grained way, so that different thinkers rarely if ever share concepts, and one's concepts very often change over time, as one's beliefs and inferential dispositions change (e.g., Block 1986). But there are other purposes for which it makes sense to individuate concepts in a less fine-grained way, so that a thinker can retain the same concept over time despite changes in belief, and different thinkers often share the same concepts.

One such purpose is tracking the kind of facts about certain kinds of agreement and disagreement, as well as closely related facts about successful communication and argument. We have already conceded that some genuine disagreements do not involve parties asserting contradictory contents, so that there can be genuine disagreement despite difference in meaning; but we have also seen that disagreements that support a full range of first-order arguments are not like this. The most natural way to make sense of these disputes will involve individuating concepts in a less fine-grained way.

Here is a simple way of developing the picture. Suppose that you tell me, "Fido is a dog". You are expressing a certain belief—a mental representation of a particular type—and I understand you only if I entertain a mental representation of the same type. If I believe you, I am justified only if the mental representation I form is a representation of the same type you expressed (at least in typical cases). My utterance of "Fido is a dog" expresses or constitutes agreement with your utterance of the same-sounding sentence just in case the two utterances are expressing beliefs of the same type. And so on.

Similar sorts of points can be made even if we are paying attention to a single thinker. If at one time I say, "Fido is a dog", but at a later time I say, "Fido is not a dog", have I changed my mind? The answer depends on whether I reject at the later time a thought of the type that I had earlier accepted, and this amounts to whether 'Fido' and 'dog' (and perhaps also 'is' and 'a') express the same concepts in both utterances. If I am inclined to say both "Fido is a dog" and "Fido is not a dog" at the same time, am I incoherent? Again, the answer depends on what types of belief I would express by these sentences.

If we want concepts to do this kind of work, we have to individuate them so that thinkers can have the same concepts over time and across changes of view. I may disagree with an interlocutor about many weighty matters, matters central to our discussion; I may despite this understand her, and agree with much of what she says. This requires that we share concepts of the same type. I may change some of my beliefs on a given topic while still retaining others, and again this requires individuating concepts in such a way that I can retain concepts of the same type despite my change of view.

Let's not dispute whether there are other ways of treating these phenomena; I am happy to concede that there could be. There are, after all, many ways of individuating mental representations. Our purpose now is simply to lay out some stipulations about one useful way of using the word 'concept'; I am stipulating that concepts are individuated in at least a somewhat coarse-grained way, and indicating (admittedly roughly) how the individuation is to go.

If we are individuating concepts in this way, then we should not think of the manifest "concept" and the operative "concept" as being distinct concepts (at least in typical cases). If they were, then we should see no conflict between our judgment that democracy requires that all citizens have equal access to polling places, and our judgment that (say) the USA is a democracy, even if we discover that many regions of the USA have enacted voter identification laws specifically designed to reduce the ability of certain groups to vote. Since my judgments about the theoretical principle (which brings to bear my "manifest concept") and my judgment about the USA (which brings to bear my "operative concept") involve different concepts, there is no inconsistency between them. We could simply rest content with our view.

It is quite possible that some cases fit this model. But the most natural understanding of the "democracy" case does not seem like this. If we judge that the USA is a democracy, but fails to meet some of the general conditions we judge necessary for democracy, this would generate a real feeling of conflict. The judgments just seem inconsistent. Anyone who finds themselves in this position and fails to take steps to reconcile her view—coming to believe that the USA is not a democracy, modifying the principle that democracy requires equal access, etc.—is *pro tanto* irrational.

So if the result of an "analysis of our manifest concept of democracy" and the result of an "analysis of our operative concept of democracy" are not analyses of distinct concepts, what are they? On the way of individuating concepts that we are now considering, we should think of these not as analyses of different concepts, but *as competing proposed analyses of one and the same concept*. To get clear on why this makes sense, we will need to turn to the notion of analysis.

3. Analyticity

Haslanger proposes (W) as an analysis of 'woman', and I take it that this proposal entails that (W) is an analytic truth.[5] But what exactly does it mean for a claim to be analytic, or to be proposed as an analysis?

Analysis and *analyticity* came under heavy scrutiny after Quine's (1953, 1966a,b) critique. Traditional accounts of analyticity combined metaphysical and epistemological elements, usefully distinguished by Boghossian as follows:

> **Epistemic analyticity** A sentence S is epistemically analytic iff "mere grasp of S's meaning by T suffice[s] for T's being justified in holding S true". (1997: 334)
>
> **Metaphysical analyticity** A sentence S is metaphysically analytic if "in some appropriate sense, it owes its truth-value completely to its meaning, and not at all to 'the facts'". (1997: 334)

It's hard to see how metaphysical analyticity could be relevant to revisionary analyses (like Haslanger's), and that's a good thing: with Boghossian and most others, I take Quine's attack on metaphysical analyticity to have been decisive. Our meaning that Hesperus is Hesperus by "Hesperus is Hesperus" doesn't make it the case that Hesperus is Hesperus; what makes it the case that Hesperus is Hesperus is a thought- and language-independent fact.

Boghossian himself defends epistemic analyticity, and although this cannot be the right account of what it is to be analytic in the sense relevant to our debate—manifest analyses, for example, will typically be known on the basis of empirical investigation, not justifiably believable by anyone who grasps them—investigating Boghossian's case will help reveal another element in notion of analyticity that will provide a framework for understanding revisionary analyses like Haslanger's.

The foundation of Boghossian's defense of epistemic analyticity is a view about what it takes to possess certain concepts, such as the concepts of geometry: "grasp of the indefinables of geometry consists precisely in the adoption of one set of truths involving them" (1997: 348). More generally:

> **Adoption Grasp** A term t is *grasped by adoption* (or *a-grasped*) just in case there is some set of truths S which is such that one understands t iff one accepts each member of S.

Boghossian's main interest is in logic; there, his idea is that (e.g.) one possesses the concept of conjunction just in case one accepts principles like, "If p and q, then p". He claims that this account of grasp "generates" (1997: 348) the following account of how logical constants get their meanings:

> It is by arbitrarily stipulating that certain sentences of logic are to be true, or that certain inferences are to be valid, that we attach a meaning to the logical constants. More specifically, a particular constant means that logical object, if any, which would make valid a specified set of sentences and/or inferences involving it. (1997: 348)

[5] I will also extend the term 'analytic' to obvious entailments of analyses, so that (e.g.) if (W) is an analysis, then it is an analytic truth that women are oppressed).

We can generalize the proposal in the following way:

> **Implicit Definition** A term t is *implicitly defined* iff we attach a meaning to t by arbitrarily stipulating that certain sentences involving t are to be true. More specifically, t has that meaning, if any, which would make true a specified set of sentences involving it.

If *Implicit Definition* is true, then there is a sense in which certain sentences are true by convention. But Boghossian argues convincingly that there is no threat that implicit definitions will result in conventional truths of the sort that Quine's arguments show to be problematic. Perhaps the simplest way to see the point is on the view that propositions—the meanings of sentences—are the primary truth bearers, so that a sentence is true only if it expresses a true proposition. We can imagine propositions having their truth values fixed antecedently to any human linguistic activity. Stipulation makes it the case that a particular sentence expresses a particular proposition; it does not make it the case that the proposition expressed by the sentence is true.

Consider an example. Suppose that the length of a certain rod r is l, so that the proposition that the length of r is l is true (and true independently of human thought and linguistic activity). Then we can introduce the term 'meter' by stipulating: "The length of r is one meter". Given the meanings of the other words in the sentence ('The', 'length', 'of', etc.), the syntactic structure of the sentence, and the composition rules of the language, there are a variety of propositions that the sentence might express, depending on what 'meter' means: the proposition that the length of r is h, the proposition that the length of r is l, the proposition that the length of r is m, etc. By stipulating that "The length of r is one meter" expresses a truth, we fix on a particular candidate—the only true candidate in the vicinity: the proposition that the length of r is l. This in turn fixes the meaning of 'meter': given the extant meanings of 'The', 'length', etc., and the facts about syntax and composition, "The length of r is one meter" expresses the proposition that the length of r is l only if 'meter' picks out l.

It is clear in this example that our linguistic stipulations have not made true any non-linguistic facts. The facts about length were fixed prior to our stipulation. Something like this model of Implicit Definition provides the foundation of what we will develop in what follows. But in order to avail ourselves of this foundation, we will need to rethink what it takes to grasp the meaning of an implicitly defined term, as well as the notion of stipulation.

3.1. Implicit Definition and Adoption Grasp

Implicit Definition would fit neatly with *Adoption Grasp*. But the two are in principle separable; an *Implicit Definition*-style account of how words get their meaning is compatible with any number of accounts of how meaning can be grasped.

Here is an example. Consider Kripke's (1980) causal picture of meaning: a word is introduced at an initial baptism, and later uses get their meaning in virtue of their connection to this initial baptism. A typical instance of such a baptism might involve pointing at a certain baby and saying something like, "I hereby name this baby 'Ansel'". This is an arbitrary stipulation that attaches a meaning to the word 'Ansel'. But it does not exactly fit the template described in *Implicit Definition*, since we are not stipulating that any sentence is true. So let's reformulate our initial baptism: "I hereby stipulate that the following sentence is true: 'The baby in room 110, Ninewells Hospital, at 10:00am GMT 25th April 2013, is Ansel'". Plausibly, then, the meaning of 'Ansel' has been fixed by arbitrarily stipulating that a certain sentence involving 'Ansel' is true. So 'Ansel' is implicitly defined.

But note that this does not entail that *Adoption Grasp* is true of 'Ansel'. On the contrary: Kripke's view has it that later speakers grasp 'Ansel' in virtue of their causal connection to the initial baptism. He specifically rejects the idea that speakers' competence with a term depends on entertaining the stipulation with which a word is introduced. So a word can be implicitly defined but not a-grasped.

It might be objected at this point that on Kripke's story, the meaning of 'Ansel' as it is used by later speakers is not determined solely by the stipulation. Instead, it is determined by the stipulation *and the later speakers' causal connection to that stipulation*. If that is right, then 'Ansel' is not implicitly defined.

This is a fair point. In general, we should allow that a stipulation can play a significant role in fixing the meaning of a word, while allowing that other factors may also play a role. So we should allow for the possibility that stipulation *partially* fixes the meaning of a word:

> **Partial Implicit Definition** A term t is *p-implicitly defined* iff t has the meaning it does at least in part in virtue to an arbitrary stipulation that certain sentences involving t are to be true. More specifically, if t has a meaning, it has a meaning which would make true a specified set of sentences involving it.

What we have established so far, then, is that partial implicit definition and adoption grasp can come apart, at least in principle; a term can be p-implicitly defined but not a-grasped. And this is a good thing: Boghossian's defence of epistemic analyticity has come under sustained attack, notably in the work of Timothy Williamson (2006, 2007). Williamson points out that experts with idiosyncratic theoretical views can dissent from even the most basic and plausible truths involving a word. For example, Williamson suggests that a thinker might reject an elementary logical truth like "Every vixen is a vixen"—an instance of a principle acceptance of which might well be held to be constitutive of understanding of 'every'—because she maintains that 'every' has existential import, and denies that there are vixens. Similarly, Williamson points to logicians who reject *modus ponens*. Such thinkers seem to be counterexamples to the claim that logical terminology is a-grasped: these experts seem to understand 'every' and 'if', while rejecting principles in which this understanding putatively consists.

There have been, of course, a variety of attempts to resist Williamson's examples (e.g., Boghossian 2011); but I, for one, continue to find the examples convincing. But there is need for care. Williamson's examples refute the claim that Adoption Grasp is true of words like 'every' and 'if'. They do not refute the claim that these words are p-implicitly defined. For example, one could imagine a Kripke-style view on which 'if' means what it does because of our causal connection to an ancient stipulation. (I do not claim that this is a plausible story, only that it is unscathed by Williamson's examples.)

The crucial lesson of Kripke's causal theory, for our purposes, is just that the stipulation that p-implicitly defines a word need not be grasped (known, believed, entertained) by competent users of the term. The meanings of my words can be fixed by stipulations that are spatially and temporally distant from me—stipulations that I am in no position to know of.

3.2. From Implicit Definition to Metasemantic Analyticity

A further thread in Quine's critique is that the notion of analyticity is unclear, and that attempts to clarify or explain it are either circular or rely on equally unclear notions of meaning or synonymy. Carnap's response was to offer an account that suited the formal languages that are the primary focus of his *Meaning and Necessity* (1956).[6] In developing such a formal language, we may stipulate that certain sentences are to be regarded as *meaning postulates*. Carnap's idea is that a sentence is analytic just in case it is entailed by the set of meaning postulates.

Carnap's use of meaning postulates is meant to ground *metaphysical* analyticities: analyticity is understood as truth in virtue of meaning. But meaning postulates were also taken up by proponents of Montague grammar in the 1970s and 1980s (see, e.g., Montague 1974: 212–13; Dowty et al. 1981: 224–5). In these works, the metaphysical ambitions have largely dropped away. The role of meaning postulates is to help to fix the meaning of certain terms. The basic idea is simple. In Montague grammar, expressions are assigned semantic values relative to a *model*. To say that a certain sentence is a meaning postulate is to say that we will only consider models that make that sentence true. Since logical truth is defined as truth at every model, meaning postulates will be logical truths.

Let's consider a simplified example. Suppose that the semantic values of one-place predicates are sets of individuals (the individuals of which the predicate is true), so that a sentence like (2) is true if and only if the semantic value of 'bachelor' is a subset of the semantic value of 'unmarried':

(2) For all x, if x is a bachelor then x is unmarried

By letting (2) be a meaning postulate, we restrict our attention to models relative to which the sentence is true, so models relative to which semantic value of 'bachelor' is a subset of the semantic value of 'unmarried'. The idea is that although this does not completely determine the meanings of 'bachelor' and 'unmarried', it puts a constraint on, and so partially fixes, them.

The traditional metaphysical notion of analyticity had it that some sentences were true in virtue of their meanings. The use of meaning postulates that we are now considering flips this on its head: some expressions have the meanings they do in virtue of our holding certain sentences true. Instead of *truth in virtue of meaning*, we have *meaning in virtue of truth*: by stipulating that we will only consider interpretations of our words that make a certain sentence true, we have put a constraint on, hence partially determined, the meanings of the words that compose that sentence.

It is no accident if meaning postulates remind one of partial implicit definitions. They play an extremely similar role; in effect, meaning postulates p-implicitly define

[6] Carnap characterizes his project here as *explication*, which he sees as involving the replacement of one concept by another; but (of course) on my view we need not accept this characterization.

the words that compose them. I suggest that sentences that p-implicitly define (some of) the words that compose them deserve to be thought of as *analytic* in the following sense:

> **Metasemantic Analyticity** A sentence is metasemantically analytic with respect to a word (or a use of a word) iff the meaning of that (use of the) word is partially fixed by the stipulation that the sentence is to be true (i.e., iff the sentence p-implicitly defines the word).

Metasemantic analyticity has, perhaps, received less attention in the literature on analyticity than epistemic and metaphysical analyticity, in part because it plays a less important role in the metaphysical and epistemological debates to which analyticity has often been turned (e.g., there is no clear role for metasemantic analyticity in giving an account of metaphysical modality, or of a priori knowledge). But there is nonetheless a notable tradition of using the word 'analyticity' to pick out metasemantic analyticity. Meaning postulates are, after all, what is analytic on Carnap's account, and one of their primary roles (the role picked up on by Montagovians) is to fix meaning. And metasemantic analyticity is the notion most relevant to the many discussions of analyticity in the work of Fodor and his collaborators. To take just one representative example, Fodor and Pylyshyn write, "analytic beliefs can't be revised without changing the content of (some or all of) their conceptual constituents; that is, they can't be changed without equivocating" (2015: 57). This passage only makes sense if what is at issue is something like metasemantic analyticity (applied to beliefs and the representations that compose them rather than to natural language sentences and words): the idea is that some of our beliefs *determine the content* of their representational constituents, so that changing these beliefs would correspondingly change the content of the constituents. What is at issue is how the contents of certain mental representations are determined: that is, the metasemantics of those representations.

So there is an established pattern of using 'analyticity' to pick out metasemantic analyticities in the literature. On my view, proposed revisionary analyses, such as Haslanger's (W), should be seen as attempts at metasemantic analysis.

4. Timing

Let's pause to take stock. I began by showing that the subject-change view cannot explain how we argue for revisionary analyses, and I have now developed the notion of metasemantic analyticity, which I claim to be relevant to revisionary analysis. But at this point it may be unclear how my rejection of the subject-change view can be reconciled with my embrace of metasemantic analyticity. For isn't the proponent of the subject-change model precisely endorsing the claim that revisionary analyses are a kind of stipulation that introduces a new meaning for a word (and hence changes the subject)? After all, what makes it the case that post-revision uses of the word have a different meaning? Surely it is the revisionary analysis—the stipulation—but that is just to say that the revisionary analysis partially fixes the meaning of the word and so is metasemantically analytic.

I propose to grant that revisionary analyses—*if successful*—are metasemantically analytic. (We will return momentarily to the question of what makes a revisionary analysis successful when it is.) But we can grant this without embracing the subject-change view. On the *subject continuity model* that I advocate, there is no new meaning; a successful stipulation fixes the meaning of the word as it was used all along.

It must be admitted that the subject-change view would fit well with our standard conception of stipulation. Stipulated definitions are typically thought of as coming when a word is first introduced. But what motivates this conception? One possible motivation comes from Adoption Grasp: if users of a word must grasp the stipulated definition, then the stipulation must occur before the word is used. But as we have seen, there are popular and plausible views on which p-implicit definition (and hence metasemantic analyticity) can be prised apart from Adoption Grasp. On these views, the stipulation that p-implicitly defines a word can be made by an unknown person in the distant past, and the stipulation need not be known by current users of the word as long as they are causally connected to it in the right way. So users of a word need not accept the sentence that p-implicitly defines that word; a word I use can get its meaning (at least in part) from a stipulation made by someone else, of which I am not even aware.

The subject continuity view is a development of this idea. Although the Kripkean causal picture rejects the view that users of a term must grasp the stipulation that introduced, it retains the idea that stipulations that fix the meanings of my words must be temporally prior to my use of those words: *first* someone stipulates, and that gives *later* uses of the words meaning. In my view, the key to understanding revisionary analyses is that they involve a stipulation that (partially) fixes the meaning of prior uses of the word: in this kind of case, *first* we go about using a word, then *later* we make the stipulation that gives meaning.[7]

On this view, in typical cases of revisionary analysis there is no change of meaning and no change of subject. Words are used univocally throughout a discourse; each word expresses the same concept throughout, even when one party proposes a revisionary analysis. What happens when a proposed revisionary analysis is successful is that stipulations that fix the meaning of a word (throughout the discourse) have emerged late in the discourse.

The subject continuity view may seem like a radical departure from familiar conceptions of stipulation and analyticity—perhaps it constitutes a somewhat revisionary analysis of stipulation and of analyticity—but it is inevitable given the conclusions that we have already reached. The argument argument gives us reason to reject the subject-change view: we should maintain that words are univocal—that no meaning-change occurs and no new concepts are introduced—before and after revisionary analysis. But we have also granted that successful revisionary analyses are in fact (metasemantic) analyses; they partially fix the meanings of words. Since they

[7] The general term for views on which the meaning of an expression or the content of a thought can depend on facts about later times is *temporal externalism*. For discussion and defence, see Jackman (1999, 2005); Ball (forthcoming).

do not introduce new meanings or new words, they must fix the meanings that words have had all along. And this is just the core claim of the subject continuity view.

5. Successful Analysis

But the subject continuity view may seem to leave us with a serious problem. On traditional views of stipulation, such as the subject-change view—views on which stipulation precedes use—in typical cases there will be no serious difficulty in determining whether a given stipulation was successful: if we make a stipulation, and then go on to use the word stipulatively introduced, the stipulation was successful.[8] But on the subject continuity view, the picture is more complex. We may have multiple parties to a dispute, each proposing a competing revisionary analysis, all of which purport to fix the meaning of the same set of uses of a term. What makes it the case that one of these proposed analyses fixes the meaning, while the others do not?

In order to answer this question, I want to begin by considering a metasemantic story that some philosophers have taken to support the subject-change view that we have just rejected: the view that meaning is use. As this slogan is typically interpreted in this debate, the view that meaning is use amounts to the view that the extension of a term as used by a speaker is determined by that speaker's dispositions to apply the term and to withhold application of the term. Call this *meaning is use now* (MIUN).[9]

MIUN suggests that revisionary analysts are changing the subject, since (at least to the extent their dispositions to apply the term are consistent with their proposed analysis) revisionary analysts will exhibit quite a different pattern of applying and withholding the term at issue; for example, the proponent of same-sex marriage will apply the term 'marriage' to same-sex unions while her opponent will not, and Haslanger will withhold 'woman' from females who are not subordinated (if such there be).

The problem with MIUN is similar to the problem with the subject-change view: it cannot make sense of our argumentative practice. We have already seen that the subject-change view has trouble making sense of the full range of first-order argument, and to the extent that MIUN is committed to the subject-change view, it will share this problem. I now want to emphasize an additional problem with the view that meaning is use now: it cannot make sense of how we respond to arguments. In particular, it cannot make sense of apparently rational and correct change of view. We may be convinced by a revisionary analysis. Suppose that I once believed that same-sex marriage was impossible (and said things like, "A man can only marry a woman and vice-versa"), but I was convinced by some argument that same-sex marriage is possible after all. To keep the case simple, suppose that this last belief is stable (and so not overturned by further argument), and that most or all of my fellow speakers come to share my view.

[8] Perhaps we need to make some further allowances to prevent the stipulative introduction of 'tonk' and the like, but these complexities are not relevant to our present concerns.

[9] See Sundell (2012) for a defence of this kind of view from a proponent of the idea that revisionary analysis requires meaning change.

How would I describe my change of view? One thing that I would take to have changed is my metalinguistic beliefs; I used to take "Same-sex couples can marry" to express a falsehood, but I came to believe that it expresses a truth. But this is neither the only nor the most central belief that I would take to have changed; I would also take myself to have believed that a man can only marry a woman, and to have come to reject this belief and to believe instead that same-sex couples can marry. (I might say things like, "I used to think that two women could not marry each other, but now I see that I was wrong.") That is, I would take my first-order views about marriage to have changed.

In section 1.1, I suggested that as a matter of methodology, we should look for an interpretation of our debates about revisionary analysis that makes sense of what participants in the debate are doing. The same is true here: we should look for an account of the debates that makes sense of our informed judgments about them. Of course, we are not infallible about what we believed and meant. But other things equal—in cases when we are not obviously confused, irrational, suffering from memory failure, and so on—an interpretation that makes sense of our judgments is to be preferred to an interpretation that does not. And since all that is at issue here is remembering my past views, and judging whether my current view is inconsistent with them, the most obvious way an interpretation can make sense of my judgment is by making it correct. (Perhaps this is the only way to make sense of it: since nothing is wrong with my memory, it seems hard to explain how I could fall into error.)

This is something that MIUN cannot deliver. According to MIUN, my dispositions to apply 'marriage' before I heard of the revisionary analysis make it the case that 'marriage' as I used it picked out a relation that men can stand in only to women and vice-versa, while my dispositions after I accept the revisionary analysis make it the case that 'marriage' as I use it picks out a relation that men can stand in to other men. So MIUN cannot make me right when I say, 'I used to think that two women could not marry each other, but now I see that I was wrong'. As I use 'marriage' at the time of this utterance, it isn't true that I used to think that two women could not 'marry' each other, and what I used to think wasn't wrong. So in order to make me correct, what is wanted is a view that predicts not that the meaning of 'marriage' has changed, but that what I meant all along is the relation that same-sex couples can stand in. In this case, the revisionary analysis has succeeded: claims like "Two men can marry each other" are metasemantically analytic of 'marry' as I used it all along.

Of course, there are cases in which MIUN delivers more plausible results. But even in these cases, it gets the right answer for the wrong reason. Suppose I accept that being a woman is a matter of biological fact, and say things like "Necessarily, a person is a woman iff she has two X chromosomes"—until a philosopher puts forward (W) as an analysis of 'woman'. I consider the analysis, and despite some intriguing arguments, I decide that the balance of considerations pull against accepting it. So I judge that (W) is false, and that "Necessarily, a person is a woman iff she has two X chromosomes" is true. To keep the case simple, suppose that my rejection of (W) and my continued acceptance of the biological view are stable (and so not overturned by further argument), and that most or all of my fellow speakers share my views (so also reject (W) and accept the biological account). And suppose further that in the end even the proponent of (W) changes her mind and becomes convinced that (W) is false.

Again, on the assumption that I am being reasonable—not unduly stubborn, not neglecting relevant evidence, etc.—then my judgment is something we should aim to make sense of, and the obvious way to do this is by making it correct. So we want a view that predicts that in this case, 'woman' as I use it picks out a biological concept that makes no reference to social status. (W) purports to be an analysis of 'woman', but in the end fails to do metasemantic work.

MIUN delivers this result; but we have already seen that it fails elsewhere, so that we need to look for a different view. One suggestion would be to appeal to a wider range of dispositions: not only dispositions to apply the term, but also dispositions to respond to new information and so on. For example, perhaps the difference between the described 'marriage' case and the described 'woman' case turns on my dispositions to respond to revisionary analyses. This strategy may make sense in some cases. But in general, our dispositions are messy. I may be disposed to respond favourably to some arguments (in some moods, when presented by some interlocutors, in some tones of voice) but not others. I am not simply disposed to come to believe that same-sex marriage is possible no matter what: whether I come to believe this will depend on exactly which arguments I encounter, and in which circumstances. My dispositions just don't fix my meaning—or at least, not until they are activated in particular circumstances. (That is why accepting a view can count as a stipulation, even though it is also a response to an argument.)

The obvious difference between the success of the revisionary analysis of 'marriage', and the failure of the revisionary analysis of 'woman' (in the described cases) is that the successful analysis is *accepted* and the failed analysis is *rejected*. I suggest that this obvious difference is what does the metasemantic work. At least in kind of simplified case we are considering here (in which the debate is resolved), what makes a given putative metasemantic analysis successful is that it is accepted.

Further, I suggest that once we have given up the subject-change view (and so accepted that there a single meaning/concept throughout the discourse), this is an extremely natural conclusion. We seem to have two possible accounts of what is going on in simple cases like the ones we are discussing:

> **Beginning of Debate** The facts that fix meaning—including the stipulations and beliefs that are metasemantic analyticities—must be in place at the beginning of a discourse. They fix the meanings of our terms and the content of our concepts throughout the rest of the inquiry. So revisionary analyses always fail; if, at the beginning of inquiry, we accept "Marriage is the union of one man and one woman", then we cannot successfully advocate same-sex "marriage" without changing the subject.
>
> **End of Debate** The facts that fix meaning can be determined by the discourse. So revisionary analyses can succeed; if everyone comes to accept on reasonable grounds that same-sex marriage is possible, this goes toward making it the case that 'marriage' picks out a relation that can hold between two men or two women.

But surely the End of Debate picture is much to be preferred. It seems absurdly uncharitable—perverse, really—to maintain that if everyone goes from rejecting the possibility of same-sex marriage to accepting it (and regarding their prior first-order views as incorrect), they are making a mistake, even if this transition is grounded in apparently reasonable argument. It looks like we have a choice of preferring our views *prior to inquiry* or our views *after inquiry*: but surely (other things equal) our views after inquiry are to be preferred. (After all, it is only then that we have had a chance to consider the evidence.)

Of course, this is not to say that discussion can never go awry, or that we can never be mistaken or confused. It is only to say that the views of people in a better epistemic position are *ceteris paribus* to be preferred to the views of those in a worse epistemic position.

5.1. Agreeing that a Dispute is Verbal, and Other Complications

We have made simplifying assumptions about the cases we have discussed: that everyone comes to agree, so that the debate is settled once and for all. And of course in many cases this kind of assumption will be entirely unrealistic. What predictions do we make about more realistic cases? One kind of complication we can easily make sense of. Some debates end not in agreement, but in acknowledgement that the disputants were talking past each other, so that the debate was merely verbal. We already mentioned that opponents of same-sex marriage sometimes claim that proponents are trying to redefine 'marriage'. Typically (and reasonably, on the view we are developing) proponents resist this charge. But proponents might also (and again, reasonably) agree that they are using 'marriage' differently, and proceed to dispute about how the word should be used. In this case, too, I think that the considered judgments of the parties to the conversation should be taken seriously: if everyone comes to the conclusion that the subject has changed, that can make it the case that the subject has in fact changed.

The case where the (first-order) debate is resolved (or perhaps dissolved) when we agree that we are talking past each other reveals a final data point worth attending to. In this case, we may re-evaluate our earlier reactions to arguments and judge some of them to have been confused. To revisit our example from section 1.1, if P has decided that the dispute about same-sex "marriage" is merely verbal, then P should also decide that O's assertion of (1) did not put any pressure on her view. Instead of arguing against it, she should have rejected it as irrelevant.

As before, an interpretation of the debate should seek to vindicate this considered judgment. And that is exactly the result we get when we take our end-of-debate claims to be metasemantically analytic. When we are evaluating (1), for example, there are three possibilities to consider:

1. If O convinces P—so that O and P end up agreeing that same-sex marriage is by definition impossible—then the assertion of (1) will seem to have been a good argument: an argument that bears on the issue, and expresses a reasonable belief. We vindicate this appearance: the assertion of (1) bears on the issue because it is inconsistent with the view that P held, and was a reasonable commitment for someone with O's view. (If P and O end up agreeing on (1), it too may end up as a metasemantic analyticity, hence true.)
2. If P convinces O—so that O and P end up agreeing that same-sex marriage is possible—then the argument will seem flawed. If the assertion were true, it

would have been relevant (since it is incompatible with P's position), and perhaps O had good reason to think it true; but a fuller consideration of the evidence revealed it to be false. We vindicate this appearance: since there is no meaning change, the assertion will be relevant—it is inconsistent with P's position—but since P's position is correct, the assertion will turn out to be false.

3. If P and O agree that they are talking past each other, then the argument will seem irrelevant: either a truth that does not bear on P's position, or a falsehood (and an obvious falsehood, given their current evidence). We vindicate this appearance: since in this scenario the word 'marriage' is used with two distinct meanings in the conversation, there are two relevant interpretations of (1), one of which is an irrelevant truth and the other of which is a relevant falsehood.

But many typical disputes about revisionary analyses—at least in philosophy—do not resolve at all: we do not agree, and we do not agree that our dispute is verbal either. And in this kind of case the considerations we have given so far make no definite prediction. The basic moral of our story is that facts about our future use can play a meaning fixing role—can be metasemantically analytic—with respect to our present use. Exactly how meaning gets determined in any actual case will no doubt be a complicated matter that requires weighing a variety of factors: facts about past, present, and future dispositions and uses, as well as (probably) other sorts of facts about the social and physical environment (which may include facts about metaphysical naturalness and so on). The view we have developed does not answer every metasemantic question. It only introduces a new factor to be considered—albeit a factor that is decisive in many significant cases.

6. Conclusion

This chapter has defended the following claims:

1. Revisionary analysis does not typically involve change of meaning or the introduction of new concepts.
2. Revisionary analyses can be seen as a kind of (attempted) stipulation, and so as attempts at giving metasemantic analyticities; but if they are successful, they fix the meaning of words as we have always used them (even before the stipulation took place).

These claims together constitute a view that repudiates the idea that philosophy is purely descriptive—in making judgments, we are also making our meanings and concepts, not merely "leaving everything as it is"—but at the same time repudiates the idea that we are changing the subject, or choosing between different concepts. (Making judgments, forming beliefs, choosing theories? Yes. Conceptual engineering or conceptual ethics (understood as something other than this)? No.)

Let me conclude by returning to Jackson's worry: does the stipulation that "a free action is one such that the agent would have done otherwise if he or she had chosen to" result in a change of subject that turns "interesting philosophical debates into easy exercises in deductions from stipulative definitions together with accepted facts"? Jackson's worry is not completely unfounded, since it is *possible* that this stipulation would change the subject; it would change the subject if the parties to the debate treat it as changing the subject, for example by refusing to engage with first-order argument for and against it. But philosophers who make this kind of claim do not usually have this attitude. It is perfectly possible to introduce a particular account of free action, recognizing it as a stipulation that will have a metasemantic effect if accepted, but still engaging in interesting philosophical debates. This is exactly what proponents of revisionary analyses usually do.

So could women be analytically oppressed? Yes: (W) could be metasemantically analytic. Are they analytically oppressed? That depends on whether proponents of (W) make a case that convinces us.

Acknowledgements

I am grateful to audiences at the Universities of St Andrews and Oslo, to the participants in Emanuel Viebahn's seminar at Humboldt University, to the faculty reading group at the University of Texas at Austin, and to Brice Bantegnie, George Bealer, Liz Camp, Herman Cappelen, Daniel Elstein, Patrick Greenough, Mark Hinchliff, Gail Leckie, David Plunkett, Sarah Sawyer, and Kevin Scharp.

References

Ball, D. forthcoming. Relativism, Metasemantics, and the Future. *Inquiry*. Forthcoming in a special issue edited by Henry Jackman.

Barker, C. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy* 25 (1):1–36.

Barker, C. 2013. Negotiating Taste. *Inquiry* 56 (2–3):240–57.

Bigelow, J. and Schroeter, L. 2009. Jackson's Classical Model of Meaning. In I. Ravenscroft (ed.), *Minds, Ethics and Conditionals: Themes from the Philosophy of Frank Jackson* (pp. 85–110). Oxford: Oxford University Press.

Block, N. 1986. Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy* 10:615–678.

Boghossian, P. 1997. Analyticity. In B. Hale and C. Wright (eds.), *A Companion to Philosophy of Language* (pp. 331–68). Oxford: Blackwell.

Boghossian, P. 2011. Williamson on the A Priori and the Analytic. *Philosophy and Phenomenological Research* 82:488–97.

Burgess, A., and Plunkett, D. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, A., and Plunkett, D. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Carnap, R. 1956. *Meaning and Necessity* (2nd edn). Chicago: University of Chicago Press.

Dowty, D. R., Wall, R. E., and Peters, S. 1981. *Introduction to Montague Semantics*. Dordrecht: Kluwer.

Fodor, J. A., and Pylyshyn, Z. W. 2015. *Minds without Meanings: An Essay on the Content of Concepts*. Cambridge, MA: The MIT Press.

Haslanger, S. 2012a. Gender and Race: (What) Are They? (What) Do We Want Them to Be? (pp. 221–47). *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Haslanger, S. 2012b. Language, Politics, and 'The Folk': Looking for 'The Meaning' of 'Race' (pp. 429–45). *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Haslanger, S. 2012c. What Good Are Our Intuitions? Philosophical Analysis and Social Kinds. (pp. 381–405). *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Jackman, H. 1999. We Live Forwards but Understand Backwards: Linguistic Practices and Future Behavior. *Pacific Philosophical Quarterly* 80:157–77.

Jackman, H. 2005. Temporal Externalism, Deference, and Our Ordinary Linguistic Practice. *Pacific Philosophical Quarterly* 86:365–80.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. New York: Oxford University Press.

Kripke, S. A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Montague, R. 1974. English as a Formal Language. *Formal Philosophy* (pp. 188–221). New Haven: Yale University Press.

Plunkett, D., and Sundell, T. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Quine, W. V. O. 1953. Two Dogmas of Empiricism. *From a Logical Point of View* (pp. 20–46). Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1966a. Carnap and Logical Truth (pp. 100–25). *The Ways of Paradox and Other Essays*. New York: Random House.

Quine, W. V. O. 1966b. Truth by Convention (pp. 70–99). *The Ways of Paradox and Other Essays*. New York: Random House.

Saul, J. 2006. Philosophical Analysis and Social Kinds: Gender and Race. *Proceedings of the Aristotelian Society* 106:119–43.

Scharp, K. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Schroeter, F., and Schroeter, L. 2009. A Third Way in Metaethics. *Nous* 43:1–30.

Sundell, T. 2011. Disagreements about Taste. *Philosophical Studies* 155 (2):267–88.

Sundell, T. 2012. Disagreement, Error, and an Alternative to Reference Magnetism. *Australasian Journal of Philosophy* 90 (4):743–59.

Thomasson, A. L. 2016. Metaphysical Disputes and Metalinguistic Negotiation. *Analytic Philosophy* 57 (3):1–28.

Williamson, T. 2006. Conceptual Truth. *Aristotelian Society Supplementary Volume* 80:1–41.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

3

Minimal Substantivity

Delia Belleri

1. Introduction

Ontological debates may be characterized as disagreements about what there is. Philosophers debate on whether there are numbers, propositions, universals, species, genders, and much more. Two of these ontological disputes have sparked significant controversy in recent years. The first dispute revolves around questions of *persistence*, asking how an object can exist through time, undergoing changes while remaining the same entity. An influential account, called Perdurantism, has it that objects persist through time by having temporal parts (Armstrong 1980; Lewis 1986; Heller 1990; Sider 1997). Endurantism, by contrast, rejects the existence of temporal parts, holding that objects are wholly present at each moment of their existence (Haslanger 1989; Van Inwagen 1990; Merricks 1994; Wiggins 2001; Fine 2006). The second debate is about *composite material objects* and it revolves around the question whether these objects exist and, if they do, which ones exist. Nihilists hold that composition never occurs, and that only simples exist (Unger 1979; Wheeler 1979; Dorr 2005). Universalists, by contrast, hold that it is sufficient for two objects to exist in order for them to compose something (Lewis 1986; Van Cleve 1986; Armstrong 1997; Rea 1998). Finally, moderatists argue that composition obtains in some cases and not in others: for instance, only when the parts concur together to form a life, or when they are otherwise connected (Van Inwagen 1990; Merricks 2001; Markosian 2014; Carmichael 2015).

I will refer to these debates as *the Persistence debate* and *the Composition debate* respectively. Also, although for reasons of space my examples in this chapter will draw only on the Persistence debate, I submit what I say could apply, *modulo* relevant distinctions, to the Composition debate too (I will briefly return to it in section 6).

In metaontology, diverging views about the Composition and Persistence debates have emerged. *Ontological Deflationism* holds that these disputes are not substantive. Some deflationists argue that these are merely terminological discussions, to be resolved by reaching an agreement about which linguistic expressions belong to plain English (Hirsch 2005, 2009). Others argue that these disputes involve questions that are either easily answered within a language, or are otherwise unanswerable because asked outside of a language (Thomasson 2009, 2015). Other deflationists yet hold that the competing views are all somehow equivalent (Sidelle 2002; Miller 2005), where this might be taken to entail that we have too little evidence to choose one

Delia Belleri, *Minimal Substantivity* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Delia Belleri.
DOI: 10.1093/oso/9780198801856.003.0003

ontological view rather than another (see especially Bennett 2009). The opposite approach may be labelled *Ontological Anti-Deflationism*, according to which these ontological disputes are substantive. A *pragmatist-naturalist* Anti-Deflationism along the lines dictated by Quine (1948, 1951, 1960) will presumably say that the Composition and Persistence disputes are substantive insofar as one of the two options should be recognized as more compatible with our best scientific description of the world. More radically, a hardcore *realist* along the lines of Sider (2011) will say that these disputes are substantive insofar as one of the competing options carves reality at its joints better than the other.

It is not clear that either of these views satisfactorily accounts for the status of the Persistence and Composition debates. On the one hand, the deflationist dismissal does not do justice to the following facts: first, that opting for either of these theories implies embracing certain commitments about what *there is,* thus assuming a position about non-linguistic, worldly facts; second, that each party to the debate can cite evidence that is sufficient at least by their *own* lights. On the other hand, embracing a full-blown Anti-Deflationism may not help either, and the Endurantism-Perdurantism debate nicely illustrates this point. First, a naturalist option may not be viable, for it is not clear that, for example, Perdurantism better fits with our current science. Indeed Sider (2001: 80ff) and Miller (2005: 110—13) precisely show that Endurantism can be formulated in a way that is compatible with special relativity, and is therefore not inferior to Perdurantism in this respect. Second, a realist position would be supremely difficult to establish too, since it remains unclear (at least to me) whether it can be successfully argued that there is conclusive, convincing evidence as to which option is more joint-carving. With respect to the existence of temporal parts, indeed Sider admits that he has "no good epistemology of metaphysics to offer" (Sider 2001: xv).

My project in this chapter is that of articulating a position intermediate between Deflationism and Anti-Deflationism. Specifically, I wish to show that a *Minimal Anti-Deflationism* about these debates is defensible, whereby the Composition and Persistence disputes are *minimally substantive.* As I will try to outline it, minimal substantivity is incompatible with kinds of defectiveness like mere verbality and lack of sufficient evidence; however, at the same time it implies no naturalistic or realistic commitments. As such, the notion of minimal substantivity could help restore the ontological respectability of the Composition and Persistence debates, with no need to submit oneself to much more committal views about what it is for a dispute to be ontologically substantive.

Before proceeding, I should say something concerning the relationship between the notion of minimal substantivity that I am about to outline and the main theme of the present volume—namely, what Burgess and Plunkett (2013) call *conceptual ethics* and what Cappelen (2018) calls *conceptual engineering.* I believe the notion of minimal substantivity helps explain how and why some debates that are (either overtly or covertly) about the engineering of some concept are ontologically relevant, even when these kinds of endeavours do not aim at identifying concepts that are naturalistically adequate, or let alone joint-carving.[1]

[1] For a proposal alternative to mine, see Díaz-León (Chapter 9, this volume).

The engineering of concepts like Gender, Race, Woman, Black, Marriage may belong precisely to this category. Indeed, as Barnes (2014: 337–8) points out, the philosopher may even assume from the start that these concepts track social constructions, where this by no means entails either (i) fundamental joint-carving; or (ii) reduction to facts described by the natural sciences.[2] If these enterprises are to be accepted as still ontologically substantive, the relevant notion of substantivity needs to be free from the above metaphysical preconceptions. This is where the notion of *minimal substantivity* can help: it can shed light on a way in which these tasks of conceptual engineering can be regarded as ontologically substantive, even when the purpose of the engineering itself is neither joint-carving nor naturalistic adequacy.

Last, the present contribution may also itself be viewed as an exercise in the engineering of the concept Minimal Substantivity. I realize this concept is far from belonging to our ordinary repertoire; consequently, I see myself as trying to articulate a novel notion rather than as trying to ameliorate a pre-existing one. My hope is that of offering a useful conceptual tool, which will help trace finer-grained distinctions and sharpen our philosophical understanding of the status of some disputes about what there is.

2. The Minimum Requirement for Substantivity

As already mentioned in the Introduction, I believe a *Minimal Anti-Deflationism* is available in logical space: Ontological disputes are substantive in (at least) a "minimal", yet interesting sense. My goal is to bring out, and elaborate on, this sense.

For purely operative purposes, we might begin our inquiry by considering what it is that we require, at a minimum, from a full-blown substantive dispute. This consideration will guide us in the process of unearthing and delineating a notion of minimal substantivity.

It seems to me the minimum necessary (and by no means sufficient) requirement for a genuinely substantive dispute to obtain is that it is *not verbal*, that is, such that the disputants are not simply talking past each other:

[*Substantivity*] If an ontological dispute is substantive, then it is non-verbal.

With this formulation in mind, let us ask: Do ontological disputes meet this necessary, minimal criterion for substantivity? There is a story, due to Eli Hirsch, according to which they do not. Hirsch's view has it that a verbal dispute is such iff each side ought to agree that, on the most plausibly charitable interpretation, the other side speaks the truth in its own favoured language (cf. Hirsch 2005, 2008a,b, 2009, 2011).[3]

Let us examine what Hirsch would say with regard to an ontology dispute such as that between Endurantism and Perdurantism. Suppose that Emma the endurantist

[2] It could be pointed out that social constructions are the subject of *social* sciences, which count as "sciences" in the Quinean methodology. Even conceding this, it would still seem possible to deem these questions as substantive in a lesser sense. The present proposal offers a way of bypassing such issues of ranking.

[3] Other influential accounts of verbal disputes are offered by Chalmers (2011); Jenkins (2014); and Balcerak-Jackson (2014).

utters: “There is a tree in front of us”, while Percy the perdurantist rebuts her claim: “No, there is a sequence of temporal parts of a tree”. In order for Emma to regard what Percy says as true, charity requires that she interprets him as speaking a perdurantist version of English (call it P-English), where “there is” ranges over a domain of objects that includes temporal parts (this is very rough[4] but details can be glossed over for present purposes). Symmetrically, Percy should interpret Emma as speaking an endurantist version of English (E-English) where “there is” ranges over ordinary objects, but not over temporal parts. If this happens, each speaker is uttering something true in their language, but they are not contradicting each other: the dispute is merely verbal.

One could question that charity considerations should demand that each side regards the other as saying something *true*. After all, regarding the other side as saying something *reasonable* (although, for all one knows, false) would seem enough in the way of honouring charity, and it would not require concluding that the adversary is speaking a different language.[5] So, the proponent of the Hirsch strategy owes an explanation as to why they would so heavily insist on attributing to the interlocutor true, rather than merely rational beliefs.

I think the following argument could be invoked to this end: in general, charity enjoins us to attribute to our interlocutors the best doxastic state they could find themselves in given the information available to us.[6] Unless there are defeating considerations, the best such doxastic state is one that involves a reasonable, true belief. Therefore, if the subject is in a position to regard an interlocutor’s belief as reasonable and true (and Emma and Percy are in such position), and instead regards it as merely reasonable, but for all one knows also false, one is not being fully charitable. If this argument is sound, and unless ontological disputes are shown to give rise to defeating considerations,[7] then assuming Emma and Percy wish to be fully charitable, they should consider each other as speaking truly in different

[4] There is a complication here: Emma rejects the existence of temporal parts, so it would be incoherent for her to believe that the quantifier ranges over temporal parts. Addressing this point, Hirsch (2009: 249—51) seems to allow that Emma fails to subscribe to a Tarski semantics for P-English, even though she subscribes to a Tarski semantics for E-English. More accurately, Emma can believe that in her own language (E-English), truth depends on reference, but need not accept that this holds *for every possible language* (including P-English). So Emma can consistently believe that in E-English, the truth of “Fa” depends on the singular term referring to some object—which cannot be a temporal part—but need not accept that, in P-English, the truth of “Fa” depends on the singular term referring to an object which could be a temporal part. She can simply believe that the truth of “Fa” in P-English depends on such-and-such being the case, where “such-and-such” is a description issued in E-English.

[5] Balcerak-Jackson suggests (2012: 17–18) that the endurantist could regard the opponent’s statement as false in ordinary English because based on a conceptual error, however excusably so, because the matter under discussion is complex enough that the speaker might have committed a mistake along the way without losing her rationality (for a similar point, see Horden 2014: 237–8).

[6] I take this formulation to be broad enough as to accommodate various ways of spelling out the principle of charity, as formulated by, among others, Quine, Grandy, Davidson, and Lewis, as well as Hirsch himself.

[7] It could be argued that, if the cost of regarding the opponent as having a true and reasonable belief is positing a linguistic difference, this cost constitutes a defeating consideration. But it is not obvious to me that the cost in the case of Emma and Percy is *so high* as to defeat the requirement for full charity. The issue is at best debatable.

versions of English, where "there is" has different senses. Therefore, ontological disputes seem doomed to be diagnosed as verbal.

In light of these considerations, what could be done in order to save the non-verbality of ontological disputes? One option is to "move one level up", so to speak: that is to say, to regard the dispute as taking place at the metalinguistic level, and thus regard the parties as advocating each a different "idiolect" as the idiolect that is best to use in that philosophical context. Hirsch seems to hint at similar considerations. Addressing Sider's idea that ontologists can speak Ontologese, Hirsch writes:

> My advice is that [the philosophers who purport to speak Ontologese] should *stick to the meta-level and engage in disputes about which sentences are true in the philosophically best language*, rather than attempting to speak that best language. (2008b: 520, my italics)

In order to secure a non-verbal conflict, moving the dispute at the metalinguistic level would seem to be a viable option. Indeed, this suggestion gains support from some recent developments in the meta-philosophical literature. David Plunkett and Timothy Sundell (cf. Sundell 2011; Plunkett and Sundell 2013; Plunkett 2015) have recently maintained that a great number of philosophical disputes which look merely terminological are instead covertly metalinguistic and concern "a distinctive normative question—how best to use a word relative to a context" (2013: 3). Plunkett and Sundell contend that this metalinguistic, normative question is not overtly expressed, but it is pragmatically communicated, so that the metalinguistic character of the exchange may not be immediately recognizable. Following Plunkett and Sundell's lead, more than one author has proposed to interpret at least some philosophical disputes as being covertly metalinguistic: Megan Wallace (MS) focuses on ontological debates about ordinary material objects, while Amie Thomasson (2016) considers debates about composition, essence, identity, and persistence, but also disputes about what art is, about free will and determinism, and about the existence of races. In light of what these authors propose, Hirsch's suggestion could be honoured by envisaging the possibility that these disputes be *explicitly formulated* in metalinguistic terms, thus moving to what Plunkett and Sundell (2013: 6–7) and Plunkett (2015: 836) would call a "canonical dispute", centred on a literally expressed disagreement about a metalinguistic, normative question.

Going back to the exchange between Emma the endurantist and Percy the perdurantist, the friend of metalinguistic disputes could say that, although the two disputants would seem engaged in a first-order communication about whether or not there are temporal parts, what they are really doing is negotiating the sense of "there is". This opens up the possibility of making the metalinguistic conflict explicit: In the newly conceived dispute, Emma would be overtly advocating an endurantist (commonsensical) sense of the term, which allows one to existentially quantify over trees, tables, and mountains but not over temporal parts; and Percy would be overtly advocating a perdurantist sense of "there is", which permits existential quantification over all these objects *plus* their temporal parts.

Let us therefore assume that it is possible to recast first-level ontological disputes as metalinguistic debates. This move has a problematic consequence which needs to be dealt with: once the disagreement has moved at the metalinguistic level, it *is not* *ontological any longer*: it is about language, or at least about which semantics for a given language should be chosen. This seems to have little to do with ontological substantivity! What should we say, then, in order to preserve the idea that these are ontologically interesting disputes—at least to a minimal degree?

3. Rescuing the Ontological Significance of Metalinguistic Disputes

In this section, I aim to show that, even if the only non-verbal disputes on matters of persistence (or composition) were metalinguistic ones, not all hopes are lost for those who wish to rescue the "minimal substantivity" of these disputes. This is because the selection of certain linguistic options would seem to enjoin a certain degree of *ontological commitment* once the choice is made. Discussing the best linguistic resources for doing ontology may therefore have downstream ontological implications, to the extent that it may lead us to underwrite certain ontological commitments.

3.1. First-Order Issues and Ontological Commitments

Maintaining the focus on persistence debates, in what follows I wish to show that some central arguments that have been proposed by supporters of perdurantism can be reformulated as metalinguistic proposals to introduce some piece of terminology, for example "temporal part", where this should sanction an ontological commitment to temporal parts. These newly reformulated arguments may be regarded as starting with problems that arise at the first order, that is, where language is *used* to talk about the existence of objects and their properties through time, and as advancing a metalinguistic proposal so as to obviate to these first-order problems. The purpose of these reconstructions is demonstrative: my aim is to show that it is possible to look at specific arguments offered in the literature in the material mode, as it were, in a new metalinguistic light. The reader who already got the gist of the strategy may safely skip the examples.

To start with, we may choose the problem of temporary intrinsics. The first-order problematic statement that needs philosophical consideration may be rendered as follows: "If a certain object *o* changes by being *F* at a certain moment *t* and subsequently being *not-F* at $t+1$, then *o* has incompatible properties at different times; if this is so, then *o* at *t* is not identical with *o* at $t+1$." David Lewis' (1986) proposal, at the metalinguistic level, could be phrased as follows: let us introduce the notion of "temporal part". Once we do this, our way of talking at the object-level changes: it is no longer one and the same object *o* to which we ascribe the incompatible properties of being *F* and being *not-F*; it is two numerically distinct temporal parts of *o*, call them *temp-1* and *temp-2*.

As a second example, we may consider the case of arbitrary undetached parts. At the object level, the predicament to be dealt with may be thus expressed: "If Descartes is deprived of a leg at *t*, he then becomes (say) Descartes-minus. Now, it seems that Descartes-minus before *t* is the same as Descartes-minus after *t*; also, Descartes after *t* is the same as Descartes-minus after *t*; and also, that Descartes is the same before and after *t*; it would then follow that Descartes-minus before *t* is the same as Descartes before *t*. But this is *not* the case." The solution proposed by Heller (1990) could be stated as follows at the metalinguistic level. If we introduce the term "temporal part", then at the object-language level there is no tension between the following statements: "The temporal part of Descartes-minus after *t* and the temporal part of Descartes after *t* are the same"; and "The temporal part of Descartes-minus before *t* and the temporal part of Descartes before *t* are *not* the same".

As a last example, let us pick the argument from vagueness in defence of Perdurantism. Here is the problematic inference that needs philosophical attention. If objects gradually go out of existence, it is vague whether they form diachronic units—that is, units that start and cease to exist. This is unacceptable, for it would entail that existence is vague, while existence as expressed through the logical vocabulary is not vague. Sider (1997, 2001) offers a response which, at the metalinguistic level, could look like this: if we introduce the notion of "temporal part" and if we subscribe to unrestricted composition, we avoid vague existence, and consequently also indeterminate statements at the object-language level. For now whatever diachronic sum of temporal parts we consider is a genuine composite object with clear-cut temporal boundaries.

It is important to note that, at least for the advocates of Perdurantism, acceptance of the term introduced seems to imply an *ontological commitment* to temporal parts,[8] at least in the sense of supposing that temporal parts are in the domain of one's existential quantifier (Quine 1948: 32; 1960: 242). For it would be odd to just *talk* in terms of temporal parts, while at the same time denying that *there are* temporal parts. The best way for avoiding ontological commitment to temporal parts would be to expunge the term "temporal part" from one's terminology (e.g., by paraphrasing it away); but since the proponents of Perdurantism do exactly the opposite, it seems safe to say that term introduction also marks (and should mark) an ontological commitment to temporal parts.

To be sure, the perdurantist could introduce talk of temporal parts as a useful *fiction*, where this would indeed imply just speaking in terms of temporal parts while denying that there are any temporal parts. Although this option is available, this is importantly not what authors who identify as perdurantists either do or should do, if anything because the ensuing position could not ultimately count as a form of *bona fide* Perdurantism, but would rather count as a form of nominalism about temporal parts. Otherwise put, if Perdurantists wish to hold a position that they could legitimately describe as *realism* (of some form), a fictionalist move would simply be a non-starter.[9]

[8] It is perhaps not fully accurate to say that the whole question at stake in a metalinguistic ontological dispute on "temporal part" is *just* whether or not we should introduce the term "temporal part". For if this were the case, then it seems that by merely conceding that one *could* introduce the term "temporal part", the endurantist would be committed to temporal parts. To avoid this result, the question at stake should be made more precise, for instance: "Should one introduce the term «temporal part», where this implies that one wishes to quantify over certain entities, thereby ontologically committing to them?" (Thanks to Richard Woodward for stressing this point.)

[9] There are also motivation issues. Notice the contrast between a version of Perdurantism that expects to be fictionalist from the outset and standard fictionalist proposals in other domains. Standard

There is an additional element regarding the commitments following from a certain linguistic proposal. When two metalinguistic proposals enter into conflict, it seems reasonable to expect that the proposal which eventually prevails in the dispute (if any) will imply some ontological commitments for *all* the disputants. Thus for instance, were the perdurantist to "win" the dispute, this would imply a commitment for all parties to use "there is" in such a way as to quantify over temporal parts. This would have various implications about which statements are evaluated as true or false among the parties: for instance, a statement of "Yesterday's temporal part of this tree is not the same as today's temporal part" would now be considered as true. By contrast, were the endurantist to "win" the dispute, it seems that the existential quantifier would have to be used by all parties in such a way as to not quantify over temporal parts, and statements like "This tree is the same as yesterday's tree" would consequently be deemed true.

It therefore seems that the ontological import of the metalinguistic dispute comes from the joint contribution of two factors. The first is that introducing certain linguistic expressions (like "temporal part") would seem to imply an ontological commitment to certain objects for the proponents of the linguistic introduction. The second is that a resolution of the dispute in favour of one option or another would seem capable to affect the ontological commitments of all the parties to the dispute. If these considerations are sound, we now seem to have a first-pass proposal regarding a sufficient condition for the obtaining of minimal substantivity in ontological disputes:

> [*Minimal Substantivity-1*] An ontological dispute is minimally substantive if it is linked with a metalinguistic, non-verbal dispute whose resolution can affect the parties' ontological commitments.

It should be stressed that being ontologically committed to, say, temporal parts need not mean believing, for instance, that "temporal part" refers to temporal parts in any robust sense, or that temporal parts are objects in any inflated, heavyweight sense. There are ways of conceiving reference and objecthood that are sufficiently deflationary so as to guarantee that ontological commitment stays suitably deflationary too.

For instance, one may have a sufficiently broad notion of reference which not only implies that it is a language-world relation, but also that it could obtain when certain *intra-language* relations obtain. According to this less demanding notion, reference need not obtain only when an object is picked out by a certain term *t*, but it could also obtain when the sentence *S* containing a certain term *t* is considered equivalent to

hermeneutical Fictionalism, for example, concerns linguistic practices that are *already in place*, from talk about numbers (Yablo 2002: 87) to talk about ordinary objects (Rosen and Dorr 2002: 171). An important motivation for the view seems to be that since we cannot get rid of these already established practices, we need a way of salvaging as much as possible of a set of linguistic usages that are otherwise regarded as metaphysically and semantically confused. The practice proposed by a fictionalist perdurantist would, however, be a *new* one. The purpose of legitimizing a defective practice that we cannot get rid of would be utterly absent. This seems to deprive the fictionalist perdurantist of an important motivation that is commonly associated with standard forms of fictionalism. One could then wonder what the motivation for going fictionalist from the outset could be: why not be a *conventionalist*, or a *deflationist realist*, about temporal parts? (Thanks to Alexis Burgess and David Plunkett for prompting me to clarify these issues).

another sentence S^* that does not contain t and that is deemed true. Following this strategy, the perdurantist might say that, if the sentence: "An object o is F at time t_1" is true, the sentence: "The t_1-temporal part of o is F" is true too and hence, the term "temporal part" refers. The latter proposal would amount to a deflationary approach to the reference of terms like "temporal part" along Neo-Fregean lines (cf. Hale and Wright 2001, 2009).[10]

Analogously, one need not believe that temporal parts are objects in any robust sense—for example, instantaneous or even just very short-lived spatio-temporal objects.[11] Indeed, the perdurantist may presume right from the start that temporal parts are to be considered objects in a more deflationary sense. For instance, one could say that a temporal part is an object in the "covering" sense of the term individuated by Thomasson (2007, 2009, 2015), in that from the sentence "Today's temporal part of this tree is blossoming" it is possible to infer "Something is blossoming" where "something" plays the role of a dummy sortal.

Of course, whether or not the perdurantist will be happy with this minimal set-up will depend on her stance with relation to what reference and objecthood amount to. However, it seems entirely possible and not *ad hoc* to opt for a combination of ontological commitment *plus* deflationary reference and objecthood, and still be considered as somebody who engages in a dispute that has some ontological significance. Otherwise put: it seems possible, and not *ad hoc* or let alone inconsistent, for someone to accept: (i) that the dispute about temporal parts has a minimal ontological significance insofar as it concerns the selection of a language from which certain ontological commitments are to be extracted; and (ii) that reference to temporal parts themselves, or that the kind of objects they are, are construed in a deflationary fashion.

So far, I have argued in what sense metalinguistic proposals—such as that of introducing the term "temporal part"—may have an ontological significance, for they might create an *ontological commitment* as a result of the dispute. The ontological commitment can be formulated in a sense that is (i) minimal enough as to exclude robust views of reference and objecthood; but still (ii) sufficient for ensuring that the proposal has *some* ontological import.

3.2. Resisting Two Attempts at Downplaying the Proposal

One way in which this proposal may be downplayed is by arguing as follows: the metalinguistic negotiations considered so far are disagreements that only obtain between philosophers. Some of the competing linguistic options could at best imply ontological commitments for a few, initiated ontologists (e.g., the commitment

[10] It is also possible to follow other deflationary accounts of reference: Horwich's (1998) view would have it that the term refers so long as the meaning of "temporal part" is constituted by certain use-features, and the expression occurs in instances of a disquotational schema like: "(x) (<n> refers to x iff x = n)". Field (1994: 261–3), would presumably emphasize the term's computational role; Brandom (1994: 360–70) would stress the expression's inferential role and the substitution patters in which it enters; while Burgess (2015) suggests that, in an inferentialist framework, a term refers if it occurs in a simple atomic truth.

[11] One could have qualms with instantaneous objects (as noted by Fine 2006: 700), or with short-lived temporal parts (like my temporal part between October 1, 2016, and October 2, 2016), on account that physical objects like these could not "pop in and out of existence".

to temporal parts). Given what Eli Hirsch calls “the lack of authority that philosophers have in our culture” (Hirsch 2008a: 181), it seems that some of the competing options could not make it out of the “ontology room” and be absorbed into ordinary usage. This seems to significantly reduce the interest of minimal substantivity, because even if the dispute produced a change in ontological commitments, these may well remain confined within the narrow conversational setting of a few academic specialists. In response to this downplaying attempt, let me try to more precisely pin down the connection between the philosophers’ (metalinguistic) dispute and the broader ontological “language games” practised in ordinary language.

To start with, it seems uncontentious that, when we think and talk about what there is—in *both* ordinary life and in philosophy—we engage in “language games”[12] which feature certain types of expressions and certain rules about how to use those expressions. These linguistic resources mainly consist of the quantificational apparatus, the numerals, the identity sign, *plus* nouns, predicates, and variables. There is therefore an undeniable continuity between the linguistic resources employed by the ontologist and the linguistic resources employed in order to talk about what there is in ordinary contexts.

Now, the language game of talking about the existence of temporal parts is one that almost exclusively philosophers engage in; however, nothing prevents that, *at least in principle*, the linguistic usages established by the philosophical discussion “leak” into (some fragments of) the ordinary language, thus affecting the ontological commitments and interpretations of the relevant key vocabulary. Note that this does not require us to imagine a world where philosophers have authority over ordinary linguistic usages; it only requires to imagine a world where the uses in the ontology room *causally contribute* to a change (or consolidation, where applicable) of ordinary language games about what there is. This seems like a much less demanding scenario than the one Hirsch seems to expect to obtain. As to the contention that the notion of minimal substantivity is not interesting enough because the link with ordinary linguistic practices is too feeble, I will not pursue the issue further, partly because it is not clear what “interesting” in this context might exactly mean, partly because it seems to me obvious that there is at least *one* sense in which this notion is interesting enough, given appropriate aims and purposes.

There is another way in which [*Minimal Substantivity-1*] might be downplayed, this time linked to a specifically (Neo)-Carnapian way of portraying the role of language in ontology. The potential worry could be illustrated with an example drawn from Amie Thomasson’s work. Thomasson finds it objectionable that there may be more than one sense of the existential quantifier—a joint-carving sense in addition to its standard, first-order predicate logic sense (conferred by introduction and elimination rules). If the existential quantifier only has one sense, then any sentence purportedly formulated with the other sense of the quantifier would be semantically defective, either because it has no clear meaning or because it is straightforwardly false (Thomasson 2015: 317; 2016: 8–9). So, even if the dispute

[12] The expression “language games” is simply meant to designate rule-governed practices of language use; it should not be understood as a way of endorsing Wittgensteinian ideas on meaning or rule-following.

eventually led (or seemed to have led) to a change in ontological commitments, the only parties uttering semantically acceptable sentences at the first-level would be the parties that stick to the standard semantics of the quantifier. The parties deviating from such standard semantics would be committing a form of semantic mistake. In this case, any such dispute would be a disappointment: for only one party could ever be right, because only one party would be upholding the semantics of English.

If the focus is restricted to the semantics of the existential quantifier, Thomasson may have her reasons to look unfavourably at the notion of minimal substantivity, which I will not discuss here. However, she need not reject that metalinguistic negotiation might concern other kinds of expressions, for example nouns like "table", "number", "person", "corporation", or "marriage". In particular, she may well accept that metalinguistic negotiation is sometimes needed in order to clarify or precisely determine the conditions of application of a certain term, or to change them so as to overcome a number of semantic flaws. In Thomasson's own words:

> Conceptual work needn't be simply explicative:...at times we may have work to do to determine how best to fill in the details of our concept of 'same person' or 'same work of art', consistent with some (ethical, aesthetic, or pragmatic) purpose. Conceptual work is also involved in determining whether tacit contradictions or incoherencies beset parts of our conceptual scheme.... Ontologists may also be engaged in what Carnap would have called 'conceptual engineering': revising or devising systems of categories to help them better serve some practical purpose. (Thomasson 2015: 327–8)

It is compatible with this picture that, sometimes, a disagreement may arise concerning the rules of application of a certain term; or concerning whether or not a new term (with application rules to be fully worked out) should be introduced in the linguistic framework used to talk about matters ontological. If ontological commitments would flow from the adoption of these terms—for instance via analytic entailments, in accordance with her own account (Thomasson 2007: 167; 2015: 145–58)—then it seems that if the dispute could lead to a change or establishment of ontological commitments, the dispute would be minimally substantive after all, even in Thomasson's framework. Therefore, setting aside her reservations on the semantics of the existential quantifier, minimal substantivity turns out to be compatible with the set-up proposed by Thomasson, at least for some expressions.

To be sure, Thomasson may still not accept that the ontological commitments extracted from a certain framework be understood in terms of "robust" notions of existence, reference, and object; they would rather have to be understood along deflationary lines. Still, it seems that [*Minimal-Substantivity-1*] is sufficiently schematic and independent of considerations of ontological robustness in order for it to be effortlessly imported into her Neo-Carnapian account. As it turns out, then, Thomasson's views need not be hostile to the notion of minimal ontological substantivity.

In general, the schematic character of [*Minimal-Substantivity-1*] makes it compatible with several ways in which ontological commitment could be cashed out, ranging from "robust" to "deflationary". We could therefore say that [*Minimal-Substantivity-1*] is compatible with *various ways of being a realist*, ranging from hard-line approaches whereby certain objects or facts are taken as ontologically "thick" and wholly language- or mind-independent, to minimal approaches where their existence is "thin" and often tied to semantic considerations. Such approaches can be held not only with regard to material objects, temporal parts or mathematical entities, but also with regard to—for example—moral properties and facts, and in general in normative domains where realism can assume the nuances just outlined. Meta-ethics provides an example here, for one can either be a robust realist about moral properties and facts; or one can be a "minimal realist" who confines herself to claiming that moral discourse is truth-apt and by-and-large true, where this gives rise to no inflationary views about the nature of moral properties or facts. As Gideon Rosen puts it, "the minimal realist holds that in a thin and metaphysically unambitious sense, the doctrine correctly represents . . . a genuine domain of fact: at least some of the objects the discourse posits really exist, and the corresponding singular terms refer . . . " (Rosen 1994: 281). If what I have argued is right, [*Minimal Substantivity-1*] serves the minimal and non-minimal moral realist equally well.

To conclude this section, I have argued that metalinguistic disputes related to certain ontological matters (like persistence or composition) can be considered as "minimally substantive". The metalinguistic proposals in play (as, for example, that of introducing talk about temporal parts) imply a certain degree, no matter how minimal or deflationary, of ontological commitment, so minimal substantivity has to do with the potential the dispute has to affect the parties' ontological commitments. Two attempts at downplaying minimal substantivity have also been deflected, by (i) pointing out that the dispute could in principle affect the ontological commitments underlying ordinary language games and (ii) showing that the notion of minimal substantivity is compatible with metalinguistic negotiation within a (Neo)-Carnapian set-up.

4. Rescuing the Epistemic Significance of Metalinguistic Disputes

The deflationist could concede that the metalinguistic disputes associated with the Persistence and Composition debates comply with [*Minimal Substantivity-1*]. Yet, she could insist, the disagreement suffers from another kind of defectiveness, one whereby we do not have enough evidence to choose between one linguistic option and the other—for example, between E-English and P-English. This position is dubbed by Karen Bennett (2009: 73) *Epistemicism.*

I will now reconstruct a potential epistemicist argument that may apply to the metalinguistic negotiation between P-English and E-English (Bennett's version concerns the very theories, like Endurantism and Perdurantism). The epistemicist about the metalinguistic dispute between E-English and P-English could say, first of all, that neither of the two languages seems to guarantee greater simplicity. The perdurantist will obviously have to increase the complexity in her language by adopting the term "temporal part". However, the endurantist will also complicate her language by turning monadic predicates like "being blue" into polyadic predicates like "being blue-tly". Therefore, it is not clear who wins on the front of simplicity. Second, it is not clear that the perdurantist's linguistic innovations help completely and uniformly to solve Endurantism's problems. Recall that the endurantist struggles with co-location, in that her account leaves it open whether, in the Lump-Statue case, there are two objects or just one. The standard worm-theory perdurantist claims to solve the problem by saying that the temporally extended lump and the temporally extended statue share some temporal parts, exactly like two roads can share a stretch of road (Sider 2001: 153). However, we could imagine that the statue and the lump are created and destroyed at the same times (Gibbard 1975: 191), where this revives the puzzle for the perdurantist as well: are there two temporally extended objects or just one (cf. Magidor 2016: 524)? Provided this line of argument could go through, absence of sufficient grounds to select one option as opposed to the other would undermine the significance of the dispute.

In arguing against this epistemic criticism, I will assume that there are two angles from which to look at a dispute: one is the "external" angle, to be identified with the bird-eye perspective of the neutral onlooker. This is the perspective of someone who accepts that, in adjudicating between different languages (like E-English or P-English) that are associated with ontological theories (Endurantism and Perdurantism), we should keep in mind that a number of "theoretical virtues" ought to be honoured—like simplicity or explanatory power. However, one has *no determinate views* as to which virtues should take priority, thus remaining somewhat "neutral" or "open-minded". The other perspective is the "internal" one, to be identified with the point of view of an engaged participant to the dispute. This is the perspective of someone who, for instance, has clear views about which theoretical virtues should take priority when adjudicating between E-English as the language of Endurantism and P-English as the language of Perdurantism. For instance, one may believe that ontological simplicity/parsimony should be maximized even if this meant increasing ideological complexity.

Having distinguished between the "external" and "internal" perspective, one may grant to the epistemicist that, from the "external" point of view, evidence does not favour any of the linguistic options, because they all stand on a par with respect to a number of features. However, it seems to me plausible that from the "internal" perspective, there will often be *some admissible consideration* that appears conclusively to favour one particular choice.[13] For instance, the endurantist may concede that she has to make her predicates more complicated, but also consider that as a bearable cost at least if she is a nominalist about properties. The cost of complicating the predicates may be outweighed by the benefits of not introducing terms that imply commitments to new objects—like temporal parts. Avoidance of a certain object-talk may therefore count as a sufficient reason to adopt E-English, at least *relative* to these theoretical considerations.

Also, all participants to the dispute would seem to be *entitled* to the theoretical considerations they favour, provided these are *admissible* by the lights of the

[13] I am saying "often" and not "always" (or, let alone, "necessarily"), because cases should be allowed where these admissible considerations are lacking, and there is therefore no internal justification. Absent internal justification, the subject would therefore not be fully rational in believing her favoured theory. Presumably, in this case the subject should suspend judgement, or believe with a very low credence.

philosophical community's standards, and are not merely idiosyncratic motivations. So, the perdurantist would seem to be entitled to her preference for the ontologically committal idiom of temporal parts, if, for example, maintaining a monadic conception of temporary intrinsic properties is more important for her; for such a ranking of preferences would count as an admissible consideration. Analogously, the endurantist seems entitled to a more complicated ideology, if she is willing to trade that with a smaller, temporal-parts free ontology; for this too strikes as an admissible consideration. If all parties are entitled to such preferences, then if their views honour them, believing these views will be justified at least relative to the adopted perspective. Their views would therefore enjoy an "internal" type of justification. (A further question, which I cannot settle here, is whether "internal" justification has any interesting relation with truth-conduciveness. In Belleri (2017) I argue that it does).

So far, we have simply argued that the internal perspective is available besides the external one. Now we need to be convinced that the external perspective is not relevant in the assessment of ontological disputes. The epistemicist seems to assume exactly this, so countering this thesis may constitute a strategy for undermining her case. I wish to suggest that recognizing the internal perspective allows a more *charitable* reconstruction of the epistemic status of the debate, and should then be preferred to the external perspective. So for instance, countenancing the internal perspective allows one to say that each party to the debate has sufficient justification (*of the internal kind*) to believe their view, while the external perspective would view them as insufficiently justified, *tout court*. Furthermore, according to the internal approach, the considerations each party invokes in favour of their view offer epistemic support to it at least relative to their *own* perspective. By contrast, in the external approach, the reasons invoked by the parties could not count as conclusive even from each party's own point of view. In virtue of the greater charity afforded by the internal approach, and assuming that *ceteris paribus* an account that maximizes charity is to be preferred, I then conclude we should deem the internal approach more relevant than the external approach.

The upshot, then, is that even if metalinguistic ontological disputes were faced with the epistemicist critique, as long as each of the linguistic options could be seen as internally, conclusively justified, these metalinguistic disputes *would not be completely epistemically deflatable*, in that it would not be possible to claim that we lack sufficient justification *tout court* to believe any side.

In light of the foregoing considerations, we should qualify the sufficient condition spelled out in [*Minimal Substantivity-1*] by making it a conjunction of two conditions, one about the dispute's potential for commitment-change and the other about the non-epistemic deflatability of the dispute.

> [*Minimal Substantivity-2*] An ontological dispute is minimally substantive if (a) it is linked with a metalinguistic, non-verbal dispute whose resolution can affect the parties' ontological commitments; and (b) the latter dispute is not epistemically deflatable.

5. Dispelling a Threat of Excessive Proliferation

It could be objected that [*Minimal Substantivity-2*] generates way too many minimally substantive ontological disputes. Suppose Carla argues that there are such things as "Schmartinis", which come into existence whenever an alcoholic drink is served in a V-shaped glass, while Farida rejects this existential claim. What should we say about this case?[14]

Suppose for the sake of the argument that there are sufficient grounds for believing that Carla and Farida are having a metalinguistic dispute. At first sight, it may seem that the objector is right: in principle, this dispute could affect the parties' ontological commitments, and Carla could be internally justified, if she provided admissible (not excessively idiosyncratic) considerations to which justification could be relativized. The dispute would then be minimally substantive: a rather unpalatable consequence.

The problem with this objection is that it trades on an *underdescribed* scenario. What we learn from the Schmartini case is that it seems wrong to call "minimally substantive" a dispute that, because of the *very few* details offered by the objector, seems to arise out of nowhere, seems not motivated by the need to solve a specific problem, and where no immediately recognizable valuable consequence seems to result from the prevailing of any of the options at stake. However, it seems obvious that all these elements *should* normally be specified. So it seems disingenuous to base the objection on an abnormally underdescribed case and not on a normally, adequately described case, in which it would indeed be possible to appreciate why, for example, Carla and Farida are having the Schmartini dispute in the first place, what problem they are trying to solve, and what benefits they pursue with their respective claims. If all these elements were specified, I am confident that the verdict as to the dispute's minimal substantivity would greatly differ.

Of course, it is open to the critic to find an example in which all the details are specified and still it is counter-intuitive to call the exchange minimally substantive. My expectation would be that (i) either the specified details make the dispute so silly and idiosyncratic that (at least) one of the two conditions contained in [*Minimal Substantivity-2*] is not met; or (ii) the specified details make the dispute reasonable enough as to turn out as minimally substantive.

The overgeneration charge contained in the Schmartini objection therefore reveals what seems to be a *background condition* that has to be satisfied in order for the elements spelled out in [*Minimal Substantivity-2*] to suffice for minimal substantivity: that the dispute occurs in a sufficiently rich context, where the rationale, aims, and prospective benefits of the dispute can be clearly identified. Luckily for us, we can specify all of these ingredients in the case of ontological disputes, as well as in countless other philosophical and non-philosophical cases.

[14] Thanks to Esa Díaz-León for discussion on this point.

6. Conclusion: Minimal Anti-Deflationism

The characterization encoded in [*Minimal Substantivity-2*] could be adopted to spell out a minimal form of Anti-Deflationism about some, if not all, ontological disputes. The position is formulated in the following way:

> [*Minimal Anti-Deflationism*] Some (if not all) ontological disputes are minimally substantive, because even if the dispute at the first-level is verbal, it can be linked to an explicitly metalinguistic dispute such that: (a) it could affect the parties' ontological commitments; and (b) it is not epistemically deflatable.

As I have shown, as long as there is sufficient evidence that the Endurantism-Perdurantism debate is a metalinguistic negotiation in the sense coined by Plunkett and Sundell (2013), we could move from an implicit, merely pragmatically conveyed negotiation to an explicit disagreement, which Plunkett and Sundell would call a "canonical dispute", about whether or not to employ certain linguistic items. If it is true (a) that this dispute could in principle affect our ontological commitments; and (b) that it is not epistemically deflatable, then the dispute is minimally substantive.

As hinted at in the Introduction, similar remarks apply, I submit, to the dispute between the views competing in the Composition debate, involving such positions as Nihilism, Universalism, Common-Sense ontologies, Organicism, and so on. Although further work will have to be done in order to substantiate this claim, it is already possible to offer a "preview" of how Minimal Anti-Deflationism could be applied to the aforementioned dispute.

To illustrate, consider a supporter of common-sense ontology declaring "There are composite objects (such as chairs, trees, mountains)"; a nihilist replies: "There are no composite objects". Suppose it were conceded that the commonsensical theorist is uttering something true in plain English, where the domain of the existential quantifier ranges over *ordinary objects*; while the domain of quantification of the nihilist were restricted to *simples* (e.g., subatomic particles). The two contenders would therefore be using two different senses of "there is", call them "there is$_c$" and "there is$_n$". Suppose one could successfully argue that this dispute be interpreted as a metalinguistic negotiation, where the parties are pragmatically communicating their advocacy of "there is$_c$" and "there is$_n$" respectively. It would then be easy to recast the dispute as an explicit metalinguistic negotiation, where the commonsensical theorist were proposing to use "there is$_c$" and the nihilist were proposing to use "there is$_n$" when discussing matters of composition.

At this stage, the proponent of Minimal Anti-Deflationism about the Composition debate may argue that this dispute: (a) could affect the parties' ontological commitments concerning parthood and material composition. For instance, were the nihilistic version of English to prevail, this could result in the parties' withdrawing their commitment to the existence of tables and chairs. Additionally, it could be urged that (b) the dispute is not epistemically deflatable, because each party enjoys at least an internal form of justification. Successfully arguing for (a) and (b) would imply that the dispute in question is minimally substantive. Although many details will have to be filled in, this brief outline already shows that a Minimal

Anti-Deflationist project can be sensibly pursued in the Composition case as well as in the Persistence case.

Before closing, we should go back to the question "What can the notion of minimal substantivity do for conceptual engineers?" As I anticipated, the notion of minimal substantivity can help clarify what is ontologically substantive in disputes where parties need not be engaged in "tracking" any fundamental aspect of reality, nor need to pursue a naturalistic project.

For instance, we might say that a debate on whether there are (or there can be) same-sex marriages is ontologically substantive in a relevant, "minimal" sense, *although* marriage-facts are not fundamental or are not naturalizable. Provided we have enough evidence suggesting that the dispute is a metalinguistic negotiation, we may argue that: (a) debating over whether we should use the expression "same-sex marriage" can affect our ontological commitments (to types of marriages) and the language games we play when we speak about what there is and what there is not, especially with regard to marital unions (actually, these disputes have *already* contributed to a massive change in the relevant language games. So the possibility claim is easily ascertained to be true.) As to condition (b): it is arguable that the dispute is not epistemically deflatable either since, *contra* the epistemicist, there is indeed sufficient evidence to favour one linguistic option rather than another. For instance, it seems that numerous arguments drawing on normative considerations, sociological studies, reports, and so on, favour use of the word "marriage" to denote same-sex couples as well as heterosexual ones. Similar considerations arguably apply to several other ontological or metaphysical disputes on matters that are neither "metaphysically fundamental" in Sider's favoured sense nor obviously naturalizable in the Quinean sense, including disputes about GENDER, WOMAN, RACE, BLACK, RAPE, SEXUAL HARASSMENT, and so on.

Last, I should stress that labelling these disputes "minimally substantive" should not be understood in a demeaning sense, but simply as signalling that the dispute is substantive in a way that requires no naturalistic or realistic assumptions. If you do not like the word "minimally", you could say that the dispute is "non-naturalistically ontologically substantive" or that it is "ontologically-committing substantive". In general, the idea of minimal substantivity should be compatible with a dispute being highly important and worth pursuing for cultural, civil, ethical, or political reasons. It seems like the notion I have been delineating can fulfil this task.

Acknowledgements

This chapter was presented at the Conceptual Engineering Workshop held at the University of Oslo in June 2016. I am grateful to the organizers, Alexis Burgess, Herman Cappelen, and David Plunkett, for involving me in the project, as well as to the audience for their helpful comments, especially Tim Sundell, Patrick Greenough, David Braddon-Mitchell, Esa Díaz-León, Torfinn Huvenes, Matti Eklund, and Derek Ball. Thanks also to Julia Zakkou, Richard Woodward, David Plunkett, and Alexis Burgess for their feedback on the final draft. Work on the final version of this chapter was funded by the Humboldt Foundation, which allowed me to collaborate with the DFG Emmy Noether Research Group "Ontology After Quine" (WO-1896/1-1) led by Richard Woodward at the University of Hamburg.

References

Armstrong, D. M. 1980. Identity Through Time. In Peter van Inwagen (ed.), *Time and Cause: Essays Presented to Richard Taylor* (pp. 67–78). Dordrecht: Riedel.

Armstrong, D. M. 1997. A World of States of Affairs. *Philosophical Perspectives* 7 (3):429–40.

Balcerak-Jackson, B. 2012. Metaphysics, Verbal Disputes and the Limits of Charity. *Philosophy and Phenomenological Research* 86 (2):412–34.

Balcerak Jackson, B. 2014. Verbal Disputes and Substantiveness. *Erkenntnis* 79 (1):31–54.

Barnes, E. 2014. Going Beyond the Fundamental: Feminism in Contemporary Metaphysics. *Proceedings of the Aristotelian Society* 114 (3pt3): 335–51.

Belleri, D. 2017. A Pluralistic Way Out of Epistemic Deflationism about Ontological Disputes. In Annalisa Coliva and Nikolaj Jang Lee Linding Pedersen (eds.), *Epistemic Pluralism* (pp. 317—44). London: Palgrave MacMillan.

Bennett, K. 2009. Composition, Colocation, and Metaontology. In David John Chalmers, David Manley, and Ryan Wasserman (eds.), *Metametaphysics: New Essays on the Foundations of Ontology* (pp. 38—76). New York: Oxford University Press.

Brandom, R. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Burgess, A. 2015. An Inferential Account of Referential Success. In S. Gross, N. Tebben, and M. Williams (eds.), *Meaning Without Representation*. New York: Oxford University Press.

Burgess, A., and Plunkett, D. 2013. Conceptual Ethics I and II. *Philosophy Compass* 8 (12):1091–110.

Cappelen, H. 2018. *Fixing Language: Conceptual Engineering and the Limits of Revision*. Oxford: Oxford University Press.

Carmichael, C. 2015. Toward a Commonsense Answer to the Special Composition Question. *Australasian Journal of Philosophy* 93:475–90.

Chalmers, D. J. 2011. Verbal Disputes. *Philosophical Review* 120 (4):515–66.

Díaz-León, E. (Chapter 9, this volume). Descriptive versus Ameliorative Projects: The Role of Normative Considerations. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Dorr, C. (2005). What We Disagree About When We Disagree About Ontology. In Mark Eli Kalderon (ed.), *Fictionalism in Metaphysics* (pp. 234–86). Oxford: Oxford University Press.

Field, H. 1994. Deflationist Theories of Meaning and Content. *Mind*, New Series 103 (411):249–85.

Fine, K. 2006. In Defense of Three-Dimensionalism. *Journal of Philosophy* CIII (12):699–714.

Gibbard, A. 1975. Contingent Identity. *Journal of Philosophical Logic* 4:187–221.

Hale, B., and Wright, Crispin. 2001. *Reason's Proper Study: Essays Towards a Neo-Fregean Philosophy of Mathematics*. Oxford: Oxford University Press.

Hale, B., and Crispin Wright. 2009. The Metaontology of Abstraction. In David John Chalmers, David Manley, and Ryan Wasserman (eds.), *Metametaphysics: New Essays on the Foundations of Ontology* (pp. 178–212). Oxford: Oxford University Press.

Haslanger, S. 1989. Endurance and Temporary Intrinsics. *Analysis* 49 (3):119–25.

Heller, M. 1990. *The Ontology of Physical Objects: Four-Dimensional Hunks of Matter*. Cambridge: Cambridge University Press.

Hirsch, E. 2005. Physical-Object Ontology, Verbal Disputes, and Common Sense. *Philosophy and Phenomenological Research* 70:67–98.

Hirsch, E. 2008a. Ontological Arguments: Interpretive Charity and Quantifier Variance. In John Hawthorne, Theodore Sider, and Dean Zimmerman (eds.), *Contemporary Debates in Metaphysics* (pp. 367–81). Cambridge, MA: Blackwell Publishing.

Hirsch, E. 2008b. Language, Ontology, and Structure. *Noûs* 42 (3):509–28.

Hirsch, E. 2009. Ontology and Alternative Languages. In David John Chalmers, David Manley, and Ryan Wasserman (eds.), *Metametaphysics: New Essays on the Foundations of Ontology* (pp. 231—59). Oxford: Oxford University Press.

Hirsch, E. 2011. *Quantifier Variance and Realism: Essays in Metaontology*. New York: Oxford University Press.

Horden, J. 2014. Ontology in Plain English. *The Philosophical Quarterly* 64 (255):225–42.

Horwich, P. 1998. *Truth*. Oxford: Clarendon Press.

Jenkins, C. 2014. Merely Verbal Disputes. *Erkenntnis* 79 (1):11–30.

Lewis, D. K. 1986. *On the Plurality of Worlds*. London: Blackwell Publishers.

Magidor, O. 2016. Endurantism Vs. Perdurantism? A Debate Reconsidered. *Noûs* 50 (3):509–32.

Markosian, N. 2014. A Spatial Approach to Mereology. In Shieva Kleinschmidt (ed.), *Mereology and Location* (pp. 69–90). Oxford: Oxford University Press.

Merricks, T. 1994. Endurance and Indiscernibility. *Journal of Philosophy* 91 (4):165–84.

Merricks, T. 2001. *Objects and Persons*. Oxford: Oxford University Press.

Miller, K. 2005. The Metaphysical Equivalence of Three and Four Dimensionalism. *Erkenntnis* 62: 91–117.

Plunkett, D. 2015. Which Concepts Should We Use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 58 (7–8):828–74.

Plunkett, D., and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Quine, W. V. O. 1948. On What There Is. *Review of Metaphysics* 2 (5):21–36.

Quine, W. V. O. 1951. On Carnap's Views on Ontology. *Philosophical Studies* II (5):65–72.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: The MIT Press.

Rea, M. C. 1998. In Defense of Mereological Universalism. *Philosophy and Phenomenological Research* 58 (2):347–60.

Rosen, G. 1994. Objectivity and Modern Idealism: What Is the Question? In John O'Leary-Hawthorne and Michael Michaelis (eds.), Philosophy in Mind (pp. 277—319). London: Kluwer Academic Publishers.

Rosen, G., and Dorr, C. 2002. Composition as Fiction. In Gale, R. (ed.), *The Blackwell Companion to Metaphysics* (pp. 151–74). Oxford: Blackwell.

Sidelle, A. 2002. Is There a True Metaphysics of Material Objects? *Philosophical Issues* 12 (1):118–45.

Sider, T. 1997. Four-Dimensionalism. *Philosophical Review* 106 (2):197–231.

Sider, T. 2001. *Four Dimensionalism: An Ontology of Persistence and Time*. New York: Oxford University Press.

Sider, T. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Sundell, T. 2011. Disagreements about Taste. *Philosophical Studies* 155 (2):267–88.

Thomasson, A. L. 2007. *Ordinary Objects*. Oxford: Oxford University Press.

Thomasson, A. L. 2009. Answerable and Unanswerable Questions. In David Chalmers, David Manley, and Ryan Wasserman (eds.), *Metametaphysics* (pp. 444–71). Oxford: Oxford University Press.

Thomasson, A. L. 2015. *Ontology Made Easy*. New York: Oxford University Press.

Thomasson, A. L. 2016. Metaphysical Disputes and Metalinguistic Negotiation. *Analytic Philosophy* 57 (3):1–28.

Unger, P. 1979. There Are No Ordinary Things. *Synthese* 41 (2):117–54.

Van Cleve, J. 1986. Mereological Essentialism, Mereological Conjunctivism, and Identity Through Time. *Midwest Studies in Philosophy* 11 (1):141–56.

Van Inwagen, P. 1990. Four-Dimensional Objects. *Noûs* 24 (2):245–55.
Wallace, M. MS. Modal Parts and Ontological Disagreement.
Wheeler, S. 1979. On that which is Not. *Synthese* 41 (2):155–73.
Wiggins, D. 2001. *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.
Yablo, S. 2002. Go Figure: A Path to Fictionalism. *Midwest Studies in Philosophy* 25 (1):72–102.

4

Reactive Concepts
Engineering the Concept Concept

David Braddon-Mitchell

1. Introduction

This is a volume about the enterprises known variously as conceptual engineering[1] and conceptual ethics.[2] Exactly what is the nature of the relationship between these terms is still fluid; but a good way to rationalize it might be this. Conceptual engineering is the business of changing existing concepts and devising new ones. Conceptual ethics is the business of evaluating existing concepts and ways of talking, along with newly engineered ones, and making normative judgments as to whether they are fit for purpose. Jointly they promise reform through innovation and selection of how we think and talk in the pursuit of various ends—whether moral ends, practical ends, or alethic ends.

Thought and talk are the promised targets of these conceptual reforms. I'll set aside talk for now, though I'll come back to it. But 'thought' suggests—as does the word 'conceptual'—that at least in part concepts are a key part of the reformist's ontology. She will be engineering new ones, and recommending the adoption of some new ones, and the revision or sidelining of others. But what about the concept concept itself? Could there be ways of thinking about what concepts are that are different from standard understandings, but which have benefits that the conceptual reformer might take to be significant enough to merit recommending that we think and talk in terms of them?

This chapter recommends such a revision, or perhaps addition, to our stock of ideas about concepts. Classical and neo-classical accounts of concepts are thought to be connected in various ways to regular beliefs[3]—they enable us to have certain kinds of beliefs with certain kinds of content. The key idea in this chapter will be that there is a kind of mental concept which is connected to a mental state that is a little different from the usual conception of belief. Beliefs—straightforwardly in the functionalist tradition[4] but also in many others[5]—are individuated by what one might call

[1] Floridi (2011); Cappelen (2018).

[2] Burgess and Plunkett 2013a.

[3] For a taxonomy of accounts, see chapter 1 of Margolis and Lawrence (1999).

[4] Braddon-Mitchell and Jackson 2007.

[5] Consider, for example, the covariational program that starts with Dretske (1981) or the teleonomic one which has perhaps its best expression in Shea (2013).

David Braddon-Mitchell, *Reactive Concepts: Engineering the Concept CONCEPT* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © David Braddon-Mitchell.
DOI: 10.1093/oso/9780198801856.003.0004

their 'input' conditions. At its crudest these are just what distally causes them, but various traditions include what they co-vary with, what ideally causes them, and so on. At its most general, we might describe them as word-head connections. The connection to behaviour (and thus the head-world connection) is purely via their interaction with desires: distinct states which interact with beliefs to produce behaviour.

Let me give a motivating example which I will discuss later: an unspeakable word and associated concept which, in the case of racist speech in my culture (I'm an Anglo-Celtic Australian), represents an indigenous person. In order not to distract by mentioning the word (mention sometimes has spillover effects from use) I'll pretend that word is "Arthur" and the corresponding concept ARTHUR. What's racist about it? It's not just linguistic. It's that corresponding to that word there is a racist *concept*—or at least a mental entity in the concept family. And that mental entity does not merely represent skin colour or information about ancestry. Rather, it's one which when tokened itself *directly* causes negative affect of a range of types, which in turn both directly and indirectly motivates discrimination and other behaviours; none of this happens indirectly via discrete racist beliefs of the form 'Arthurs are lazy' and so forth. Recognizing that there are such mental entities in the concept family gives us the power to challenge them. If there is such a concept it's not one we should possess (although this chapter is of course asking you to possess the concept of such a concept—more on that distinction later as well). Perhaps, too, it will be explanatory: certain kinds of implicit bias may be explained by the fact that such a concept is possessed, even though no explicit negative beliefs are held, and the concept may be given a linguistic label the same as the anodyne representational concept (perhaps the tacitly prejudiced person possesses this special kind of concept ARTHUR while using the word 'Aboriginal'.

This chapter will recommend, then, that we consider a kind of concept which bears a relation like the one traditional accounts of concepts bear to beliefs, but instead bears it to states individuated not only by their causal inputs, but also by their *direct* causal outputs. I'll call these states **reactive representations**, **RR**s for short. They are partially representational states which are reactive inasmuch as they bypass interaction with distinct desires to directly motivate behaviour. In a later section I'll discuss how this relates to well known objections to 'besires': states that combine features of beliefs and of desires.

I will argue that there seems to be an important difference between two sorts of representational mechanisms: ones that really do act like mere beliefs where the representational state is discreet, and waits for a motivational state to come along and interact with it, and cases where the act of recognition—the act of representing—has itself *immediate* motivational force, as in the case of 'thick' moral terms,[6] certain sorts of phenomenal concepts and, I think, cases of bias and hate. When the motivating force is immediate, the motivation itself becomes part of how we taxonomize the

[6] It might seem that by explaining thick moral terms and slurs in terms of concepts I'm coming down firmly on the side of a debate about whether these are pragmatic or semantic phenomena (see Väyrynen 2013) for a defence of the pragmatic account). Certainly that's one way to taxonomize the view I will defend. But another would be to somewhat problematize the pragmatic/semantic distinction.

world: being apt for motivating in that way is part of what makes us see something as falling under the concept. I'll call a concept that is connected to an RR in the way an ordinary concept is connected to an ordinary belief a **Reactive Concept,** or **R-concept**. Whether there are such things, as I understand them, is an empirical matter which my research team is investigating. Here I argue only that the hypothesis that there are does a lot of explanatory work, and most importantly for current purposes, makes sense of the focus on the reform of concepts on normative grounds that has been called conceptual ethics (Burgess and Plunkett 2013a,b).

2. The R-concept Hypothesis

In this section I will first elaborate a little more on the hypothesis that there are R-concepts and then consider how they compare to ordinary concepts, ordinary beliefs, and desires. I then go on to consider whether the idea can be easily subsumed by global belief-desire psychology.

2.1. Concepts, R-concepts, Beliefs, and Desires

Exactly what the relationship is between beliefs and desires and representational content is complicated. But for our purposes here let it be assumed that concepts are part of what makes it possible to possess certain beliefs or desires, or what makes it possible to token a representation with a certain propositional content.[7] Insofar as you possess the concept 'triangle' you can token states that represent triangles. So perhaps to possess a concept is to possess a kind of ability. There are complications that don't matter for now: there is a tradition according to which a concept is what allows one to explicitly and consciously represent a certain state, others on which they are required to represent something *de dicto* rather than merely *de re* (on such views mental states might have non-conceptual content[8]). But the simple formulation will do for our purposes.

Concepts traditionally, then, fall firmly on the representational side of the mental family. If I possess the concept 'possum' then that is *part* of what it takes to be able to form at least *de dicto* beliefs about possums[9]. Rational people require, in addition, evidence of various sorts to form beliefs. In the case of beliefs, possession of the concept together with evidential or causal connections makes, in normal subjects, possession of the belief in some cases almost automatic. If I have the concept POSSUM and I am presented in experience with the features of possums that are encoded in POSSUM then at very least I will form the belief that it seems that there is a possum here. If I take myself to be in a position to trust my perception, then I'll form the belief that there is a possum nearby.

The idea here is firmly attached to standard belief-desire psychology. Beliefs are states which when formed have no power to motivate by themselves. Rather they sit in the head, waiting to be paired up with a state which will interlock with them. Their

[7] This is close to the rough characterization in Burgess and Plunkett (2013a) except that it's a necessary condition, rather than a sufficient one for thought formation.

[8] See Bermúdez (2007) and Roskies (2008) for some up to date discussion of this idea.

[9] And desires, but the point which follows about automatic production may not apply in this case.

semantic values, while they depend on the overall cause roles they play, perhaps together with desires, are invisible to the desire mechanism.[10] Once the belief that, for example, a possum is present is formed, a mental entity hangs around with a local syntactic shape that can be recognized by desire-like objects. If the desire to be near a possum encounters a belief that a possum is near, then (*ceteris paribus*) the agent doesn't move. If it is encountered by a desire for possum flesh, quite other behaviour is caused.

It is not part of the individuation condition of a belief that it causes any particular behaviour; the behaviour depends entirely on the agent's desire profile. And a standard concept is, inter alia, something that is required to form beliefs.

But notice that it's a contingent fact that the *token* physical state which is sitting around waiting for desires doesn't itself directly cause behaviour. In principle there could be such states; states formed by certain input conditions, which indeed might sit around and wait for desires, but in the meantime directly shortcut the belief-desire system to produce behaviours. There could be a state, co-varying and generally having whatever your favourite foundational semantics says makes it about possums, which directly motivates running away from possums. Let's call this state 'UGH-POSSUM'. The state itself, when formed, has the power to initiate movement away from the possum. Insofar as it has belief like features, it can also combine with other desires about possums (if you independently desire to sing a silly song about possums whenever near one, so you will sing as you leave). But it doesn't need any such extra desire to do *certain* things. This token state realizes[11] a belief that there is a possum nearby. But it does not *only* realize a belief that there is a possum nearby: it has extra powers that not every belief that there is a possum nearby would have. A kind to which this token state belongs is not individuated by the purely representational semantics, it's individuated as well by its output powers: it's powers to directly produce behaviour. Such a state is what I'm calling a reactive representation. An immediate concern that you might have is that this sounds an awful lot like a besire, and there are well known objections to the very possibility or coherence of such states. But nothing I say here depends on showing those objections don't work, as in section 4.2 I argue that in fact RRs are not besires.

The extension of the idea of a concept I'm recommending here bears the same kind of relationship to RRs as the traditional idea of a concept does to beliefs. The RR 'UGHPOSSUM' is a state which is belief like inasmuch as it responds to inputs; is caused by possums; and tracks perceptual evidence of possums. But rather than waiting for desires to come along and interact with it, it directly induces aversion of the detected object. The R-concept UGHPOSSUM is a kind of ability. It's the ability to form these states. Possession of the R-concept UGHPOSSUM is required to have the ability to form the RR 'UGHPOSSUM'. And just as in the case of regular concepts, where there were conditions in which possession of the concept makes forming beliefs automatic, there will be conditions in which possessing a certain R-concept will make tokening the relevant RR automatic. If someone who possesses a certain

[10] A point vigorously stressed in Fodor (1987) and countless subsequent publications.

[11] The reader wondering why I have said 'realizes' rather than 'is' will have her questions answered in section 4.2.

R-concept sees presentations as of the representational features encoded in the R-concept UGHPOSSUM then they will token the RR UGHPOSSUM, and this will in turn produce the downstream causal effect of fleeing the marsupial infested neighbourhood.

2.2. Comparison with Standard Accounts of Concepts

So in using R-concepts to taxonomize the mind we will draw distinctions differently. Two concepts that have the same input conditions will count as the same ordinary intensional concept. But if two states which have the same input conditions, but one gives the capacity to form regular beliefs, and the other gives the capacity to produce RRs which bypass the standard desire mechanisms to motivate behaviour, they count as different along the R-concept dimension. Is there any motivation to use the term 'concept' for R-concepts at all, rather than engineering an entirely different term for this sort of mental entity?

There are plenty of *prima facie* reasons for using the word 'concept' here: clarity, ease of explanation, and so on. But another is that we already need to vary the standard intensional individuation of concepts to explain mental phenomena which are amongst those whose explanation it is the job of a theory of concepts to provide.

Ordinary concepts are often thought to be fully individuated by their input conditions—or at least what it takes to fall under them is. But something other than the 'ordinary' idea of individuation of concepts by the features of the things which makes them fall under that concept may already be needed to account for merely hyperintensionally distinct concepts.[12] Some such hyperintensional distinctions might, for example, be marked by *method* of detection, rather than features required for the thing to fall under the concept. It takes the same features in the world for something to count as *falling under* TRIANGLE and TRILATERAL, but one can imagine someone disposed to try to identify these figures by counting angles insofar as she is deploying the TRIANGLE concept, and angles insofar as she is deploying TRILATERAL.

So are TRIANGLE and TRILATERAL the same concept or are they different concepts? Surely there is no need to make that call: they are intensionally the same but hyperintensionally distinct. If one wanted terminology, one could say they are same concept but different H-concept (friends of the hyperintensional often like to think that concepts are individuated hyperintensionally, and they use another world for intensional concepts, but that's a fight about words).

This seems to be the right thing to say in this case, and no-one I know of has suggested that insofar as concepts are hyperintensionally distinct they are not conceptually distinct in some sense (if anything, as above, the other move is more often heard—that concepts are hyperintensional through and through, and that intensional distinctions are coarser grained than the usual conceptual ones). And if it is the right thing to say in this case, then it seems to apply to the case of R-concepts as well. For here also we have intensionally (and maybe hyperintensionally) distinct

[12] For a sketch of an account of this that does not invoke metaphysically robust impossibilities, see Braddon-Mitchell (2009).

states, which because they function causally in slightly different ways, get to count as different in a different dimension from the intensional or hyperintensional.

2.3. R-concepts versus Belief Desire Psychology with Irrationality or a Divided Mind

So if the distinction between R-concepts and standard concepts is a good one, the easy job of saying that they deserve to be thought of as a kind of concept is done. But our comparison with standard accounts of the relationship between beliefs and concepts is not yet done. For there is a danger that R-concepts and RRs might be redundant given the way belief-desire psychology is often understood.

So far I have introduced the idea of RRs in terms of the different behaviour of token concepts. The thought is that there are tokens of the required types in the head, and there is a difference in behaviour between RRs and standard beliefs. Standard beliefs stand as if at a dance, looking for a desire partner with whom to generate behaviours. RRs do the work themselves, as it were, without a partner.

But on a global conception of belief-desire psychology this doesn't make sense. On this conception, thinkers don't have individual beliefs and desires. Rather they have global input output profiles, complete belief-desire states (Braddon-Mitchell and Jackson 2003: chapter 3). Individual beliefs (or desires) only make sense as abstractions that describe the minimal differences between global states.

On this conception having the R-concept UGHPOSSUM is indistinguishable from having an overall belief-desire profile which includes the belief that there are possums around, and a desire to vacate the possum rich environment. For you are globally set up such that when you encounter possums you flee them.

For many purposes that's the right conception of belief and desire, in my view. But I'm less sure than I once was that it's *always* the most useful conception. In particular, it doesn't play well with typical actual behavioural profiles in that it can't make sense of irrationality.

Suppose, for example, that you are disposed to apparently randomly flee possums or not. It turns out that what is going on is that you approve of possums; you have a standing desire to observe them. But nonetheless you possess the UGHPOSSUM R-concept, so your categorization process produces an immediate impulse to flee. At times when you are reflective, you overcome this because of your desire to observe and be in the company of possums. At other times you just react unreflectively and flee the neighbourhood. Or consider someone who sincerely claims that certain behaviours are likely to cause eternal torture, who sincerely believes that infinite torture massively outweighs fleeting pleasure, yet performs behaviours which would, if this were true, bring about the infinite pain and only modest pleasure—and does all this with only modest regret. The global story here has unilluminating fixes:[13] you have context dependent desires (you desire-when-reflective to be with possums, you desire-when-unreflective to avoid them) or the divided mind hypothesis (Lewis 1982; Stalnaker 1984; Davidson 2004; Egan 2008). It's unsurprising that these are a little unilluminating about the exact behaviour of mental states in less that perfectly

[13] For a nice account of the ways in which these are unilluminating, see Norby (2014).

rational beings like us: for all of these fixes are not designed to taxonomize the mind in terms of token states, but rather to preserve *global* accounts of the way the mind is, removing contradiction either by creating two separate consistent accounts (the fragmentation story), or one in which our desires and perhaps beliefs are so massively context dependent they don't play the kinds of roles in our ordinary discourse that they were intended to (e.g., a desire-when-reflective as against a desire-when-unreflective, or a belief-when-talking-about-the-act as against a belief-when-acting).

When we aren't talking about an abstract global theory of what it is to be a rational being (for which global belief-desire theory seems eminently suited) but rather interested in *token* explanations, the hypothesis that sometimes a single token state both motivates and represents, and at other times the work is divided between distinct token states, is both illuminating and empirically tractable. It seems to explain a lot if true, and it shouldn't be beyond our powers to find out if it is true.

3. General Explanatory Benefits of R-concepts Including Ones for Conceptual Ethics

In this section I argue that, if the R-concept hypothesis is correct, it would not only provide direct benefits in explaining the importance of conceptual ethics but would also provide general explanatory benefits.

3.1. Three Unified Benefits: Hate Speech, Crypto-evaluative Terms, Phenomenal Concepts

Here's another benefit of the idea of the R-concept. It provides a unity of explanation over a range of three apparently very different areas. I'll briefly survey the advantages that this story might have in each of the areas, suggesting that the unity over the areas it brings is a theoretical advantage, and that in each case understanding that there are R-concepts involved makes better sense of the projects of conceptual engineering and conceptual ethics in these domains.

3.1.1. HATE SPEECH

I've already given the example of hate speech: here the idea is that to hate speech there might correspond hate R-concepts; which give you token states which both represent the external world a certain way, and cause you to react in a certain way. That the reaction is so immediate also gives the impression that being apt for the reaction is part of what it takes to fall under the R-concept. So if we wanted a theory of content for R-concepts, then this would likely come with the territory. The R-concept *Arthur causes you to token states which represent indigenous Australians, to respond in a certain way and also to form the view of them being represented as rightly producing that response.* You do not truly fall under that concept, the racist is likely to think, unless you are deserving of the response. This makes richer sense of the recommendation that a practitioner of conceptual ethics might be doing when she recommends against using the word "Arthur". If it was just an issue of the *word* then the focus seems to be on two possible worries. One is not using the word because it gives offence. This is a weak justification for two reasons. One is that there is a liberal conception of offence as a very weak kind of harm, which hardly justifies restriction of this kind of speech unless we extend the harm into a kind of verbal violence. Perhaps a reasonable move, but not an uncontroversial one. The other is that it gives *evidence* of racist attitudes on the part of the speaker. But this would be superficial evidence: if the underlying *concept* is just the same as the representational concept "Aboriginal" the only racism is being careless with possible offence. But the intuition many have is that it's not an unqualified benefit to have racists hide their colours by careful choice of words. So there's something more going on than just choice of words. On the R-concept story, though, it is indeed evidence: but one which shows that the user of the word may possess a certain R-concept, something which other things being equal disposes them to prejudicial action. And making a recommendation against such a concept, even to someone in the grip of it, might sometimes help. Seeing how hard it is not to talk in those terms, she can become aware of it and, if in respect of her considered desires they are reasonable, she may try to remove the concept from her mental repertoire. I discuss in section 3.3 the prospects for finding a connection between use of words and R-concepts.

Perhaps here a few words are appropriate to discuss the extent to which accepting my view requires you to take sides on existing views about pejorative language. You might think that my view, tied as it is to R-concepts, falls firmly on the side of analyses which place the emphasis on the content of the expressions (see, e.g., Hom 2008). But that of course depends on what you mean by 'content'. I talk of the 'content' of R-concepts, and call R-concepts a kind of concept because I think that there are similarities in the way these notions do explanatory work to purely descriptive concepts and contents. But of course it's an *extension* of both ideas. And as such it's ultimately terminological (though not therefore unimportant) whether the extensions deserve the original terms. Another way of using the ideas I propose would be to accept that the mental entities I call "R-concepts" exist, and think of them as the underlying mechanisms that explain much of the pragmatics of thought and talk. The view is inconsistent, though, with any account according to which there is nothing mental that distinguishes what's going on in these cases from cases where there is no pejorative or hateful language. A very flat-footed version of Anderson and Lepore's view in their (2013) according to which the fact that the words are taboo is the *only* explanatory factor would be inconsistent with it, and the explanation would be entirely outside the head of the user. Finally, much of the literature concentrates on things that have speech does. There is a tendency in philosophy, as elsewhere, to promote single factor explanations, so that the promoter—should she win—gets to have a monopoly on being right. To the extent that linguistic behaviour has many important effects, one might think that this should flourish. So, for example, Camp's view (2013) that pejorative language serves to signal alignment with a certain perspective is something, which if just baldly stated like that is likely often true, and certainly consistent with what I say. Even stronger versions of it—such as the view that the distal explanation of its prevalence is to do with such signalling—are more empirically risky, but also consistent. What isn't consistent would be the view that none of the proximal psychology is as I describe.

3.1.2. PHENOMENAL CONCEPTS

At first blush, it might seem as though phenomenal concepts have very little to do with the topics we have been discussing. But I think that R-concepts explain important features of them as well. Consider a discussion which went on over many years between me and my friend the sadly late philosopher Jonathan McKeown-Green. Jonathan was blind from birth, and as a philosopher interested in perception, he was a natural person to talk to about colour concepts. At some level he thought he couldn't possibly have the same concept RED that I had. But intensionally he possessed a concept (let's call it Red-J) that was identical to mine: let's say it tracked a certain reflectance profile. And yet, he and I thought, it wasn't the same concept.

When there is similarity at the intensional level, but difference of concept, you might diagnose hyperintensionality. But that's not what's going on here. If one of our concepts was hyperintensional, it would categorize in a more fine grained way. But there's nothing more fine grained here in either of our concepts.

Perhaps it's response dependence that's at issue (Pettit 1991). But moving to a typical response dependent conception doesn't help either. If Red is to be characterized as "apt in normal conditions to produce response R in species-typical individuals" then again we have an ordinary intensional concept, and something that Jonathan and I could agree on, while not dispelling the sense that we don't possess the same concept—that somehow, while we agree on this, he didn't understand what sighted people *mean* by a term like 'red'.

But notice that there is an R-concept and RR in the neighbourhood. Recall that an RR is something which is responsive to information about the world, but part of its individuation is via not just those inputs, but its outputs—the things it causes. So far the outputs we have considered are ones of reactive attitudes and behaviours. But reactive phenomenologies are just yet more causes.[14] To possess the R-concept RED is to possess a capacity (and perhaps a disposition) to token the RR 'red'. The RR 'red' is tokened when the agent receives certain information about the world[15] but also has the effect of producing the phenomenology.[16]

This, then, is something Jonathan and I do not share. He does not possess the R-concept Red as he does not possess the ability to token the RR 'red', even though he does share the intensional concept Red. The R-concept RED has as a feature an effect on what it takes to be Red in this sense—that is, had produced this effect in me. So while there is a sense (derived from the intensional concept RED) in which we can

[14] By phenomenologies here I don't mean anything substantive. If you are a dualist who believes in rich qualia, let them be the causal product of tokening certain RRs. If you are at the other extreme, and think it's all about dispositions to make verbal claims about experience, let those dispositions be possessed by the RR. And for every positions in between the same can be said.

[15] If this were a chapter about phenomenal concepts, then some discussion of whether the input channels are relevant here would be appropriate—maybe it makes a difference if it is tokened by the perceptual system.

[16] There are interesting issues here about how integrated an RR has to be. Perhaps empirically we may find the cause of phenomenal component in the perceptual system, and the cognitive component elsewhere. And yet somehow there is a "binding" of the two. The bound pair might be thought to be a cause of the phenomenology, but possibly not the cognitive component alone.

both judge that some flower is red (me by looking, he by consulting a colour meter) there is another in which he cannot judge that it's red in the same way as me.

3.1.3. CRYPTO-EVALUATIVE CONCEPTS

Finally a certain kind of puzzle in giving an account of certain concepts is explicable if there are R-concepts in the vicinity. These are concepts I call 'crypto-evaluative'. They might include HAPPINESS, FREEDOM, DEMOCRACY, and many others. I call them crypto-evaluative because on the surface they seem amenable to analysis in terms of what features of the world are required for them. We might, for example, give a story about what decision making processes count as democratic. But they are all vulnerable to normative objections that expose the hidden evaluate features. Suppose an agent considers an excellent account of what democratic decision making amounts to (and thinks herself a democrat) but judges that that's not a good way to organize society. Likely she will make the 'true democracy' move: she will be less inclined to judge that that procedure really is democratic.

Consider the debate about the concept of happiness. Various descriptive accounts of happiness—ones which feature objective lists, or which feature psychological states—have been criticized on the grounds that these things could in principle turn out to be undesirable. (Nozick 1989; Nussbaum 2008). The obvious alternative is explicitly normative accounts: happiness is the greatest prudential good, or happiness is flourishing of some kind. These in turn are criticized as leaving out a substantial account of what happiness is (Haybron 2008; Feldman 2010). (Maybe the concept is one which is committed to mental ease, or some other psychological state, as part of what it takes to be happy—no-one is happy, such a response goes, however much she has the objective list if she is miserable.)

If there is an R-concept HAPPINESS to be had (perhaps in addition to a standard one) then what that R-concept amounts to is the ability to form the RR 'happiness'. And, perhaps, the RR 'happiness' is tokened when there is a state which is caused by (or otherwise represents) some list of features—perhaps in this case internal psychological ones, perhaps also elements of an objective list—but which also causes motivation towards these features. So no-one would judge that she is happy (insofar as the R-concept is playing a role in the judgement) unless her representation of these states motivated her towards it. Thus we get an explanation of how someone can both be prone to think that there is substantial empirical content to the concept of happiness, and yet be drawn to the idea that some normative features are essential to it. Crytpo-evaluative concepts are just R-concepts.

3.1.4. TACIT BIAS

I'll add one very brief speculation. A lot has been written about implicit bias; bias where agents seem to have both explicit beliefs and explicit desires which don't appear to have any bias towards a group of people. And yet, the agent behaves in a biased way: perhaps being more inclined to vote against someone from a minority group in a job selection, for example. One hypothesis worth testing would be that R-concepts are involved here. If the biased motivation is part of the R-concept—a direct causal product of the representation—and not a result of stand alone desires metaphorically looking for beliefs to pair with, then it's more likely that such a pattern might be something the agent herself could miss when interrogating her discreet beliefs and desires. There may be no stand alone desire not to see members of that group employed, and no stand alone belief that they shouldn't be, but the R-concept makes it likely that, confronted by members of a group, an RR will be formed which directly causes the biased behaviour.

3.2. Conceptual Difference Where There Is No Disagreement About Fact

So far I have done some setup about the idea of the R-concept, and discussed some of benefits of having the notion to hand. In this section I'll talk about another case that I think makes R-concepts worth exploring for conceptual engineers.

Sometimes there seem to be disagreements which matter a lot, but which don't seem to depend on any disagreement of fact.

Consider the concept of 'survival'[17] It's very contentious what exactly survival amounts to. I'm tempted myself to think that it's a difficult notion, which admits of degree (Braddon-Mitchell and Miller MS). But for current purposes we need only consider two alternatives; a physical continuity theory and a psychological continuity theory. The simple physical continuity theory is just one that says you survive over some interval iff you are related to the entity at the end of the interval by the ancestral of the physical similarly relation. That relation in turn holds when there is object causally connected to you which overlaps with your current self physically to a very high degree. The psychological continuity theory is much the same, substituting psychological similarity for physical similarity. It's usually assumed that our account of psychological similarity is one which allows psychological similarity across discontinuous physical change—as with material destruction followed by reconstruction, or downloading.

It's notorious that people disagree vigorously about the plausibility of these two theories, which in part is what makes them good things to discuss in introductory courses.

I've tracked the opinions of undergraduate students on these topics for a number of years, and found that roughly speaking 40% of classes of 400 students or so are attracted to psychological continuity theories, and 45% or so are attracted to physical continuity theories, at least insofar as they take this difference to explain their different reaction to teletransporter cases.

The well known teletransporter cases are ones where there is bodily destruction but recording of the psychological information, followed by bodily recreation at a distant location, down to the relevant degree of detail required for the psychological information to be preserved.

Those who are attracted to the psychological continuity theory say that teletransporters, should they come to exist, will provide painless easy travel, and will be happy to use them if the price is right. Physical continuity theorists are genuinely baffled by this, and tend to scream at their fellow students of the other persuasion "you don't get it: you are being paid money to be killed".

[17] The *locus classicus* is of course Parfit (1984).

Now the question that I'm hoping R-concepts can help with is this: what is the disagreement about?

This question arises because it is possible to divide up the sample of students in ways that factor out obvious sources of disagreement.

Here are some obvious sources:

(1) Perhaps some students believe in souls, and really hold a soul continuity theory. But they think that souls are directly connected to bodies, at least insofar as boldly destruction severs the soul from the earth.

(2) Perhaps some students are dualists, and they think that no matter what I try to tell them about how a recreated brain will support consciousness, they don't accept this, and thus think they'll no longer exist after teletransportation, because there'll be no experience.

(3) Perhaps students think that there are substantive metaphysical facts about what survival is grounded in. Survival is worth having, they think, and it is a natural feature of the world (other things people might call survival are mere gerrymandered things of no interest) and the job of metaphysics is to find what grounds survival: so these two possible candidates, the two continuity theories, are alternative accounts of what the grounds of survival are. The disagreement is about the metaphysical fact about the nature of the grounds of survival.

So in order to eliminate these possibilities as sources of disagreement, I excluded from my sample people who believe in souls, people who are dualists, and people who believe in grounding. Each of these groups have an answer to what the disagreement of fact might still obtain between people who agree on the fundamental physical facts. The people who believe in souls think that while we might agree about the physical body and its constituents, there is something else—the soul—which might or might not be present, and we could be tacitly disagreeing about that (your body will survive, and your brain, but not your soul). Dualists, very similarly, might take the view that the dualistic component of the mental may survive some (identity preserving) physical changes, but not others. And finally, these undergraduates were taught about metaphysical grounding: the version in which grounded entities are ontologically distinct yet grounded in underlying physical nature. On many such views it's a substantial metaphysical fact which fundamental states ground the non-fundamental, and if persons are non-fundamental, there can be a substantial disagreement as to which physical states ground them.

On eliminating these people whose views take them to think there is a substantial disagreement there was only a slight change. Of the remaining group (luckily for current purposes my university is in a culture with few soul believers amongst undergraduates) the split is now about even. And now the puzzle is more serious. All of them agree, for the sake of the example, on all the base metaphysical facts. They agree about the nature of physical continuity. They agree about the nature of psychological continuity. They agree about background physics and science more generally insofar as it is relevant. They agree that there are no substantial grounding relations which connect more abstract things like 'grounding' with the basic facts in the universe. So what is left?

The remaining obvious possibility is something like merely verbal or lexical disagreement. They disagree about what 'survival' means. But that seems too thin a thing to justify the reactions. The physical continuity theorists still hold that it's mad and bad to teletransport, because it doesn't preserve survival. If it were a mere lexical disagreement, then you'd expect that it could be solved by dictionary makers. And yet no physical continuity theorist thinks she would be dissuaded by the discovery that the *OED* says that the psychological continuity theory is the right definition of the English word 'survival'. Indeed, not all physical continuity theorists think that their own theory *does* give the right account of the English word, or that it matters that it does.

Parallel with a lexical story is what one might call a "mere" conceptual story. The disagreement is about which concept to possess. But the problem with this is that *both parties already possess both concepts*. Neither party thinks that one shouldn't possess the other concept. Indeed the physical continuity theorist couldn't think about what she is right about were it not for the fact that she possessed the concept 'the thing that is preserved in psychological continuity, and that PC theorists call "survival"'.

So what is the disagreement about, if it's not fundamental metaphysics, grounding, semantics, or concepts?

Well of course the striking difference is a difference that looks like it's not one of belief, but of motivation. The psychological continuity theorist is very much motivated to preserve psychological continuity at almost any cost. The physical continuity theorist is similarly motivated with respect to physical continuity.

So what's going on here might be the possession of different R-concepts. The teletransporters (those who believe they survive teletransportation) have an R-concept which takes facts about psychological similarity as input, and *directly* motivates behaviour seeking psychological similarity at all costs. The others have an R-concept which takes facts about physical similarity as input, and similarly motivates behaviour seeking *physical* similarity at all costs. The first category is motivated to provide benefits to their psychological successors; the second will only be so motivated if they are physical continuers.

How does the R-concepts strategy help? If it were merely a matter of each side using different ordinary concepts, it would be still mysterious. For as I will explain in section 4.1, successful explication of an ordinary concept to an agent entails it's possession by that agent, so both sides would possess both concepts. But as we will see, R-concepts are different: you can explain the R-concept ARTHUR successfully to someone without that person thereby possessing it (even if she possesses a kind of meta-concept, as we will see). So it's quite possible for the groups to possess different R-concepts, for this to explain the difference, and for that fact not to change even when we explain the idea of the R-concepts to both sides.

3.3. R-concepts and Words: How to Engineer R-concepts

A principal part of the justification for a theory of R-concepts is not just that it's a better way to taxonomize parts of our cognitive economy, but that it helps in the task of conceptual revision under the guidance of conceptual ethics.

The thought so far is that it gives us as conceptual ethicists a story of why certain R-concepts are bad ones, and as engineers, a motivation to engineer new ones and try to persuade people to adopt them.

But short of as yet practically impossible, and likely highly immoral, direct neural intervention with our conceptual structures, the principal handle we have on our conceptual life—at least our more complicated concepts—is via words. So it may pay to consider the relationship between concepts and words, and R-concepts and words.

The use of words amongst those who understand them tends to token the concept required for their understanding. If I say to you "be alert for triangles" you'll be more likely to look at your environment in a way which is alert to their presence, and thus form beliefs that there are (or aren't) triangles in the area.

The interesting question for our current concerns is how the word-concept pairing works in the case of R-concepts. I suggested earlier that R-concepts may play an explanatory role in the case of thick moral concepts. The question remains whether the word-concept pairing is sufficiently robust that influencing verbal behaviour can be part of an intervention strategy to minimize the harmful effects of harmful R-concepts.

If it were, then we should expect something like this. The word "Arthur" is associated with the R-concept Arthur. The phrase "Indigenous person" is associated with the R-concept Indigenousperson. If you ask someone (who is neither a committed racist nor alert to conceptual and linguistic contributions to discrimination and oppression) to "look out for Arthurs" then, on sighting an indigenous person, you might expect her to be more likely to form an RR with the content Arthur which in turn produces automatic fear responses and general negative affect. If on the other hand she is primed differently, and asked to look out for indigenous people, then on sighting such a person, you might expect her to be more likely to token a representation only of the regular concept Indigenous person, and thus have more neutral responses.

Of course all of this is tricky, since word-concept pairings are not entirely robust. Perhaps part of what it is to be a committed racist, aside from explicitly held racist beliefs, is to have the racist concept Arthur so firmly entrenched that all words for indigenous people will token representations with the content Arthur, either directly or via a very short chain of inference.

This, of course, is a second empirical matter concerning R-concepts which needs investigating: not only the crucial one of whether we actually possess them, but also the way in which the tokening of RRs under their influence is, or is not, mediated by related words.

This is perhaps a good point to add one further speculation. A lot has been written about implicit bias; bias where agents seem to have both explicit beliefs and explicit desires which don't appear to have any bias towards a group of people. And yet, the agent behaves in a biased way: perhaps being more inclined to vote against someone from a minority group in a job selection, for example. One hypothesis worth testing would be that R-concepts are involved here. If the biased motivation is part of the R-concept—a direct causal product of the representation—and not a result of stand alone desires metaphorically looking for beliefs to pair with, then it's more likely that such a pattern might be something the agent herself could miss when interrogating her discreet beliefs and desires. There may be no stand alone desire not to see members of that group employed, and no stand alone belief that they shouldn't be, but the R-concept makes it likely that, confronted by members of a group, an RR will be formed which directly causes the biased behaviour.

4. Two Objections to the Theory of R-concepts

4.1. Failing to Possess the Concept and Failing to Understand the Concept

Here is a *prima facie* puzzle. With regular, purely representational concepts, understanding a concept is roughly equivalent to possessing a concept. Once I understand that the concept of TRIANGLE is the concept of a plane figure with three sides, it follows that I automatically possess the concept. Understanding what that concept is gives me that concept. So by understanding what the concept TRIANGLE is, I will end up with any of the powers that are associated with the concept: the ability to form thoughts about triangles, for example. By understanding the concept MOTOR CAR I come to possess the concept, and form the ability to form thoughts about them. But by understanding the R-concept ARTHUR we do not, we hope, thereby possess it, and do not have the ability to form the relevant RR. This fact might make one think that an R-concept is not really behaving like a concept.

It is helpful here to distinguish between a concept, and a meta-concept: the concept of that concept. In the case of ordinary concepts they are distinct, but for most purposes that distinctness doesn't matter. The meta-concept of triangle is, for example, the concept of some kind of mental entity or power that enables one to form thoughts about three sided plane figures. That's not the same as the first order concept TRIANGLE. But acquiring the meta-concept is sufficient for acquiring the concept, so it rarely matters. Perhaps it matters in animal cognition or developmental psychology, where plausibly some animals, and humans at some stages of development, can possess concepts but not meta-concepts.

But the distinction matters here in realm of R-concepts. A possible complaint about the idea of an R-concept is that the idea that we might be able to resist such concepts is absurd—conceptual engineering will never be able to bring it about that we do not have these concepts, especially not if we go about explaining them. For typically, to explain a concept successfully is to bring it about that the person to whom you have explained it possesses the concept. If I do a really good job of explaining an alien concept to someone, perhaps the concept associated with the Danish word "Hygge"—a concept of cosiness and domestic wintry comfort which is at the time of writing fashionably held to be a key contributor to wellbeing—then I will have succeeded insofar as she now possesses the concept. So if I somehow explain the R-concept ARTHUR or the R-concept perhaps associated with the word for which "The N word" is a euphemism—then it might seem that I will bring about its possession. Then, as an objection, this can cut one of two ways. Either this is true, and so there's no benefit to talking about R-concepts, or else it's an objection to the possibility of there being any such thing as an R-concept.

The latter horn of the dilemma goes this way:

(1) (Premise) if R-concepts are concepts then explanation will result in possession.
(2) (Premise) Possession of R-concepts has motivational force and casual power.
(3) (Premise) Mere explanation cannot bring about by itself motivational force.
(4) (from (2) and (3)) Explanation will not bring about possession of R-concepts.
(5) (Conclusion—from (4) and (1)) R-concepts are not concepts.

The conclusion comes from applying *modus pollens* to the conditional in the first premise. But of course this is not a compulsory move. Instead we should consider how generally true the first premise is.

We saw that the first premise is indeed true of purely representational concepts. But it is not true of R-concepts. But this is because the relationship between first order concepts and meta-concepts is different in the case of R-concepts.

Mere explanation about the nature of an R-concept will create a meta-concept. But it will be a purely representational meta-concept. Not an R-concept. And purely representation meta-concepts do not have the automatic power to create the first order R-concepts they represent, in the way that they do have the power to create first order representational concepts.

Consider our racist example again. Giving lots of information about the R-concept ARTHUR will tell someone that there is a mental state which responds to evidence of Aboriginality, which is able to interact with standing beliefs and desires, but that without any such interaction motivates fear, distrust, and dislike. This information creates in a comprehending recipient, a concept, for sure. It creates a concept (a representational concept) of the R-concept ARTHUR. So it's a representational meta-concept. But having the concept of such a concept does not allow you to form thoughts of the kind "there's an Arthur". Because thoughts of that kind require you to have an RR, usually brought about by having the R-concept ARTHUR. And indeed nothing purely informational will bring that about. Of course there are meta R-concepts too: these might be concepts of R-concepts which come with built in motivational force for the acquisition of the first order R-concept. Perhaps these sorts of meta R-concepts may have some power to bring about possession of the R-concept, but these meta R-concepts cannot be acquired by purely informational means either.

So premise (1) should be rejected. It's not true of all concepts that explanation of them results in possession of them. Of course this does show that there is a difference between purely representational concepts of which this may be true, and R-concepts for which it is false. But that's unsurprising. R-concepts are different from regular concepts, and this difference exactly tracks the key, motivational, difference.

4.2. Reactive Representations and Besires

There is a substantial body of literature which denies the possibility of so-called **besires**: states which are both beliefs and desires (e.g., Lewis 1988, 1996; Price 1989; Smith 1994; Zangwell 2008). R-concepts are defined in terms of reactive representations: and RRs at first blush look like they play both the belief role and the desire role.

This would make them a kind of besire. There can be no R-concepts without RRs, so if there are no besires, there are no R-concepts.

Let's lay this out:

(1) (Premise) R-concepts are defined in terms of RRs.
(2) (From 1) If there are no RRs there are no R-concepts.
(3) (Premise) RRs are a kind of besire.
(4) (Premise from the Besires literature) Besires are impossible.
(5) (3 and 4) There are no RRs.
(6) (CONCLUSION from 5 and 2) There are no R-concepts.

It is far from uncontroversial that premise (4) is true. The besires literature is loaded with arguments for and against the possibility of besires. But I'm going to accept the premise: partly because I believe it to be true, and partly it puts my suggestion in a dialectically better position if it's compatible with the impossibility of besires.

The strategy I'll adopt is perhaps surprising. I'll deny premise (3) which says that reactive representations are a kind of besire.

Reactive representations certainly play some the causal roles that beliefs and desires do. But recall that a besire is supposed to be both a belief and a desire. Importantly, as I hope to make clear, it's a type of state that is both a belief with a certain content and desire with certain content. This is in contrast with the idea that there might be a *token* state which is both a belief and a desire.

To see this consider this characterization by Smith of the problem that besires pose for Humeans:

> W]hat Humeans must deny and do deny is simply that agents who are in belief- like states and desire-like states are ever in a single, unitary, kind of state.... And their argument for this claim is really quite simple. It is that it is always at least possible for agents who are in some particular belief-like state not to be in some particular desire-like state; that the two can always be pulled apart, at least modally. This according to Humeans, is why they are distinct existences. (Smith 1994: 119)

The thought here is that if you are in a belief-like state, there is no necessary connection to any desire-like state. If I believe that there are possums around me, and I desire to flee, it's always (logically) possible for me to be in the same belief state but some other desire state: perhaps wanting to cuddle them. But if the belief that possums are around, and the desire to flee, are the very same state—a besire which is both a belief and a desire—then that isn't logically possible. For this state (which is the belief that possums are around) is in part individuated by also being the desire to flee possums. Should I no longer desire to flee possums, I am no longer in the same state: so I no longer have the belief that there are possums around. This violates the Humean doctrine that for any belief, I may combine it with any desire.

This is a doctrine I subscribe to, so this thought should affect me. And it does. If besires are both beliefs and desires in the sense that the belief that P and the desire that Q could both be type identical to a particular state which has belief-like features and desire-like features, then I am persuaded that there are none.

David Lewis' more complicated arguments (Lewis 1988, 1996), which I shan't discuss in detail here, nevertheless depend I think on a similar thought. The rules for evolution of beliefs and desires are different. By "rules" here we mean something like the patterns of evolution which are constitutive of being a belief or a desire. If there is one state which is both a belief and desire in this type sense, the rules of evolution for belief allow it to change in a way which has no effect on its motivational aspects. But that means it's still the same desire—except it isn't, because desires were meant to be identical to states which have the representational features we have evolved away from. Similarly the rules for evolution of desire allow us to evolve away from our desire profile without affecting the world-head aspect of the state. Since the world-head aspects have not changed, we have the same original belief. But by hypothesis our desires have evolved, so we do *not* possess the same desire. But the original belief we possessed was supposed to be identical with this desire, so we cannot have the original belief and fail to possess the desires. So the desired hypothesis in its most flat footed form leads to contradiction. So that's why there can't be any besires, if by besire you mean a type of state that is both a certain belief and a certain desire, and by "is both a certain belief and a certain desire" you mean type identical to both. But are reactive representations besires?

No. Because they aren't type identical to beliefs or desires. Consider the reactive representation 'UGHPOSSUM'. Someone who tokens it perhaps does token, in belief and desire terms, the belief there is a possum and the desire to run away.[18] So 'UGHPOSSUM' plays a role, perhaps, of being the realizer of the belief that there is a possum, and the desire to avoid possums. But it is not type identical to either. Should we evolve in such a way that we no longer take there to be possums around, but are still disposed to run should there be some, then we no longer have the belief that there are possums, we no longer have the RR UGHPOSSUM but we still have the desire to avoid possums, albeit now realized differently. So the RR UGPOSSUM is not identical to the desire to avoid possums, even though it may *realize* it on occasion. Equally, if we remain in a state which covaries with possums, while somehow coming to tolerate them, we may still be in a state which we can call a belief that there are possums around, but we are no longer in UGHPOSSUM, nor do we desire to avoid them. Thus UGHPOSSUM is not identical to the belief that there are possums around, though again it may realize it from time to time.

So a reactive representation is not type identical to any belief, or any desire. A besire, on at least some understandings—the ones that make them problematic—is something that is type identical to both a belief and a desire. Thus RRs are not besires, thus premise 3 of the above argument is false and the argument fails.

Why do I keep saying type identical here, as though I have a contrast in mind? If fact this is optional. I could just say that the RR is not identical to a belief or desire, but that at some times the RR is a realizer of both the belief and desire (or perhaps that some neural state realizes both the RR, the belief and the desire). When desire

[18] Actually I'm not sure we should say that she necessarily tokens the desire to run away. Maybe she only has the desire to run away if the basis for her disposition to run from possums is a more general state that is capable of interacting with various different beliefs.

evolution happens so that we no longer desire to flee possums, but still believe there are some, the token physical state that realized the belief that there were possums and a desire to avoid them is no longer present, but some other physical state exists which realizes the belief that there are possums without realizing the desire to avoid them. The RR too no longer exists. Problem solved.

But type identity talk, as the reader may have guessed, is useful for its contrast with token identity. Exactly how much sense the type token identity distinction makes is controversial. Nothing I say here depends on it being a good distinction, but I take talking about it to be helpful. This is because it explains the sense in which it's easy to see that people think that any kind of talk which is even close to besire talk, is taking seriously the idea of a belief being identical to a desire. And if you think that, you might think the same about RRs, and therefore be puzzled why the standard objections to besires don't apply.

Here then is the thought: you might think that the RR is token-identical to a token of the belief that possums are near, and a token of the desire to avoid possums. Another way of putting this is of course to say that there is just one token state: in virtue of certain features it's right to call it the RR 'UGHPOSSUM'. In virtue of others it's right to call it a token of the belief that there are possums, in virtue of yet other features its right to call it a token of the desire to avoid possums. What happens when that stage changes, so that, for example, it no longer serves the role of generating possum avoidance? The token state at one level of individuation no longer exists. Nor does the desire to avoid possums. But the belief that there are possums does, now realized by a successor token to the original one. Perhaps it will help to think of the relationship between beliefs and desires and RRs as the underlying token states change as something like a counterpart relation: perhaps analogous to Ted Sider's temporal counterpart relation (Sider 2001), except rather than all counterpart relations being of one kind (in Sider's case person counterparts), we can help ourselves to Lewis' notion of the same thing having different counterparts under different counterpart relations. So the initial state is token identical to a belief (that there are possums around), a desire (to avoid possums), and an RR (because the connection between the representation and the behaviour is extremely direct. Suppose the state evolves in such a way as to no longer motivate possum avoidance. The new state bears the *belief-counterpart* relation to the original state (and so we say in ordinary language that the agent has the same belief) but not the desire or RR counterpart. Suppose the state evolves over time in a way that results in the agent no longer behaving as though there is a possum around, but still having the disposition to avoid possums. The new state bears the *desire counterpart relation* to the original state, so we say it is the same desire, but does not bear the belief or RR counterpart relation. Suppose the state, or the agent, evolves in such a way that the agent will act as though there are possums around, still is disposed to avoid them, but that these features are less direct and automatic: the aspect of the agent which registers the possum has to interact cognitively with an interrogable desire to avoid them to produce the behaviour. Then the state bears the belief counterpart relation to the original state, and the desire counterpart relation to the original state, but not the RR counterpart relation. So we say she has the same belief and desire, but not the same RR.

5. Conclusion

In some ways this chapter has been an advertisement for a research programme. What the advertisement does is suggest that there will be substantial explanatory benefits to the idea of an R-concept and its associated RR, and that many of these are of special interest to conceptual engineering and conceptual ethics. If there are such things they may not only explain apparent disagreements where there is not underling disagreement of facts (like the survival case) but also lay bare what is at stake in proposals to reform or change such concepts. The idea unifies puzzles about phenomenal concepts, crypto-evaluative terms, and hate speech. It might explain what is at issue between people who are arguing about crypto-evaluative terms (and perhaps other thick moral terms) and thus when the different participants can or should revise their usage. It provides a tempting explanation of how hate-concepts might be associated with hate speech, and how such concepts differ from merely descriptive ones associated with inflammatory words. It is suggestive of mechanisms where bias can be purely implicit, and inaccessible to an agent honestly interrogating her beliefs and desires. It is equally suggestive of how it might be that linguistic intervention can generate better consequences than just hiding people's racist or otherwise biased attitudes. One way in which this is a very different approach to, say, Haslanger's (2000) view is for her it's much more important what the 'input' conditions are: what it takes descriptively to fall under a concept. Reasons for linguistic intervention in the concepts we have will fall on the side of the social consequences of categorization. I add to this the consequences for the concept possessor; in what possessing such a concept makes one do over and above categorization.

All of this requires that we really have such mental states, but the explanatory merits of it seem to highlight the importance of the empirical project of determining whether we do. But this chapter has at least done some of the groundwork in ruling out two tempting *a priori* objections: that R-concepts function in such an odd way that even if things a bit like them exist, they don't do any of the work that would justify seeing them as a kind of concept, and that they require use to possess besires.

References

Anderson, L., and Lepore, E. 2013. Slurring Words. *Noûs* 47 (1):25–48.

Bermúdez, J. L. 2007. What Is At Stake in the Debate about Nonconceptual Content? *Philosophical Perspectives* 21 (1):55–72.

Burgess, A., and Plunkett, D. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, A., and Plunkett, D. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Braddon-Mitchell, D. 2009. Naturalistic Analysis and the A Priori. In David Braddon-Mitchell and Robert Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press.

Braddon-Mitchell, D., and Jackson, F. 2007. *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Braddon-Mitchell, D., and Miller, K. MS. Survival in Continuous Degrees.

Camp, E. 2013. Slurring Perspectives. *Analytic Philosophy* 54 (3): 330–49.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Davidson, D. 2004. Paradoxes of Irrationality. *Problems of Rationality*. Oxford: Oxford University Press.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Egan, A. 2008. Seeing and Believing: Perception, Belief Formation, and the Divided Mind. *Philosophical Studies* 140 (1): 47–63.

Feldman, F. 2010. *What is This Thing Called Happiness?* Oxford: Oxford University Press.

Floridi, L. 2011. A Defence of Constructionism: Philosophy as Conceptual Engineering. *Metaphilosophy* 42 (3):282–304.

Fodor, J. A. 1987. Why There Still Has to be a Language of Thought. *Psychosemantics*. Cambridge, MA: MIT Press.

Haslanger, S. 2000. 'Gender and Race: (What) Are They? (What) Do We Want Them To Be?' *Noûs* 34 (1): 31–55.

Haybron, D. M. 2008. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. Oxford: Oxford University Press.

Hom, C. 2008. The Semantics of Racial Epithets. *Journal of Philosophy* 105: 416–40.

Lewis, D. 1982. Logic for Equivocators. *Noûs* 16 (3):431–41.

Lewis, D. 1988. Desire as Belief. *Mind* 97:323–32.

Lewis, D. 1996. Desire as Belief II. *Mind* 10:303–13.

Margolis, E., and Laurence, S. (eds.) 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Norby, A. 2014. Against Fragmentation. *Thought: A Journal of Philosophy* 3:30–8.

Nozick, R. 1989. *The Examined Life: Philosophical Meditations*. New York: Simon and Schuster.

Nussbaum, M. C. 2008. Who Is the Happy Warrior? Philosophy Poses Questions to Psychology. *The Journal of Legal Studies* 37:81.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Pettit, P. 1991. Realism and Response-Dependence. *Mind* 100 (4), new series:587–626.

Price, H. 1989. Defending Desire-as-Belief. *Mind* 98:119–27.

Roskies, A. L. 2008. A New Argument for Nonconceptual Content. *Philosophy and Phenomenological Research* 76:633–59.

Shea, N. 2013. Naturalising Representational Content. *Philosophy Compass* 8 (5):496–509.

Sider, T. 2001. *Four-dimensionalism*. Oxford: Oxford University Press.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Väyrynen, P. 2013. *The Lewd, the Rude and the Nasty: A Study of Thick Concepts in Ethics*. Oxford: Oxford University Press.

Zangwill, N. 2008. Besires and the Motivation Debate. *Theoria* 74:50–9.

5

Strategic Conceptual Engineering for Epistemic and Social Aims

Ingo Brigandt and Esther Rosario

Conceptual analysis, as traditionally practiced by philosophers, consists in articulating a concept by means of imagining various scenarios and using one's intuition as to whether or not the concept (e.g., the concept KNOWLEDGE or the concept CAUSATION) applies to a given case (Jackson 1998; Brigandt 2011). This approach makes explicit the concept one currently happens to possess. In contrast, the project of *conceptual engineering* attempts to put forward the most suitable concept for a philosophical task, which may well require revising or abandoning one's current concepts. Kevin Scharp (2013) champions this approach when arguing that the concept of TRUTH is inconsistent and theoretically unsuitable and that it should be replaced with several successor concepts. In what follows we primarily use the label 'conceptual engineering' for this general approach, although another option is to employ the term *conceptual ethics*, as advocated by Alexis Burgess and David Plunkett (2013a,b), given that the latter label highlights normative questions, such as what makes one concept better than another one, which concepts should be used, and which concepts should be discarded (see also Plunkett and Sundell 2013; Plunkett 2015, 2016). A prominent illustration of this philosophical practice is the prior work by Sally Haslanger (2000, 2004, 2005, 2006) on the concepts GENDER and RACE. Her "ameliorative analysis" of gender and race, with definitions that incorporate the hierarchical social treatment of different genders and races, deliberately deviates from standard definitions (even philosophical ones), so as to provide new concepts meant to serve feminist and anti-racist aims.

We contribute to discussions on how to construe GENDER and RACE, where our reflections on these concrete cases will also provide guidelines for the general philosophical project of conceptual engineering. In this essay, we put forward various suggestions without endeavouring to articulate a specific concept of GENDER or RACE. In fact, our general position will be that there is no unique, privileged concept of either GENDER or RACE.[1] Rather than an alleged all-purpose concept, what is needed is a plurality of concepts of GENDER and RACE, each of which is geared toward certain

[1] In addition to Jennifer Saul's (2006, 2012) proposals, which we will encounter below, there have been other discussions attempting to put forward a uniquely correct concept of gender (Stone 2004; Bach 2012).

Ingo Brigandt and Esther Rosario, *Strategic Conceptual Engineering for Epistemic and Social Aims* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020).

DOI: 10.1093/oso/9780198801856.003.0005

epistemic and/or social aims. We dub the development of a concept for a specific thought and action context—combined with an openness to employing another concept when pursuing other aims—*strategic conceptual engineering*. We will illustrate this approach by sketching three distinct concepts of GENDER, all of which are needed because each serves a different social aim.

By 'concepts' in general we mean components of a person's thought which can influence reasoning as well as action. Concepts have the latter capacity by means of embodying descriptive and/or normative beliefs (some of which may be implicit). This is not a purely externalist construal of concepts, that would locate all aspects of conceptual content outside of a person's cognitive life. Our reason for including and in fact focusing on narrow conceptual content is *methodological* (for a more detailed account, see Brigandt 2011). The project of conceptual engineering consists in adjudicating why adopting and using one concept is most conducive (or in any case more conducive than another one concept) for conducting a given intellectual or practical task. For this philosophical project, aspects of content that impact thought or action (such as descriptive or normative beliefs) have to be invoked, so that one is in a position to count a person's adoption of a novel concept or a modification to a concept as a change—and possibly an improvement—in their intellectual or practical capabilities. (The same holds for the study of scientific concepts addressed in the next section.) A purely externalist approach, in contrast, might acknowledge that a person's beliefs have changed, while maintaining that the concept (e.g., the externally established referent) has not. But then such an approach would use the notion of 'concept' such that it fails to actively address what is actually relevant for the purpose of conceptual engineering and conceptual ethics.[2]

Before turning to conceptual engineering in the concrete context of GENDER and RACE as concepts that have exercised philosophers, we begin with a look at scientific concepts.

Lessons from Scientific Concepts

Not only concepts used by philosophers are in need of improvement (as the agenda of conceptual engineering underscores), but also scientific concepts are routinely revised. Indeed, the investigation of concept change and semantic variation in biology that has been conducted by one of us (Brigandt 2010, 2011, 2012) can provide guidance for the philosophical project of conceptual engineering and conceptual ethics. The motivation for attempting to understand concept change in science is the following.

Scientists are never hostage to the definitions they once used, and upon the arrival of a novel theoretical perspective they may redefine a term. For example, with the advent of molecular genetics the CLASSICAL GENE concept gave rise to the significantly different MOLECULAR GENE concept (Brigandt 2010). The HOMOLOGY concept (referring to the corresponding structures in different species) was introduced well

[2] Our motivation for not using a purely externalist construal of concepts can be seen as an instance of conceptual engineering pertaining to the philosophical concept CONCEPT, as it adopts a certain notion of CONCEPT because it is more fruitful for the philosophical task at hand (Brigandt 2011).

before the advent of Darwin's evolutionary theory, and only later came to be given a definition that appeals to common ancestry (Brigandt 2012). Yet Kuhn (1962) and Feyerabend (1962) prominently claimed that a term being used within different theoretical frameworks amounts to 'incommensurable' concepts. Outside of philosophy of science, some have likewise deemed the changing of a term's definition as illegitimate by framing it as a "change of subject" (see also Cappelen 2018; Prinzing 2018). For instance, in their argument against a naturalistic reduction of consciousness, Chalmers and Jackson (2001) acknowledge that future neuroscientists may put forward what these scientists would call a naturalistic concept of CONSCIOUSNESS. But Chalmers and Jackson insist that this different concept would be irrelevant to their claims about consciousness involving the current concept of consciousness, where they do not allow for there being good reasons to replace the concept with a revised one (Brigandt 2013).

Thus, the question is how it can in fact be legitimate for scientists to revise a concept. We do not have a clear answer to the notoriously difficult question of how to individuate concepts. But this does not matter for our purposes, because regardless of whether a change is framed as an enduring concept being modified (in features not affecting its identity) or as one concept being replaced by a different concept associated with the same term (see also Richard, Chapter 17, this volume), the project is to understand why the new concept is the better one to use. To do so, the crucial move is to view a concept as being used by scientists to pursue a *specific scientific aim*, because this aim—Brigandt (2010) dubs it the 'epistemic goal' of a concept—sets the standards for whether one definition of a concept is superior to another definition. In line with the aims that individual scientists pursue, such an epistemic aim is not a global aim of science (e.g., putting forward explanations), but a local aim that is pursued by only some scientists in a certain context, for example, explaining programmed cell death, or having the ability to experimentally manipulate molecular pathways involved in cell-cell signalling.

The novel philosophical claim is to point out that there are cases where a scientific aim can be tied to an *individual concept* in that this concept is being used by scientists to pursue this aim. For example, while the CLASSICAL GENE concept was used for the purpose of predicting (and statistically explaining) phenotypic patterns of inheritance across generations, the MOLECULAR GENE concept serves the aim of causal-mechanistically explaining how inside a cell a gene leads to the formation of its molecular product. Making explicit such an aim tied to a concept's use permits one to account philosophically for the rationality of concept change: a revised definition is an improvement over an earlier definition if the former is empirically more conducive to meeting this aim (Brigandt 2010).[3] While concepts are typically construed as beliefs about the concept's referent and philosophy of science has traditionally articulated various *representations of natural phenomena*, the inclusion of scientists' investigative, classificatory, or explanatory aims to one's philosophical study of

[3] While this account was developed independently of work outside of philosophy of science (and more explicitly emphasizes the philosophical role of aims), it aligns with Haslanger's ameliorative project about philosophical concepts: "on my view, whether or not an analysis is an improvement on existing meanings will depend on *the purposes of the inquiry*" (Haslanger 2005: 24, emphasis added).

scientific concept use adds an additional dimension, as aims are *values pertaining to what human practice is to achieve.*

In addition to diachronic conceptual differences (e.g., concept change in the history of science), another common situation is synchronic semantic variation across different scientists. The latter can likewise be captured by a philosophical framework emphasizing the notion of epistemic aims (which also provides guidelines for philosophical studies of concepts outside of science). For example, genes do not form a unique structural kind, resulting in a plethora of different (although overlapping) contemporary uses of the term 'gene,' where even one and the same biologist can use the concept differently in different contexts. Such semantic variation can be philosophically understood and seen to be conducive to scientific practice in case distinct uses of a term serve somewhat different scientific aims. In the case of the MOLECULAR GENE concept as used currently, one can make the case that there is still one *generic* epistemic aim shared across all uses of the term, while individual uses may differ in the *specific* epistemic aim pursued by a person in this context (Brigandt 2010). Another case in point is that biologists use various SPECIES concepts, which cross-classify organisms. Philosophers have argued that a plurality of SPECIES concepts is needed to serve different scientific purposes (Kitcher 1984). The conceptual diversity in biology suggests a first lesson for the philosophical project of conceptual engineering. It may well be that one definition cannot serve all the legitimate aims associated with a philosophical notion, so that philosophers should be open to a *pluralism* that permits different concepts, each of which is most conducive in some context (see also Burgess and Plunkett 2013b).

Apart from philosophers accounting for conceptual differences in biology in terms of scientific aims, there are cases where the scientists themselves are aware of the fact that their semantic differences are due to a concept being used for different purposes by different approaches or different biological fields—the diversification of the HOMOLOGY concept during the last three decades is a case in point (Wagner 1989; Roth 1991; see also Brigandt 2003). But in general, the aims for which scientific concepts are used and reshaped are only implicit in scientific practice, which also results in instances of biologists talking past each other when being oblivious of the fact that their semantic disagreement is due to individually legitimate, yet different scientific aims being at stake. This yields another lesson for conceptual engineering. It can be beneficial to explicitly articulate the aim that a philosophical concept is to serve, in particular if others implicitly use the same term for quite different purposes. Subsequently showing that one's proposed concept serves this purpose better than traditional alternatives is no easy task, given that a concept may be meant to serve several related aims, but being explicit about the aims at stake is a fruitful first step.

Philosophy of science accounts of concept change have appealed to epistemic aims and values, rather than social aims and values. But this should not be construed as an endorsement of a dichotomy between epistemic aims and social aims. Indeed, even though some do construe 'epistemic' narrowly as only those features pertaining to truth (Douglas 2009), the above examples of epistemic aims included explanatory aims—which go beyond aiming at truth given that not every true scientific representation is explanatory of a phenomenon and given that causal and mechanistic explanation is also related to the practical ability of effectively intervening in nature.

More generally, the specific aims pursued by scientists typically combine more epistemic and more practical considerations, from pragmatic constraints on research (e.g., not pursuing a question that one cannot answer given data and computational limitations) to intended applications that implicate pharmaceutical or environmental aims (Kitcher 2001). This *entwinement of the epistemic and the social* also holds for scientific concepts and categories, which can implicitly embody non-epistemic values, or at least their use can be criticized based on such values (Brigandt 2015). That said, many general discussions of scientific concepts have tended to focus on more epistemic aims, so it is now time to turn to concepts that explicitly serve social-political aims.

Engineering Concepts in the Context of Biosocial Phenomena: Gender and Race

Among the concepts that answer not only to empirical facts and epistemic aims but also to social-political aims, our discussion focuses on GENDER and RACE. We do not so much endeavour to put forward a clearly articulated concept of GENDER or of RACE (and later on will also argue that a plurality of GENDER and RACE concepts is needed). Instead, our primary agenda is to offer methodological remarks on how to engineer these concepts, with implications for other concepts relevant to philosophers engaging in conceptual engineering and conceptual ethics. Before also addressing the notion of RACE, we begin with GENDER.

Sally Haslanger (2000)—while explicitly acknowledging that there may be further concepts of GENDER—prominently put forward the following definition of WOMAN (we do not restate her analogous definition of MAN):

> S *is a woman* iff
>
> (i) S is regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction;
>
> (ii) that S has these features marks S within the dominant ideology of S's society as someone who ought to occupy certain kinds of social position that are in fact subordinate (and so motivates and justifies S's occupying such a position); and
>
> (iii) the fact that S satisfies (i) and (ii) plays a role in S's systematic subordination, that is, *along some dimension*, S's social position is oppressive, and S's satisfying (i) and (ii) plays a role in that dimension of subordination. (Haslanger 2000: 42)

One motivation for this definition is prior scepticism about the possibility of putting forward a coherent concept of GENDER. SEX and GENDER have traditionally been distinguished, with sex pertaining to biological features, while gender roughly being the social meaning of sex. But it has been questioned whether sex is independent of social construction (Butler 1990, 1993). And the idea of gender being constituted by certain social features may commit one to postulating a sort of social essence of gender, which is unacceptable given that different women can face different social expectations, occupy different social positions, and can have radically different social experiences (as shown vividly by considerations about intersectionality). By falsely maintaining that all women share certain social properties, any definition of 'woman' would effectively privilege certain forms of femininity while excluding others. This casts doubt on whether it is advisable for a feminist to ever put forward a specific definition of 'woman,' yet at the same time it is unclear *who* the target of feminist activism is if it does not pertain to women, as a socially construed category (Alcoff 1988; Heyes 2000; Stone 2004).

By not assuming that women must have a unique, shared psychological identity or identical social experiences, Haslanger's above account circumvents these problems and instead focuses on the point that, because of their gender, women tend to occupy disadvantaged social positions. This provides not only an idea that can unite feminist activists, it more generally is a concept that is geared toward addressing legitimate social aims. Haslanger's concept of GENDER and her proposed concept of RACE (which likewise highlights discrimination and differential social privilege) are meant to be "effective tools in the fight against injustice," and their development was guided by the "need to identify and explain persistent patterns of inequalities between females and males, and between people of different 'colors'" (Haslanger 2000: 36).

At the same time, Haslanger's account has been criticized for maintaining that women are *by definition* those female persons who are subordinated. Jennifer Saul (2006), responding to Haslanger (2006) and her earlier writings, charges that Haslanger's account is not even acceptable to feminists as one of her primary target audiences (see also Mikkola 2009). Haslanger favours a society where there are no subordinated women, that is, no women as defined by her. She thus is seen to commit herself to the social goal of the "elimination of women," a goal when stated in these terms "may well alienate many of those who would be most likely to support anti-oppression movements" (Saul 2006: 139). Saul's central worry is that using a term like 'woman' with a meaning that departs from standard usage is bound to create confusion. Consequently, in this early commentary, Saul (2006) prefers to use 'woman' and 'man' as denoting biological sex (in line with what she considers ordinary use and devoid of social factors like subordination), indicating that she does not think that ordinary speakers even use 'gender' as a social category that is distinct from 'sex.'

To our minds, Saul's perspective has problems, a discussion of which offers lessons for the project of conceptual engineering beyond the case of gender. First, Saul's criticism of Haslanger focuses on terminology, on how to linguistically use a given term, and on communication across speakers (and the same holds for Mari Mikkola's 2009 objections to Haslanger). This ignores the question of what *concepts* to use, where concepts are vehicles of thought—broadly understood as also motivating action—that can be employed by an individual regardless of communication by means of language. The issue is what overall set of concepts are needed within a larger community of agents (given their diverse legitimate aims), and by implication, what novel concepts may have to be developed. In their proposal of the project of conceptual ethics, Burgess and Plunkett (2013a) rightly assume that questions about what concepts to employ and questions about what terms to use to express them linguistically are separate issues.[4] This is also something that Haslanger recognized:

[4] "That we ought to use concept C does not yet settle how we should lexicalize it, for example" (Burgess and Plunkett 2013a: 1095).

> At this point we should bracket the terminological issues and just consider whether the groups in question are the ones that are important to consider given the goals of our inquiry.
> (Haslanger 2000: 46–7)

In our view, considerations about concepts are vital—even when the issue of terminology is disregarded—given that "our conceptual repertoire determines not only what we can think and say but also, as a result, what we can do and who we can be" (Burgess and Plunkett 2013a: 1091). Indeed, as the previous section on scientific concepts has hinted at, a concept is meant for certain intellectual or practical purposes and thus to be used *in some contexts only*. A community and even one person can fruitfully employ a variety of different concepts depending on the context. This makes room for the possibility that Haslanger's novel (or some similar) *concept* of GENDER is fruitful, at least when the identification and explanation of gender-based discrimination is at stake, even if other cognitive contexts require different concepts of GENDER.[5] Consequently, our recommendation is to first address questions about what concepts are fruitful, and only then to discuss the (admittedly difficult) question of what terms to use to express them, for example, whether a newly engineered concept should be associated with an existing term, to which another concept is already tied, or whether to introduce a new term for this novel concept (see also McPherson and Plunkett, Chapter 14, this volume). (We will comment in the concluding section on what terminology may be best for public discourse despite our pluralistic approach to GENDER concepts.)

Second, even when side-stepping the domain of concepts and moving on to considerations about communication, while Saul is rightly worried about the possibility of misunderstandings, this may lead to a problematic preference for sticking to the common use of terms:

> It seems to me that communication is difficult enough as it is, and that we should instead try to use ordinary terms in as ordinary a way as possible. (Saul 2006: 141)

Although there are serious questions about how one can realistically effect a widespread change of common term use and when it is wise to attempt this, Saul's bias toward the status quo may be particularly precarious for *ordinary* terms, given that in some cases such term uses are highly problematic and should be reformed, for example, the past practice to use 'rape' so as to not consider marital rape as such. In the context of concepts, such a bias is at odds with the very spirit of conceptual engineering and conceptual ethics, which endeavours not only to introduce completely novel concepts, but also to abandon or transform existing concepts to the extent to which they are empirically flawed or unsuited for their intended purpose (for other reasons to discard a folk concept, see McPherson and Plunkett, Chapter 14, this volume).

[5] Although Saul (2006) acknowledges the possibility of using the term 'woman' with different meanings in different contexts, she rejects this option on grounds of misunderstanding in *communication across contexts*, and thereby fails to evaluate whether using Haslanger's concept can be fruitful in *some contexts of a person's thought and action*.

We will get to the issue of concepts of GENDER and RACE embodying problematic empirical assumptions soon, but first we address the political purpose for which Haslanger puts forward her ameliorative analysis. Saul does have this issue in view, and argues that even with her preferred option of using 'woman' and 'man' as denoting *biological* sex (and thus as synonymous to 'female' and 'male,' respectively), a *social* explanation of inequalities is possible:

> The explanation of persistent inequalities between females and males would begin from the fact that females tend to be systematically subordinated to males. The next step would be to analyse this systematic subordination and how it is perpetuated. (Saul 2006: 136)

However, Saul's approach construes gender as a purely biological category and views pointing to the social factors underlying gender-based discrimination as a second step and thus as a *separate* issue.[6] In contrast, Haslanger's concept of GENDER rightly highlights that despite the biological aspects of gender, a proper understanding of the matter has to explicitly include that the gender differences at hand are due to social factors. There are many biosocial kinds that combine a mutual causal influence of more biological and more social factors. And if the *purpose* of a concept referring to such a kind is to be explanatory, a properly engineered concept has to include in its content all the relevant causal factors and how they are entwined.[7] Thus, for the purpose of accounting for inequalities across the sexes, Haslanger's concept of GENDER is superior to the concept SEX or a biological concept of GENDER as the one preferred by Saul (2006), even if Saul can in fact provide the social explanation by adducing additional (non-biological) concepts.

A case where not only the entwinement of more biological and more social factors matters to social-political aims, but where ordinary concepts can also embody problematic empirical assumptions, is *race*. Historical and some current conceptions of RACE assume racialism, in particular the idea that there are behavioural and cognitive differences between human races and that these are due to biological differences. Given that this is empirically false, Appiah (1996) has prominently argued that there are no races—because there is nothing in reality corresponding to this ordinary concept of RACE (see also Zack 1993). Others have instead opted for using a different concept of RACE that discards false beliefs, for example, one that focuses on the cultural construction of race. As Mallon (2006) points out, the philosophers endorsing eliminativism, constructivism, and other rival positions on

[6] Mikkola (2009: 581) says something similar: "There is nothing in my proposal that prevents feminists from identifying and explaining persistent inequalities since this is a separate empirical task." However, she never explains what this empirical task is separate from—certainly it is not separate from Haslanger's agenda of providing a concept of 'woman' that identifies and explains persistent inequalities.

[7] In the introduction we justified in general terms why a strongly externalist construal of concepts does not work for our agenda, and instead concepts have to be construed as embodying relevant empirical (or normative) beliefs. Yet a much deeper failure to appreciate that conceptual engineering is not the acquisition of novel empirical knowledge by means of analysing current concepts, but the incorporation of relevant content into a concept using *previously* acquired empirical knowledge, can be found in Mikkola (2009: 580): "I am unconvinced that politically much, if anything, hangs on analysing [sic!] the concept *woman*. In fact, to claim that by discerning the conditions for satisfying *woman* the content of feminist policies is discovered, strikes me as extremely implausible."

race actually agree on the basic empirical facts, and the preference for different positions hinges on normative considerations (see also Ludwig 2014).

Given our focus on the conceptual engineering aim of capturing race-based discrimination, what we want to highlight is the insight of Kaplan (2010) that even if there are *biological* differences among races in a country, these may well be due to *social* factors (see also Kitcher 2007: 315). His starting point is the partial success of race-based medicine, for example, the drug BiDil recommended as a treatment to prevent heart failure in African Americans. This suggests the existence of medical and thus biological differences across some races, given that differential drug efficacy presupposes different physiological features in different individuals. Yet Kaplan argues that even if there are such differences, this does not entail that they are due to alleged genetic differences across races (as shown by the different health of African Americans and recent immigrants from Africa). Instead, in a context like the US, the physiological and health differences may very well be due to social discrimination, which affects diet, exposure to pollutants and toxins, and levels of stress experienced (a risk factor for hypertension and heart disease). A further point of his is that the label 'race-based medicine' erroneously suggests that this is to be dealt with as a medical (rather than primarily social) problem. But what matters most for our purposes is that the example of race reveals that some philosophical concepts that aim at explaining social differences (or at exposing social discrimination) have to capture the reciprocal influence of social and biological features.

A common motivation for overlooking the connection between the social and the biological is the spectre of biological determinism. Some discussions of sex and gender are also predicated on a nature-nurture dichotomy. The latter assumes that as one layer there is nature, which is unaffected by how nurture builds on it as a second layer. The dichotomy between the two alleged layers also takes for granted that it is meaningful to claim either that nature constrains nurture or that nurture is largely unconstrained by nature. The nature-nurture dichotomy, however, is empirically false, and many biologists have abandoned it on the more profound ground that it is a conceptually flawed and theoretically useless perspective (and instead have come to emphasize the idea of phenotypic plasticity and other notions capturing the entwinement of organismal development and environment; Pigliucci 2001).

How to promote concepts that avoid false empirical beliefs (such as a nature-nurture dichotomy or biological determinism) among the general public is a difficult matter. Indeed, when anthropologists use the term 'race' at all, they usually use it in the context of cultural identification, as they are rightly worried that using 'race' in connection with biological (or at least genetic) features will trigger illicit connotations. Nevertheless, the lesson for the philosophical project of conceptual engineering is more obvious. Gannett (2010) points out that even discussions of race by philosophers have tended to be conducted in terms of social constructivism vs. biological realism (and a separation of the conceptual schemes of experts and laypersons), so as to result in a problematic "dichotomization of the biological and the social" (p. 368). While there are legitimate concepts of RACE that focus on cultural factors, our discussion about biomedical differences across races being due to social discrimination shows the need for some concept of RACE that captures in an empirically adequate fashion how social and biological factors are in reality interrelated. The beliefs about race currently held across the general public of course fail to reflect this emerging empirical account, but the fact that ordinary beliefs are not excluded from being empirically flawed and potentially harmful—and may even be particularly prone to be so, as racialist conceptions of race show—underscores yet again the problem with Saul's (2006) recommendation of trying to use concepts with their ordinary meanings.[8]

Gender and the Inclusion of Trans Persons

Beyond the worry that a revised concept of GENDER like Haslanger's departs from ordinary use (which above we found not to be compelling), there are important issues to be addressed. For recent discussions of gender have emphasized the need to ensure that the concept WOMAN include trans women (Bettcher 2013, 2017a; Kapusta 2016)). This is not only a problem for Haslanger's account (as we will address below), but Jennifer Saul likewise recognizes that her earlier approach (in Saul 2006) failed to include trans women because it favoured using 'woman' to denote those persons who are biologically female.[9] Now Saul (2012) favours a contextualist approach, which (in addition to an analogous definition of MAN) construes the concept WOMAN as follows:

> X is a *woman* is true in a context C iff X is human and relevantly similar (according to the standards at work in C) to most of those possessing all of the biological markers of female sex. (Saul 2012: 201)

In some contexts, the standards may indeed entail that the relevant similarity pertains to biological sex, but Saul discusses other similarity conditions. In this fashion, she can capture persons with intersex conditions as well as trans persons, where the latter fall under the gender category which with they identify.

At the same time, Saul has some misgivings about this contextualist approach, on the grounds that it is too flexible and may permit using WOMAN with any meaning. For instance, there are situations where biological and other similarity conditions are operative according to which trans women are not women. Given this, Saul worries that this makes a trans woman's assertion that she is a woman true only in *the specific context* in which she makes this utterance, and thus true in an utterly trivial fashion (as opposed to it being a substantial, context-independent truth about this women's gender in line with her self-perception). The flipside of this issue is that also the

[8] Another instance of concepts embodying empirically flawed assumptions (regardless of whether this always aligns with ordinary meanings) is Saul's suggestion to use 'woman' and 'man' as denoting biological sex. This assumes that these are the only biological categories and that there is a unique account of biological sex. Yet the diverse human conditions collectively called 'intersex conditions' show not only that sex is not a binary category (and instead a spectrum), but also that sex-related biological features such as genes, chromosomes, hormones, internal/external sex organs, and secondary sex characteristics do not always align, where even one person can possess chromosomally male and female cells (Ainsworth 2015; see also Kessler 1990; Butler 1993).

[9] While we previously objected to Saul's (2006) desideratum to use terms with their established, ordinary meanings, we also do not think that ordinary use would support Saul's earlier, exclusively biological construal of WOMAN, given that laypersons' actual use is more diverse and flexible, which often includes trans women.

statements of those lawmakers who claim trans women to not be women (and who pass laws that prohibit trans women from using women's washrooms) seem to come out as true. Saul's worry is that one may have to grant that in *these lawmakers' context* the standards are such that their biological construal of WOMAN obtains.

Saul sees the option of insisting that her own standards, rather than those of transphobic lawmakers, are the right ones. Yet she frames this approach as follows:

> On my view of "woman," I cannot argue that the lawmakers are making a mistake about how the word "woman" works. But what I can do is argue that they are morally and politically wrong to apply the standards that they do.... But we must recognize this claim for what it is: a moral and political, rather than merely linguistic, claim. (Saul 2012: 204)

Saul correctly recognizes that moral and political considerations are involved; however, she seems to view them as separate from and something over and above language use. A dichotomy between concept use and moral-political values should not only be alien to philosophers engaged in the project of conceptual ethics (see also Plunkett 2015; Díaz-León, Chapter 9, this volume; McPherson and Plunkett, Chapter 14, this volume). As we have explained above, also scientists revise their language use and disagree on how to use a scientific term in part because of the values and scientific aims they favour, which can include environmental, application-related, and other practical aims. Furthermore, there is the widespread phenomenon of ordinary discourse where a disagreement is not about the object being talked about, but about what concept should be used to describe the object. This negotiation about what particular concept (e.g., what construal of SPICY) to use in a certain context has been dubbed 'metalinguistic negotiation' by David Plunkett and Tim Sundell (2013); and they argue that metalinguistic disagreements can be *substantial* disagreements (as among other things they influence action, e.g., whether to add more spice).[10] Metalinguistic negotiation is linguistic (or about how a certain word works, as Saul puts it), but at the same time based on aesthetic, moral, or political values.

In her response to Saul's discussion, Esa Díaz-León (2016) rightly argues that opting for the employment of moral and political considerations regarding the use of 'woman' can also be seen "as a genuinely linguistic point" (p. 248). The standards of similarity that are at work in a given context matter for Saul's contextualism, yet "questions about the relevant standards might involve moral and political considerations (and in cases that concern the rights of trans women, will very likely do so)" (Díaz-León 2016: 249). Díaz-León points to the distinction between attributor factors and subject factors that has been invoked in epistemic contextualism. In a situation where person *A* wonders whether or not subject *S* counts as knowing that *p*, attributor contextualism maintains that this is contingent on the context of knowledge attributor *A*, that is, attributor factors determine the standards. In contrast, some have objected that instead it had better be subject factors (the context of

[10] In the 'Lessons from Scientific Concepts' section we critically mentioned Chalmers and Jackson's (2001) tenet that it is illicit to associate a term with a different concept (a situation which they frame as a "change of subject"). Yet such metalinguistic negotiation can very well be legitimate; in fact, the target of Plunkett and Sundell's (2013) criticism is the idea that meaningful disagreement always presupposes that two speakers use the same concept (and make different claims with it).

putative knower *S*) that are relevant for setting the standards of knowledge. Díaz-León suggests that the latter option—subject contextualism—should be used for the concept WOMAN, in which case it is not the lawmakers' context (or the context of anyone else who may make claims as to whether a trans woman as a woman) that matters, but the trans woman's context:

> according to subject-contextualism about *woman*, "X is a woman" is true iff X is human and relevantly similar to most females, where what counts as relevantly similar to most females depends on "objective" features of X's context, including instrumental, moral, and political considerations having to do with how X should be treated (regardless of who utters the sentence or what their beliefs are). (Díaz-León 2016: 251)

Díaz-León's approach strikes us as basically correct, though we add in way of clarification that when person *A* uses a concept that refers to another person *S*, it cannot always be the case (or be an *a priori* matter) that it is the latter person's context that is the relevant one. And to the extent that *S*'s context matters, it need not be her beliefs or values that settle the issue. For instance, a trans woman may happen to have a conception of WOMAN that diverges from the conception of other trans women (and other women in general) and an adoption of which would be morally and politically problematic in the given situation. Instead, in the case at hand (positions taken by lawmakers about trans women), what is at stake is whether trans women are treated equally to other women—an issue that *also* follows from the context of a transphobic lawmaker (the attributor context). Regarding the question of equal treatment, the proper social-political aim is that trans women are accorded the same rights as other women (regardless of whether given their personal values some lawmakers or trans women may disagree). Thus, the lesson is that the philosophical focus should be on the *relevant and legitimate aims* (which may not be the aims of either the attributor or the subject of an utterance), which must be identified and have to be justified as legitimate (even if they already happen to be agreed upon). We shall use this advice in the remainder of the section, in which we also highlight that a particular concept WOMAN has to be geared to a specific aim.

The desideratum that trans persons ought to be fully included by a philosophical account of gender has recently been reiterated by Katharine Jenkins (2016). Her critical target is Haslanger's earlier ameliorative concept WOMAN, and Jenkins argues that it excludes trans women by counting many trans women as men. The definition by Haslanger requires of a woman that the person be "regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction" (Haslanger 2000: 42), yet Jenkins points to several cases where some trans women would not fall under this definition: she may be respected as a woman regardless of not being deemed to have a female's reproductive role, she may not publicly present as a woman, or—particularly problematically—her public gender presentation is not respected by others as being a woman. To remedy this issue, Jenkins makes the useful distinction between two relevant concepts of GENDER. While the first one, GENDER AS CLASS, is basically Haslanger's ameliorative concept, Jenkins puts forward her account of GENDER AS IDENTITY, which is to properly include trans persons. In the next section, we will say more on the way in which both concepts of gender are needed, but now focus on the latter.

Jenkins puts forward the following definition of GENDER AS IDENTITY (for the category of woman):

> S has a female gender identity iff S's internal 'map' is formed to guide someone classed as a woman through the social or material realities that are, in that context, characteristic of women as a class. (Jenkins 2016: 410)

Although Jenkins makes the crucial claim that "this definition of gender identity entails that all trans women have a female gender identity" (p. 413), in our view this entailment claim is highly dubious, given that Jenkins articulates her notion of an 'internal map' as follows:

> On my definition, having a female gender identity does not necessarily involve having internalized norms of femininity in the sense of accepting them on some level. Rather, what is important is that one takes those norms to be relevant to oneself; . . . Her experience of social and material reality includes navigating the norm that women should have hairless legs, even though she is not complying with it. (Jenkins 2016: 411)

But even taking gender norms relevant to oneself (without necessarily accepting them) is too strong a condition. On Jenkins's account, a trans woman who happens to be oblivious of the norm that women should shave their legs is not a woman (and the same holds for cis women who are unaware of this particular norm). And from the perspective of the project of conceptual engineering and conceptual ethics, why would the inclusion of norms like shaving one's legs be relevant to a trans-inclusive concept WOMAN in the first place?

The problem is not just that with leg-shaving Jenkins has used a poor example of a relevant norm. Making explicit that this norm is meant for gender identity in Western cultures or even using a different social norm would not help. In our view, the whole point of putting forward a concept GENDER that is trans-inclusive is the social-political aim that trans women (and also cis women) are socially accepted as the gender with which they identify, and have all the moral and legal rights corresponding to their gender. One's (sincere) self-identification as a woman should suffice for this, and one should not make social and legal recognition as a woman contingent on being aware of any gender-related norm, as Jenkins does. As an analogy, take party membership, for example, being a Republican in the US. Of course, Republicans tend to have many factual beliefs, political values, and social practices in common that form their political identity, and these shared features are important to sociologists and political scientists for predictive and explanatory purposes (predictions and explanations that are useful even if there are some Republicans that do not conform to the stereotype). However, Republicans who are untypical in terms of their beliefs and values can still be party members and have concomitant rights. For instance, being registered as a Republican voter (according to state guidelines) suffices for being eligible to vote in a Republican primary—no specific ideological test is required to possess this voting right.

In the case of gender, the fact that persons of one gender tend to have similar internal maps and are aware that certain social norms and gender-related expectations apply to them is likewise relevant for explanatory purposes, and is an aspect of social-psychological gender identity that may indeed be relevant for *some* concept of

GENDER. Endorsing such social norms is required by a strong ideological test, while merely being aware of the existence of these gender-related norms (as Jenkins envisions) can be dubbed a weak ideological test. However, for the purpose of a concept of GENDER (or GENDER IDENTITY) that has the purpose of *guaranteeing gender-appropriate social recognition and rights* (and that can usefully complement Haslanger's concept of GENDER AS CLASS), not even a weak ideological test should be required to count as a woman (or as a man). In addition to the concern of excluding trans women, recent female immigrants from other cultures may not be aware of various gender-related social expectations (e.g., shaving one's legs in a Western culture) and not have them as part of their current internal map, yet this should be no grounds to deny them social recognition as women and the rights that other women have. In summary, for the specific social-political aim of achieving gender-appropriate recognition and rights, we advocate using a concept of GENDER that defines someone to be a woman or man solely in terms of this person's sincere self-identification with a particular gender, for example, as long as someone identifies as belonging to the gender commonly denoted by the term 'woman.'[11] More precisely, in many legal contexts 'man' and 'woman' are the only two categories, while for the purpose of social recognition more than two genders may well be needed—but in this case, the issue is also a person's mere self-identification with a gender.

Strategic Conceptual Engineering

It is useful to view a concept as a tool that serves certain epistemic, social, and other aims (Brigandt 2011; Burgess and Plunkett 2013b; Prinzing 2018; McPherson and Plunkett, Chapter 14, this volume). While such aims are sometimes implicit in how a concept is used and how it is revised, for the purpose of deliberate and effective conceptual engineering our discussion has repeatedly emphasized the need to articulate the aim at stake and to ensure that a concept employed actually embodies those empirical facts that are conducive to this aim. This is admittedly easier said than done. One hurdle is justifying that something is a *legitimate* aim, and to convince others of its importance. Not only were, in the past, social aims deemed to be legitimate that are nowadays repudiated, but disagreement about what the relevant aims are exist even within contemporary academic subcommunities. While we discussed several scholars (Saul, Díaz-León, and Jenkins) who endorse the desideratum that a concept of 'woman' include trans women, there are unfortunately feminists who oppose this based on a trans-exclusive stance on feminist identity politics (Heyes 2006; Watson 2016; Bettcher 2017b; McKinnon 2018).

Moreover, it can be difficult to assess whether several related aims better be viewed as separate in that they cannot all be fully met by one concept. For instance, when putting forward her concept of GENDER, Sally Haslanger (2000) singles out the need

[11] McKitrick (2007: 141) wonders about a proposal like ours: "Perhaps having a female gender identity supervenes on the psychological property of having a strong and persistent belief that you are a woman. But again, that merely pushes back the question: What exactly is it that you believe about yourself?" Here our analogy with party membership is instructive, where all that is cognitively *required* is that one identify as belonging to a collection of individuals picked out by such a *label* as 'Republican' or 'woman.'

to “identify and explain persistent inequalities” (p. 36). But these could be viewed as two distinct aims, and a concept that merely identifies patterns of discrimination may not embody the relevant psychological and social factors that actually explain these patterns. More importantly, beyond description and explanation—which could be viewed as largely epistemic aims—Haslanger explicitly adopts the social-political aim of fighting oppression, by means of providing conceptual “tools in the fight against injustice” (p. 36). But as Jennifer Saul’s (2006) criticism highlights, if those in subordinated positions rely on a concept that offers a causal explanation of systematic discrimination (even when it highlights the social nature thereof), this may very well motivate them to “become trapped in a feeling of powerless to change their own fates” (p. 138). This raises the difficult question of whether the aim of explaining discrimination and of opposing discrimination have sometimes—*though not always*—to be addressed separately when engineering concepts. In addition to identifying and explaining inequalities, Haslanger lists the “need for accounts of gender and race that take seriously the agency of women and people of color of both genders” (2000: 36), but it is doubtful that her definitions of GENDER and RACE actually capture the relevant agency, so this may be yet another aim that requires a separate treatment. We will return to this difficult issue below by opting to separate certain social aims insofar as quite different concepts of GENDER are needed to address each of them.

Regarding the question of how to construe race, it is clear that a traditional biological conception that would posit significant phenotypic differences between races and take them to be due to genetic differences between races is empirically false (Rose et al. 2009; Hochman 2013; Pigliucci 2013; Yudell et al. 2016; Spencer 2018a). Moreover, grouping humans into races is not even a major classificatory or explanatory aim for contemporary population genetics (Templeton 2013; Kopec 2014). A useful concept of race will therefore ensure to include the social factors that influence how persons come to be deemed to be of a certain race (e.g., persons of biracial ancestry considered to be black), how racial self-identity is generated, how a race is perceived by others, and how persons from different races are treated within society, so as to track the social dynamics of race. Despite the largely cultural nature of race—which suffices for many epistemic and social purposes where a concept of race is employed—our discussion has pointed out that biological features enter not only in the trivial sense of skin colour and other phenotypic features being socially conceptualized. The way that social discrimination can have actual physiological effects that result in predispositions to disease matter if the aim is to offer a thorough explanation of social dynamics, and at least when an explanation of health disparities is sought after. A concept of race that highlights social impacts on biological features is also conducive to the social aim of reducing discrimination, and of opposing erroneous biological determinist conceptions (which have been promoted by the idea of race-based medicine).

To be sure, these race-related health disparities hold for the context of the United States; and philosophical commentators have indicated that race operates differently and has another overall prominence in many other countries (Alcoff 2006), for instance, when even in biomedical contexts the notion of ETHNICITY is more relevant (Gannett 2010; Ludwig 2014). The best concept of RACE to employ in some

Caribbean or South American contexts such as Cuba or Brazil may well be different from one geared toward conditions in the US.[12] Different societies may not only *empirically* differ in their social processes, but their different identity politics concerns may result in different legitimate *aims* that a particular concept of race has to serve. Thus, it appears that there is no unique concept of race that could capture all cultural contexts and serve all epistemic and social aims.

Some may wonder whether a concept that includes *all* biological and social features and distinguishes how they operate in different cultural (or historical) contexts can function as an all-purpose concept after all. We doubt that this is an option even for a concept of RACE (see also Hardimon 2017; Spencer 2018b). One reason is that relative to one explanatory aim, a concept should only include what is explanatorily relevant and thus not adduce any feature (no matter how true) that is irrelevant to the phenomenon to be explained. Therefore, an explanatory concept must exclude some aspects of reality, which however have to be included with respect to some other explanatory aim. In the context of scientific concepts we have already covered that biologists use, for instance, different species concepts because these answer to different concrete biological aims.

In any case, a strong case against all-purpose concepts can be made in the case of *gender and political aims*. We call not only for a plurality of concepts of GENDER, but for what we dub *strategic conceptual engineering*, which is the strategic employment of a concept for certain epistemic or social aims, combined with the understanding that this concept has a limited scope of application and the openness to use another concept with respect to other aims. In what follows, we distinguish three different social-political aims, and argue that different concepts of gender are needed to meet each aim, respectively.

(1) Identifying and explaining gender-based discrimination. Haslanger's definition has been criticized, where the most noteworthy point has been that it fails to count trans women as women. Yet this tension can be resolved if in line with the idea of strategic conceptual engineering one views Haslanger's concept as restricted to particular aims.[13] At the very least, Haslanger's proposal, which highlights the oppression of women (due to being women), was calculated to be strategically useful when she put it forward, given that Haslanger reacted to scepticism within the

[12] For similar reasons, race and gender cannot be treated in a completely analogous fashion (Heyes 2006). The situation that both race and gender are widely held to be social constructions should not obscure the fact that they do not socially function in the same way even within the same sociohistorical context. In North America, gender is often treated by theorists as an individual property of the body or of a person (centering on one's psychological identity and social experiences), whereas race clearly implicates ancestry and cultural history. In such contexts, a properly engineered concept of RACE would do well to take heritage into account (while making room for the possibility of some people adapting a racial reception that disavows or ignores a part or parts of their ancestry).

[13] When distinguishing the concepts GENDER AS CLASS and GENDER AS IDENTITY, Jenkins (2016) charges that Haslanger, although acknowledging different concepts of GENDER, still privileges her ameliorative account (i.e., GENDER AS CLASS). This strikes us as false, given that Haslanger made plain that she views any concept of GENDER as tied to certain purposes: "Let me emphasize at the beginning that I do not want to argue that my proposals provide the *only* acceptable ways to define race or gender; in fact, the epistemological framework I employ is explicitly designed to allow for different definitions responding to different concerns" (Haslanger 2000: 36).

feminist scholarly community about the relevance of even articulating a concept of WOMAN (Alcoff 1988; Heyes 2000; Stone 2004), by providing a conceptual tool for the aim of uniting feminist activism. An approach similar to Haslanger's is in our view also fruitful for the aim of identifying gender-based discrimination or of explaining the social workings of this discrimination (so as to indicate how it could be remedied), provided that the following constructive suggestions of ours are being used.

First, while Haslanger's ameliorative definition—and the subsequent criticism—is actually about the concept WOMAN (and MAN), we suggest that (with respect to the aim of identifying discrimination) the focus should be on the concept GENDER, of which one can offer a revised, ameliorative definition without also redefining WOMAN and MAN (or without establishing any bold entailments about WOMAN and MAN). An advantage of this is that it shifts the focus *away* from questions about the *extension* of concepts, such as the question of whether there are females that are not discriminated against and thus not women on Haslanger's definition.[14] Sometimes the specific extension of a concept does matter—as in the case of a concept of WOMAN that attributes rights and social recognition to those classified as women (to which we turn below). But concepts may have *other* virtues (e.g., exhibiting discrimination's systemic nature) *than classificatory functions*. An account in the present context should be open to the boundaries of different genders being vague, to some persons being of more than one gender, and to there being more than two genders. An account of GENDER is definitely not committed to any definitive specification of these matters (while substantial expectations may obtain for a proposed definition of the two terms 'woman' and 'man'), and can even make room for the number of different genders varying across cultural history. Beyond the (immediate) aim of identifying discrimination, our approach would also be advantageous for Haslanger's ultimate aim of eliminating such social gender differences, as this agenda is now framed in terms of the elimination of "gender" rather than in terms of the elimination of "women."

Second, we advocate the use of *relational accounts* of GENDER (and RACE), in which the construal of one gender makes reference to other genders or to the properties used to characterize other genders (and in which different races are understood in relation to each other). In contrast, the proposals by Saul we have mentioned, including her contextualist definition of WOMAN (Saul 2012), do not make reference to MAN (and the same holds for Jenkins' 2016 construal of GENDER AS CLASS). Even Haslanger's account of 'woman' is not explicitly relational and instead her definition of WOMAN focuses on intrinsic features of this category. A general drawback of using only intrinsic features is that this way one has a hard time capturing variation within any category, and in the case of gender variation is essential because "there is no one 'women's social role'" (Saul 2012: 197). A relational account points to the *differential* social treatment of genders. It neither has to claim that all women are subordinated—as Haslanger's definition problematically does by excluding non-subordinated

[14] Recognizing that her account of WOMAN raises this question, Haslanger (2000) is compelled to respond that she "not convinced that there are many cases (if any) of the latter" (females who are not subordinated for being considered female), while also stating that her "analysis is intended to capture a meaningful political category for critical feminist efforts, and non-oppressed females do not fall within that category" (p. 46).

females—nor does it have to articulate what this allegedly shared social subordination consist of, for example, the "certain kinds of social position" to which Haslanger appeals in condition (ii) of her definition (which we restated in an earlier section). Instead, a relational construal can indicate that *compared to* men, women tend to be discriminated against because of their gender.

In our view, the decisive advantage of a relational approach is its ability to capture *intersectionality*, that is, the way in which several social identities (gender, race, class, sexual orientation, trans status, etc.) intersect for a particular person so as to aggravate or attenuate oppression and privilege (Crenshaw 1989, 1991).[15] When focusing on GENDER rather than WOMAN and articulating gender as an axis of discrimination, this engineered concept flags that gender is *only one* axis of discrimination, and is open to there being many others dimensions of discrimination (and other salient features related to social differences) beyond gender, including race and sexual orientation.[16] Recall that this concept of GENDER is to answer to the aim of identifying (and possibly explaining) gender-based discrimination, which for empirical and political reasons has to comport with intersectionality. A relational account of GENDER highlights the *organized and systemic* nature of gender-based discrimination, without claiming that every woman is oppressed and that gender is the only dimension of discrimination.

(2) Assigning legal rights and ensuring gender-appropriate social recognition. To be sure, a relational concept of GENDER fails to delineate exactly which persons are of a certain gender, which is not relevant for the purpose of identifying and explaining society-wide discrimination, but is indeed indispensable for the aim of ensuring gender-based legal rights and social recognition in line with someone's gender identity. With respect to this aim, a *different* concept of gender has to be employed, that entails a specific extension of WOMAN and other gender categories (and that has a normative prong by assigning legal rights and social status to these persons). Jenkins (2016) did put forward an account that is meant to ensure that trans women count as women. However, we have argued that it fails to do so. For by insisting that one needs to be aware of gender-specific social expectations, Jenkins's definition makes an individual person's gender contingent on an ideological test imposed by the larger culture. This is not only irrelevant but harmful, given that it would not grant gender-based legal rights and social recognition of a person's chosen gender to someone who is unaware of some social expectations. In contrast, for the aim of ascribing rights and social recognition, in the previous section we proposed a concept of GENDER according to which one's sincere belief that one is of a particular gender suffices. Although a woman's psychological identity includes more substantial

[15] Haslanger has intersectionality concerns clearly in view, regardless of whether her particular definitions of MAN and WOMAN fully comport with this. For instance, Jones (2014) questions whether indigenous Australian men (who are not privileged by any standard) count as MEN on Haslanger's definition, thereby also touching upon questions about the *extension* of concepts such as MAN and WOMAN from which we have tried to move away. (A further discussion that is critical of Haslanger but still focuses exclusively on the extensions of concepts is Mikkola 2009.)

[16] Although our below concept of WOMAN that is to assign rights (and therefore has to include trans women) is dedicated to trans issues, this does not mean that the present concept of GENDER cannot acknowledge trans status as yet a further axis of discrimination that is distinct from gender.

beliefs about what it means to be a woman, the belief that she belongs to the gender category termed 'woman' is all that is required for her to be recognized and treated accordingly.

(3) Empowering persons by means of their gender identity. A further social aim we have not discussed yet is empowering groups of persons, for instance women. This also requires a specific, strategic concept of GENDER. Our first concept of GENDER (the concept in the spirit of Haslanger's ameliorative account), when not only identifying but also offering a more detailed explanation of the social workings of gender-based discrimination, may point to public policy means of reducing discrimination within an overall society. However, this may not be helpful for empowering individual women. For this purpose, the best solution may be a concept of GENDER that incorporates substantial psychological aspects of gender identity, provided they are positive features conducive to personal empowerment (rather than internalized harmful stereotypes). Such psychological aspects of gender identity were deliberately left out from our second GENDER concept as self-identification with a particular gender, so *this* concept is clearly unsuitable for the purpose of personal empowerment. A concept serving the aim of an individual's personal and social empowerment needs to include relevant gender-related psychological resources and social affordances. A person should be able to pick and choose amongst a variety of (non-binary) gender associated behaviours and activities in order to construct their own particular gender attributes that enable agency. A clear limitation is that the psychological and social features relevant to a particular individual do not hold for all persons. Indeed, discussions on intersectionality have already shown that even different persons belonging to one gender (e.g., all who identify as women on the second gender concept) need different empowerment strategies, so that it is doubtful that the task can be achieved by a single psychological identity concept, no matter how flexible it is. But the employment of different psychological identity concepts by different groups of persons is something that our agenda of strategic conceptual engineering welcomes.

Conclusion

In this essay we have put forward the agenda of strategic conceptual engineering, which in addition to the development of novel concepts consists in the employment of a variety of concepts, such as several concepts of RACE, where each such concept is geared toward a specific epistemic or social aim, while in other contexts the use a different concept is more fruitful. We have illustrated this approach by arguing that at least three concepts of GENDER are needed, even when only social aims are in view. With respect to the *first* aim of identifying and explaining gender-based inequities, a concept similar to Haslanger's ameliorative account is needed; although instead of focusing on the extension of WOMAN we suggested a relational construal of GENDER which highlights differential social treatment across genders and thereby points to gender as one (though not the only one) axis of discrimination. A *second* vital aim is to assign gender-based legal rights and social recognition in line with one's chosen gender, so that such a concept must be trans-inclusive and offer a clear-cut account of the extension of WOMAN and other gender categories. In contrast to Jenkins's proposal, we argued that for the purpose of ensuring rights and social recognition, awareness of gender-related social norms and other substantial psychological aspects of one's gender identity have to be omitted; and instead, one's sincere belief to be of a particular gender (e.g., the category labelled 'woman') suffices to count as being of that gender. The *third* social aim is the empowerment of persons by means of their gender identity. Now a further concept of GENDER is needed, which does include psychological features relevant to personal empowerment and awareness of one's social resources, where we acknowledged that even relative to this aim different concepts may be needed given that different groups of persons have different social experiences and occupy different social situations and thus may have to use different empowerment strategies.

Beyond the specific case of gender, our strategic conceptual engineering diverges from other general approaches. In particular, it goes in several respects beyond how the known phenomenon of *semantic contextualism* is often conceived. First, strategic conceptual engineering includes the creation of completely novel concepts and meanings. Furthermore, it importantly articulates that the 'context' that determines which meaning to use is specifically the epistemic or social aims at hand. And third, whereas Saul's (2012) contextualism assumes a unified definition of WOMAN across all contexts (in terms of some similarity to most biological females, where only the specific similarity metric differs across contexts), our three significantly distinct concepts of GENDER make plain that strategic conceptual engineering is open to different contexts requiring completely different concepts. As previously mentioned, we do not have a conclusive account of concept individuation (and for our agenda it is more important to adjudicate whether a new conceptual variant—regardless of whether it qualifies as a different concept—is an improvement). Yet paying attention to the aims of concept use (in addition to a contentious similarity of conceptual content) also contributes to the question of concept individuation: If one but not the other concepts of GENDER is in a position to meet a given social-political aim, this is a reason to regard these three concepts as *relevantly* different (at least for this concrete case of conceptual engineering). For instance, our second concept of GENDER was put forward precisely because our first one (building on Haslanger's definition) does not specify someone's particular gender category (and thus this concept would not serve the purpose of being trans-inclusive and assigning gender-based legal rights and social recognition). Conversely, a concept in terms of one's self-identification with a certain gender category is not in a position to provide a causal explanation of how gender functions as one axis of society-wide discrimination.

Our approach has also the potential to enrich philosophical views of concepts in general. Discussions in the philosophy of mind and language commonly concern the role that conceptual content has for determining the concept's extension, that is, they focus on a concept's satisfaction conditions. When reflecting on some criticisms of Haslanger's account (driven by the question of who would fall under WOMAN), we have argued that (in this specific context) such considerations about the extension of concepts are actually a red herring, given that some concepts do not serve the aim of classification (or of assigning rights to a clearly delineated group of individuals), but may serve the aim of causal explanation (e.g., of discrimination within society). Thus, concepts should not only be philosophically studied in terms of their extensions and satisfaction conditions, but also in terms of the way in which the content embodied in a concept supports explanatory reasoning. Even our second concept of GENDER (where extensions matter) has not only a descriptive prong that specifies who counts as falling under a certain gender category, but also a normative prong that entails legal rights and a commitment to the social acceptance of a person's gender identity, so that this concept can function in moral reasoning (see also Reuter 2019).

The topic of our discussion was concepts as vehicles of thought which also motivate action. An issue we have not covered is that the promotion of a useful concept among a wider group of individuals (e.g., for political purposes or empowering others) requires that a concept be communicated, which makes it necessary to lexicalize it by using some term that expresses this concept. Although one option is to use a single term such as 'gender' while having its meaning vary across contexts, we have already encountered the worry that this may hamper communication across persons. Mikkola (2009) objects to Haslanger's ameliorative account of gender on the grounds that its use by feminists would only confuse ordinary speakers. But this assumes a strict separation between the language use of "feminists" and "ordinary speakers,"[17] as if either of these groups was not internally diverse regarding its language and concept use. These groups are also overlapping; and in our view, it is not implausible that a person can use our second or third meaning of GENDER (which takes a first-person view and is of benefit to oneself), while *also* employing the first meaning of GENDER (which partially takes a third-person view and is most fruitful in thinking about the situation of various persons of one's gender), at least in other contexts. Generally, common language already permits a good deal of flexibility, where a word is used with somewhat varying meanings, which need not conflict in the aspects relevant to a communicative situation, or where the context disambiguates sufficiently to not hinder discourse.

To be sure, there are drawbacks with expressing several meanings by means of an established term, given that associations tied to one (entrenched) meaning may carry over to another meaning (McPherson and Plunkett, Chapter 14, this volume). Yet a more compelling reason than confusion across speakers is political consequences. Despite our resolute pluralism on the level of concepts, we acknowledge that on the terminological level one meaning of 'gender' may often have to prevail in public discourse (and different words may have to be found for the other meanings). Among the three purposes that concepts of GENDER may serve, currently the politically most consequential agenda is to ensure legal rights and social recognition, as shown by efforts to publicly misgender trans persons, which would have major practical consequences not only in the case of successful legal measures to deny access to public facilities such as bathrooms. Thus, to the extent to which existing terms such as 'gender' and 'woman' need to be associated with a prioritized meaning, it would be our second concept of GENDER as one's self-identification with a gender category.

[17] "Quite simply, if feminists appropriated Haslanger's gender terms, this would create linguistic confusion between them and ordinary speakers... achieving this task would be hugely difficult if feminists appropriated Haslanger's gender terminology because it complicates communication between them and ordinary language users" (Mikkola 2009: 569).

Note that even such a more monistic use of the term 'gender' diverges from how some persons ordinarily use the term—our self-identification concept of GENDER definitely is not about someone's biological sex at birth. Generally, the point of conceptual engineering and conceptual ethics is to improve some concepts that are currently used. Not only does this make it unavoidable that a term is used with its current meaning while some (also) use it with a revised meaning, but the advocacy for a politically more appropriate meaning of a traditional term will always generate social friction. The widespread use of such a revised meaning has to be seen as a political ideal, which can only be achieved by a gradual, arduous process.

Acknowledgements

We thank the participants of the June 2016 'Conceptual Engineering and Conceptual Ethics' workshop in Oslo for their comments on our chapter, and Herman Cappelen for the organization and funding of the workshop. We are especially grateful to David Plunkett and Herman Cappelen for detailed comments on a previous version of this essay. Ingo Brigandt's work was also supported by the Social Sciences and Humanities Research Council of Canada (Insight Grant 435-2016-0500).

References

Ainsworth, C. 2015. Sex Redefined: The Idea of Two Sexes is Simplistic. Biologists Now Think There is a Wider Spectrum Than That. *Nature* 518:288–91.

Alcoff, L. 1988. Cultural Feminism versus Post-structuralism: The Identity Crisis in Feminist Theory. *Signs* 13:405–36.

Alcoff, L. M. 2006. Latinos and the Categories of Race. *Visible Identities: Race, Gender, and the Self* (pp. 227–46). New York: Oxford University Press.

Appiah, K. A. 1996. Race, Culture, Identity: Misunderstood Connections. Part 1: Analysis: Against Races. In K. Anthony Appiah and Amy Gutmann (eds.), *Color Conscious: The Political Morality of Race* (pp. 30–74). Princeton: Princeton University Press.

Bach, T. 2012. Gender is a Natural Kind with a Historical Essence. *Ethics* 122:231–72.

Bettcher, T. M. 2013. Trans Women and the Meaning of 'Woman.' In Nicholas Power, Raja Halwani, and Alan Soble (eds.), *Philosophy of Sex: Contemporary Readings* (6th edn) (pp. 233–50). Lanham: Rowman & Littlefield.

Bettcher, T. M. 2017a. Through the Looking Glass: Transgender Theory Meets Feminist Philosophy. In Ann Gary, Serene J. Khader, and Alison Stone (eds.), *The Routledge Companion to Feminist Philosophy* (pp. 393–404). New York: Routledge.

Bettcher, T. M. 2017b. Trans Feminism: Recent Philosophical Developments. *Philosophy Compass* 12:e12438.

Brigandt, I. 2003. Homology in Comparative, Molecular, and Evolutionary Developmental Biology: The Radiation of a Concept. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 299B:9–17.

Brigandt, I. 2010. The Epistemic Goal of a Concept: Accounting for the Rationality of Semantic Change and Variation. *Synthese* 177:19–40.

Brigandt, I. 2011. Natural Kinds and Concepts: A Pragmatist and Methodologically Naturalistic Account. In Jonathan Knowles and Henrik Rydenfelt (eds.), *Pragmatism, Science and Naturalism* (pp. 171–96). Frankfurt am Main: Peter Lang.

Brigandt, I. 2012. The Dynamics of Scientific Concepts: The Relevance of Epistemic Aims and Values. In Uljana Feest and Friedrich Steinle (eds.), *Scientific Concepts and Investigative Practice* (pp. 75–103). Berlin: de Gruyter.

Brigandt, I. 2013. A Critique of David Chalmers' and Frank Jackson's Account of Concepts. *ProtoSociology* 30:63–88.

Brigandt, I. 2015. Social Values Influence the Adequacy Conditions of Scientific Theories: Beyond Inductive Risk. *Canadian Journal of Philosophy* 45:326–56.

Burgess, A., and Plunkett, D. 2013a. Conceptual Ethics I. *Philosophy Compass* 8:1091–101.

Burgess, A., and Plunkett, D. 2013b. Conceptual Ethics II. *Philosophy Compass* 8:1102–10.

Butler, J. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Butler, J. 1993. *Bodies That Matter: On the Discursive Limits of "Sex."* New York: Routledge.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Chalmers, D. J., and Jackson, F. 2001. Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110:315–60.

Crenshaw, K. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* (pp. 139–67). Chicago: University of Chicago Press.

Crenshaw, K. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43:1241–99.

Díaz-León, E. 2016. *Woman* as a Politically Significant Term: A Solution to the Puzzle. *Hypatia* 31:245–58.

Díaz-León, Esa. Chapter 9, this volume. Descriptive vs. Ameliorative Projects: The Role of Normative Considerations. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Douglas, H. E. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Feyerabend, P. 1962. Explanation, Reduction, and Empiricism. In Herbert Feigl and Grover Maxwell (eds.), *Scientific Explanation, Space, and Time* (pp. 28–97). Minneapolis: University of Minnesota Press.

Gannett, L. 2010. Questions Asked and Unasked: How by Worrying Less about the '*Really* Real' Philosophers of Science Might Better Contribute to Debates about Genetics and Race. *Synthese* 177:363–85.

Hardimon, M. O. 2017. *Rethinking Race: The Case for Deflationary Realism*. Cambridge, MA: Harvard University Press.

Haslanger, S. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Noûs* 34:31–55.

Haslanger, S. 2004. Future Genders? Future Races? *Philosophic Exchange* 34:1.

Haslanger, S. 2005. What Are We Talking About? The Semantics and Politics of Social Kinds. *Hypatia* 20:10–26.

Haslanger, S. 2006. What Good Are Our Intutions? *Aristotelian Society Supplementary Volume* 80:89–118.

Heyes, C. J. 2000. *Line Drawings: Defining Women through Feminist Practice*. Ithaca: Cornell University Press.

Heyes, C. J. 2006. Changing Race, Changing Sex: The Ethics of Self-transformation. *Journal of Social Philosophy* 37:266–82.

Hochman, A. 2013. Against the New Racial Naturalism. *The Journal of Philosophy* 110:331–51.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

Jenkins, K. 2016. Amelioration and Inclusion: Gender Identity and the Concept of *Woman*. *Ethics* 126:394–421.

Jones, K. 2014. Intersectionality and Ameliorative Analyses of Race and Gender. *Philosophical Studies* 171:99–107.

Kaplan, J. M. 2010. When Socially Determined Categories Make Biological Realities. *The Monist* 93:283–99.

Kapusta, S. J. 2016. Misgendering and Its Moral Contestability. *Hypatia* 31:502–19.

Kessler, S. J. 1990. The Medical Construction of Gender: Case Management of Intersexed Infants. *Signs* 16:3–26.

Kitcher, P. 1984. Species. *Philosophy of Science* 51:308–33.

Kitcher, P. 2001. *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Kitcher, P. 2007. Does "Race" Have a Future? *Philosophy & Public Affairs* 35:293–317.

Kopec, M. 2014. Clines, Clusters, and Clades in the Race Debate. *Philosophy of Science* 81:1053–65.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Ludwig, D. 2014. Hysteria, Race, and Phlogiston: A Model of Ontological Elimination in the Human Sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences* 45:68–77.

Mallon, R. 2006. "Race": Normative, Not Metaphysical or Semantic. *Ethics* 116:525–51.

McKinnon, R. 2018. The Epistemology of Propaganda. *Philosophy and Phenomenological Research* 96:483–9.

McKitrick, J. 2007. Gender Identity Disorder. In Harold Kincaid and Jennifer McKitrick (eds.), *Establishing Medical Reality: Essays in the Metaphysics and Epistemology of Biomedical Science* (pp. 137–48). Dordrecht: Springer.

McPherson, Tristram, and David Plunkett. Chapter 14, this volume. Conceptual Ethics and the Methodology of Normative Inquiry. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Mikkola, M. 2009. Gender Concepts and Intuitions. *Canadian Journal of Philosophy* 39:559–83.

Pigliucci, M. 2001. *Phenotypic Plasticity: Beyond Nature and Nurture*. Baltimore: Johns Hopkins University Press.

Pigliucci, M. 2013. What Are We to Make of the Concept of Race? Thoughts of a Philosopher–Scientist. *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:272–7.

Plunkett, D., and Sundell, T. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13:23.

Plunkett, D. 2015. Which Concepts Should We Use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 58:828–74.

Plunkett, D. 2016. Conceptual History, Conceptual Ethics, and the Aims of Inquiry: A Framework for Thinking about the Relevance of the History/Genealogy of Concepts to Normative Inquiry. *Ergo* 3:27–64.

Prinzing, M. 2018. The Revisionist's Rubric: Conceptual Engineering and the Discontinuity Objection. *Inquiry* 61:854–80.

Reuter, K. 2019. Dual Character Concepts. *Philosophy Compass* 14:e12557.

Richard, Mark. Chapter 17, this volume. Conceptual Evolution. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Rose, S., Ceci, S., and Williams, W. M. 2009. Should Scientists Study Race and IQ? *Nature* 457:786–9.

Roth, V. L. 1991. Homology and Hierarchies: Problems Solved and Unsolved. *Journal of Evolutionary Biology* 4:167–94.

Saul, J. 2006. Gender and Race. *Aristotelian Society Supplementary Volume* 80:119–43.

Saul, J. 2012. Politically Significant Terms and Philosophy of Language: Methodological Issues. In Sharon Crasnow and Anita Superson (eds.), *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy* (pp. 195–216). Oxford: Oxford University Press.

Scharp, K. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Spencer, Q. 2018a. Racial Realism I: Are Biological Races Real? *Philosophy Compass* 13:e12468.

Spencer, Q. 2018b. Racial Realism II: Are Folk Races Real? *Philosophy Compass* 13:e12467.

Stone, A. 2004. Essentialism and Anti-essentialism in Feminist Philosophy. *Journal of Moral Philosophy* 1:135–53.

Templeton, A. R. 2013. Biological Races in Humans. *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:262–71.

Wagner, G. P. 1989. The Origin of Morphological Characters and the Biological Basis of Homology. *Evolution* 43:1157–71.

Watson, L. 2016. The Woman Question. *TSQ: Transgender Studies Quarterly* 3:246–53.

Yudell, M., Roberts, D., DeSalle, R., and Tishkoff, S. 2016. Taking Race Out of Human Genetics. *Science* 351:564–5.

Zack, N. 1993. *Race and Mixed Race*. Philadelphia: Temple University Press.

6

Never Say 'Never Say "Never"'?

Alexis Burgess

We theorize with the concepts we have, not the ones we hope to have eventually, at some ideal limit of inquiry. That's true even when it comes to theorizing *about* our current concepts. Such theorizing can therefore be reflexive, blending use and mention. To take one simple case: a standard semantics for conjunction just uses conjunction itself. More to the point of the present volume, this sort of reflexivity shows up in *normative* theorizing about concepts too. For instance, we'll use logical concepts to reason about which logical concepts we *ought* to be using.

Now suppose you and I disagree about whether or not we ought to use some particular concept C, and we want to start arguing about it. A question of method arises. Should we suspend use of C for the duration of our debate?[1] After all, it would seem hypocritical of me to use C in the course of arguing that we shouldn't use C. And it seems circular or question-begging for you to use C in arguing that we can permissibly use C (or that we must). So perhaps we should just agree at the outset to avoid using C. But then again, that policy might sound like a significant concession to me, the enemy of C. What's more, whether or not this suspension policy would be prejudicial, it might turn out to be quite impracticable, if C is somehow central to our conceptual schemes; as concepts of philosophical interest tend to be. There are no Archimedean points in conceptual ethics. So what exactly are the rules of engagement here?

This short chapter tries to make some headway on the methodology of conceptual ethics by offering a qualified defense of "hypocrisy".[2]

First some terminology. By an argument in conceptual ethics, I'll mean an argument in the strict formal sense, whose conclusion concerns whether or not we should use some concept C. This is already a stipulative restriction in scope. I'm ignoring evaluative conclusions, for example, to focus exclusively on deontic modals. And I'm not going to say much about arguments that target multiple concepts at once (e.g., all vague concepts, inconsistent concepts, or concepts of race and gender).

My spotlight is meant to be flexible in other ways, however. "We" whose use of C is at issue could be anyone you please: we folk, we philosophers, we atheists. "Should"

[1] Presumably we'll have to refer to C somehow, or at least quantify over it; but the question is about use, not mention.

[2] For more on the mirror-image issue of circularity, see my (2013).

Alexis Burgess, *Never Say 'Never Say "Never"'?* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Alexis Burgess.
DOI: 10.1093/oso/9780198801856.003.0006

can also be glossed in different ways: morally, prudentially, rationally; all-things-considered, or just *pro tanto*. And note that conclusions of this basic form can be combined to yield more nuanced views in conceptual ethics. Revisionism about C, for example, might be taken to decompose into a pair of views: that we should not use C (i.e., eliminativism) but should use C*; where C and C* stand in some suitable similarity relation.[3]

I'll call an argument that we should not use C an "anti-C" argument, and an argument that we can or should use C a "pro-C" argument.[4] So a "circular" argument in conceptual ethics is a pro-C argument that uses C. And a "hypocritical" argument in conceptual ethics is an anti-C argument that uses C.[5] When I say that an argument "uses" a concept C, I basically mean that a phrase whose content involves C appears in some line of the argument.

Now, actual anti-C arguments in this area rarely have conclusions as blunt as: one shouldn't use C, period. People rather tend to argue for parameterized conclusions, like: one shouldn't use C for these sorts of purposes, or in those kinds of contexts. Elsewhere, David Plunkett and I have spilled some ink on the range of relevant parameters in conceptual ethics.[6] Here I just want to think schematically. To be more realistic, we'll want to say that anti-C arguments end in the verdict that we shouldn't go in for a certain (probably improper) subset of possible uses of C. Or put another way: that we shouldn't use C is settings of type S.[7] To be explicit, call these anti-C/S arguments. This adjustment mandates a corresponding change to our definition of hypocrisy. The anti-C/S arguments of interest are just those that make use of C in what amounts to a setting of type S. (But I'll suppress this complication for readability when it's not relevant.)

So, should we abjure hypocritical arguments in conceptual ethics? Or more positively, is it permissible to argue hypocritically?

A number of my informants have reported that hypocrisy doesn't feel inherently problematic to them. Sure, there may be something odd about *initiating* a debate in conceptual ethics with a hypocritical argument. But if your interlocutor is already on record as endorsing the use of some concept C, then using C yourself in the course of trying to persuade her otherwise just feels like "addressing one's opponent on her

[3] To get closer to the conventional understanding of revisionism, one might add that C is already in circulation, while C* is not.

[4] For symmetry, one could also call arguments that we can permissibly refrain from using C "anti-C". But I won't be concerned with this broader class of conclusions. Eliminativism and revisionism, which lie at the heart of conceptual ethics, both involve the stronger claim that we should not use C. And there's nothing *prima facie* hypocritical or problematic about using C to argue that we can permissibly refrain from using C.

[5] Interestingly, perhaps, there's no analogue of hypocrisy in conceptual analysis. A circular analysis uses C to specify the content of C itself. A circular ethics uses C to argue in favor of using C. A hypocritical ethics uses C to argue against using C. But there's nothing "hypocritical" about using C to specify the content of some other concept, like C's complement, or an antonym. If anything, that would just be another kind of circularity.

[6] Burgess and Plunkett (2013).

[7] I like the word 'setting' because it's relatively free of other theoretical associations. I also like that it can be read as both environment/context and "value of a variable". I think I can enjoy those resonances simultaneously without committing any fallacies of equivocation.

own terms". Cast in this dialectical light, hypocrisy actually looks a lot like good old-fashioned *reductio*.[8] Witness where the presumption that it's kosher to use C leads us: to the conclusion that it's not!

Of course, if and when you persuade your opponent that we shouldn't use C, *via* some hypocritical argument, she may well reflect on her route to that conclusion, note that it used C, and hence retrospectively disavow the argument.[9] But that's just "kicking away a ladder" once it's been climbed. No problem there, right?

The problem with this kind of ladder-kicking is what Gilbert Harman (1986: 39) has called the principle of positive undermining. His thought is basically that a subject should stop believing P once she realizes that her basis for that belief isn't actually any good. And unless your opponent above is totally daft, she'll see that the only reason she eventually came around to your anti-C position (P) was a hypocritical argument that she's now disowned. So, according to Harman's principle, she should now stop believing that we shouldn't use C. Once she kicks away the hypocritical ladder, the rational thing to do is jump back down off the roof.[10]

Now, I don't take this line of thought to show that there is definitely something wrong with hypocritical arguments. After all, I aim to defend hypocrisy in the end. But I do take the Harman-inspired point to undermine (or at least problematize) the idea that hypocrisy is innocuous just because of its resemblance to *reductio*.[11] And we'll soon see a simple, direct argument against hypocrisy, which will take some time to evaluate.

First, though, I want to spend a moment addressing the significance of getting clearer on the status of hypocrisy. Officially, a hypocritical argument is an anti-C/S argument that uses C in a setting of type S. How often do such arguments actually crop up in conceptual ethics? And how hard is it to reformulate them without hypocrisy when they do arise? If it's not often, or not hard, then maybe we shouldn't

[8] It's not really *reductio*, though. I'm actually not sure how to describe the logical or formal relationship between the supposition that one can permissibly use C (which merely mentions C) and the first subsequent step of the argument that actually uses C.

[9] Unless the conclusion of the argument is that we shouldn't use C outside the scope of a supposition (or some such setting S), since we're assuming the argument only uses C within the scope of a supposition. But if this conclusion was derived by discharging a supposition reduced to absurdity, the supposition must have been that it's permissible to use C outside the scope of a supposition. But this supposition would not have warranted subsequent use of C within its scope.

[10] To make matters more complicated, though: once your opponent has dropped the anti-C view, her only reason for resisting your hypocritical argument is gone! So she should presumably climb the ladder once again. She could keep going around and around in this way indefinitely. But this possibility doesn't make hypocritical arguments look particularly attractive.

[11] Here's a reply due to David Taylor. Nevermind whether hypocritical arguments are themselves acceptable. The mere existence of an otherwise compelling hypocritical argument immediately provides for a compelling, non-hypocritical, anti-C argument, along the following lines. Initially adopting the pro-C position would lead to a kind of cognitive instability (in light of the previous note), where we're constantly flip-flopping between accepting and rejecting the use of C. This situation is rationally untenable, so we can only embrace the anti-C alternative. (Compare my own Reply to the Objection at the very end of the present chapter.) I'm not sure whether to accept this friendly suggestion, however. For the rational force of Taylor's non-hypocritical argument-form seems to hinge on the status of the associated hypocritical argument. If hypocrisy is a disqualifying vice, friends of C can reasonably refuse the first flip-flop. So there may be no avoiding a direct confrontation with the issue of the chapter: whether hypocrisy itself is somehow problematic.

invest too much energy in trying to understand or evaluate hypocrisy. These two issues are closely related. How "hard" it is to transform a hypocritical argument into a non-hypocritical argument depends in large part on how densely packed the field of possible anti-C/S arguments is with hypocritical ones. So let me first speak to this basic statistical issue. And let me bracket setting-sensitivity to begin with.

Conceptual ethics is a part of philosophy that's preoccupied with concepts. The concepts that interest philosophers tend to be "central" to our ordinary conceptual schemes, in the sense that they're implicated in the analysis of a whole host of other concepts. I'm thinking of things like modal concepts (necessary, contingent), logical concepts (exists, identical), normative concepts (should, right), epistemic concepts (know, rational), and so on.[12] Since we still do philosophy in something resembling natural language, central concepts like these will inevitably turn up in any sustained effort to specify the contents of the bits of language use that make up philosophical discourse. Discourse in conceptual ethics is no exception. Any given anti-C argument therefore runs a significant risk of deploying a concept whose analysis involves C itself. In a generous sense of "use"—where using the concept vixen, for example, involves using the concept fox—such an argument would count as hypocritical.[13]

To get a feel for the size of this risk, let's play with some numbers. Let's say there are 50 central concepts of philosophical interest. And let's say the average philosophical argument (in conceptual ethics) makes overt, explicit use of 30 different concepts (central or otherwise). And let's say the average conceptual analysis involves 4 central concepts. Of the $(30 \times 4 =)$ 120 token concepts implicated in the average argument, at least 70 have to be redundant (given that there are only 50 central concepts). Let's say as many as 100 are redundant, so that the average non-redundant concept shows up 6 times in the argument. Then the chance of hypocrisy would be $(20/50 =)$ 40%. That's pretty high. If there were something terribly wrong with hypocritical arguments, this level of risk might be unacceptable.

Suppose you found yourself arguing hypocritically by accident. How easy would it be to rectify the mistake, without compromising the essential upshot of the argument you were trying to make? I doubt there is any general recipe for converting hypocritical arguments into equi-plausible, non-hypocritical arguments. And case-by-case conversion won't be as simple as paraphrasing away a few offending words. The "paradox" of analysis points up just how hard it is to know in advance which concepts will be implicated in an adequate definition. Our best efforts at paraphrase could easily end up introducing new dimensions of hypocrisy, or simply preserving old ones.

[12] Some concepts of interest to actual conceptual ethicists probably won't show up in the analysis of any concepts explicitly used by those ethicists. I'm thinking of slurs, for example, or racial concepts. These are not central to our conceptual schemes in the relevant sense. So if your interest in this book stems mostly from your interest in concepts like those, this chapter may not be for you.

[13] It doesn't matter if this sense of "use" feels artificial. What's relevant is whether there's anything intuitively objectionable about using the concept vixen to argue that we shouldn't use the concept fox. Or to take a less frivolous example: whether there's anything suspicious about using the concept of reliability to argue that we shouldn't use the concept of truth (on the supposition that reliability is properly defined in terms of truth). These cases don't feel much better to me than overt, word-for-word hypocrisy.

Doesn't this whole situation look much less dire, though, once we acknowledge the setting-sensitivity of real-life arguments in conceptual ethics? Won't the typical hypocritical argument just fall outside the scope of its own S? And even if not, can't we always just replace S with some S* that excludes the argument by fiat?

Not so fast. First of all, as a purely descriptive matter, the most common settings you'll find in real-life conceptual ethics are by far, "for ordinary practical purposes" and "for serious theoretical purposes". Philosophical arguments may fall outside the scope of the former, but they certainly don't escape the latter. Second, excluding an argument from its own scope by fiat could be problematically *ad hoc*. If the modified argument featuring S* seems plausible, that's probably because it borrows luster from some more natural, less gerrymandered argument. Take a toy example. Compare a hypocritical argument that we shouldn't use the notion of numerical identity "for serious theoretical purposes" to a sanitized variant that self-consciously excludes normative theorizing about that very concept. If the second argument is any good, that's probably due to its being a special case of the first.[14]

All of which is really just to say that it would certainly be nice if hypocrisy were innocuous. Then we could go about our business in conceptual ethics without constantly monitoring the concepts we use to mount our anti-C arguments. Unfortunately, there is a simple line of reasoning to the effect that we should never accept hypocritical arguments:

Argument X.[15] For any concept C, any setting-type S, and any hypocritical, anti-C/S argument H, either H is sound or it isn't. If H isn't sound, then we shouldn't accept it. So suppose H is sound (case two). Then its conclusion is true. Its conclusion is that we shouldn't use C in settings of type S. Since H is hypocritical, to accept H would be to use C in such a setting.[16] So we shouldn't accept H. So, in either case, whether or not H is sound, we shouldn't accept it.[17]

It might sound odd to counsel (in case two) against accepting a sound argument. Soundness is usually what we tell our students to aspire to. But we all know that some sound arguments are dialectically no good: the question-begging ones. Argument X just goes to show that we should treat hypocrisy similarly.

This last observation raises a question of focus for us.[18] When reflecting on conceptual ethics, maybe we ought to be more interested in issues of soundness and truth; less interested in issues of method or disputation. Metaphysics first, epistemology later. To make the complaint vivid, consider a hypocritical one-liner (where premise and conclusion collapse into a single claim):

(Z) We should not use 'should'.[19]

[14] Here I take issue with Scharp (Chapter 19, this volume, Section 8) on defective concepts in conceptual engineering.

[15] Interestingly, there doesn't seem to be any analogous argument against circularity in conceptual ethics.

[16] More generally: to accept an argument that uses C is *inter alia* to use C yourself.

[17] This might be a confusing way to state the conclusion, if 'should' means different things in case one and case two. In case two, it means whatever 'should' means in the conclusion of H. In case one, it's just being used to track the sense(s) in which unsound arguments are verboden.

[18] Thanks to Herman Cappelen for pressing me here, and inspiring the following digression.

[19] Or never say 'never'. See Eklund (2015: section 5) on conceptual fixed points.

If Z were true, that would be headline news. It wouldn't matter much, by comparison, whether we should nevertheless avoid accepting Z. So why have I been focusing on the rules of engagement in conceptual ethics rather than the facts on the ground?

Not for any good reason, really. I think these alethic questions are interesting too. But hypocritical one-liners are few and far between in conceptual ethics. (Because there are only so many concepts involved in articulating any anti-C view.) And when it comes to full-blown arguments, my hunch is that hypocrisy should be no obstacle to validity. Soundness is a further question. I can imagine someone thinking that hypocritical arguments can't possibly be sound (and that one-liners can't possibly be true). But I have no firm view. And however these alethic issues shake out, I do think it's worth wondering whether there's something independently, dialectically wrong with hypocrisy, in the same vein as begging the question. Who knows, though, what order we should take all this in. I'm focusing on method mainly because I have something to say about it. Here, then, is the idea that precipitated the present chapter. Notice that X isn't just an argument *about* conceptual ethics, it's actually a *contribution to* conceptual ethics. Its conclusion says that we shouldn't accept hypocritical arguments. But accepting an argument involves using the concepts featured therein. So X concerns a normative question about concept use. To be more exact, we could regiment its conclusion as a prohibition against using any concept C in a certain type of setting: hypocritical arguments. (Or more narrowly: arguments that are hypocritical with respect to C specifically; the choice won't matter.) Now, the observation that X itself amounts to an argument in conceptual ethics proper doesn't immediately impugn X. But it does raise an interesting question.

Is X hypocritical? Well, we've only defined hypocrisy for arguments in conceptual ethics whose conclusions target individual concepts. So consider an argument generated from X by universal instantiation on C. And let's pick a concept explicitly used in X, like soundness, acceptance, disjunction, or even hypocrisy itself. The conclusion of this new argument would therefore be that we shouldn't use the concept of (e.g.) acceptance to mount hypocritical arguments. Call the new argument A, for acceptance:

Argument A. For any setting-type S, and any hypocritical, anti-*acceptance*/S argument H, either H is sound or it isn't. If H isn't sound, then we shouldn't accept it. So suppose H is sound (case two). Then its conclusion is true. Its conclusion is that we shouldn't use *acceptance* in settings of type S. Since H is hypocritical, to accept H would be to use *acceptance* in such a setting. So we shouldn't accept H. So, in either case, whether or not H is sound, we shouldn't accept it. In other words: we shouldn't use *acceptance* in H-type settings.

If there were something problematic about A, then presumably X would inherit the problem.[20] So, is A hypocritical? Like X, it uses the concept of acceptance. And unlike X, its conclusion does specifically say that we shouldn't use that very concept. But A only proscribes the use of that concept in settings of a certain type: hypocritical,

[20] Even if other instances of X, generated by instantiating C with other concepts, aren't problematic. After all, to endorse X in full generality is effectively to endorse each and every one of its instances. We could restrict the initial quantifier over C in X just to concepts that aren't used in X itself. But then we'd have to confront a version of the borrowed luster issue raised earlier.

anti-*acceptance* arguments. As we've just said, A is an anti-*acceptance* argument. So, given the definition of hypocrisy, the only thing left for us to check in order to determine whether A is hypocritical is therefore... whether A is hypocritical! If it is, it is; and if it isn't, it isn't.

This predicament is strikingly similar to the situation with truth-teller sentences in the literature on the semantic paradoxes, like "This sentence is true". As far as the T-scheme is concerned, all we can say is that the truth-teller is true if it's true; not if it's not. And it's not at all clear where else to look for guidance as to the sentence's truth-value. Similarly, it's unclear what might settle the status of A if not the definition of hypocrisy (together with A's intrinsic features).

What follows from all this? At the very least, I think we can conclude that there is something odd and potentially problematic about A, and therefore X. Since X is the only argument we have managed to marshal against hypocrisy, this conclusion strikes me as a significant, if vague, defensive result. I haven't offered a positive argument that hypocrisy is kosher. (Recall that I didn't endorse the argument that assimilates hypocritical reasoning to *reductio*.) I won't say that hypocrisy is innocent until proven guilty. But I will stop investing time trying to avoid hypocrisy in conceptual ethics until someone comes up with a solid reason.

Objection. The worst-case scenario you painted above was that A might be hypocritical. But you don't think there's anything wrong with hypocrisy. So you have no reason to resist A, and therefore no reason to resist X.

Reply. But if we accept X, then we will think there's something wrong with hypocrisy. So accepting X is unstable. Moreover, you've misdescribed the worst-case scenario. The worry was rather that it's somehow unsettled or indeterminate whether A is hypocritical. Whatever you make of (determinate) hypocrisy, this liminal status is just weird.

Conclusion. We should not yet conclude that we should not argue hypocritically when we argue about which concepts we should use.

Acknowledgements

Thanks to Davids Plunkett and Taylor for (separate) comments on an earlier draft. And to audiences in Oslo and Leeds. Very special thanks to Herman Cappelen for his support, encouragement, and generosity over the past few years. Without it, this essay simply wouldn't have existed.

References

Burgess, Alexis. 2013. Keeping 'True': A Case Study in Conceptual Ethics. *Inquiry* 57:580–606.

Burgess, Alexis, and Plunkett, David. 2013. Conceptual Ethics II. *Philosophy Compass* 8:1102–10.

Eklund, Matti. 2015. Intuitions, Conceptual Engineering, and Conceptual Fixed Points. *The Palgrave Handbook of Philosophical Methods*. London: Palgrave Macmillan.

Harman, Gilbert. 1986. *Change in View*. Cambridge, MA: MIT Press.

Scharp, Kevin. Chapter 19, this volume. Philosophy as the Study of Defective Concepts. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

7

Conceptual Engineering
The Master Argument

Herman Cappelen

I call the activity of assessing and developing improvements of our representational devices 'conceptual engineering'.[1] The aim of this chapter is to present an argument for why conceptual engineering is important for all parts of philosophy (and, more generally, all inquiry). Section I of the chapter provides some background and defines key terms. Section II presents the argument. Section III responds to seven objections. The replies also serve to develop the argument and clarify what conceptual engineering is.

I. Background and Explanation of Central Terms

If we use 'conceptual engineering' as I suggested above, that is, to mean *the project of assessing and developing improvements of our representational devices*, then if you think 'concepts' are the core representational devices, conceptual engineering amounts to the following: It is the project of assessing and then ameliorating our concepts.[2] For example:

- An epistemological conceptual engineer will, in an ameliorative spirit, assess epistemic concepts.
- A conceptual engineer in moral philosophy will, in an ameliorative spirit, assess moral concepts.
- A metaphysical ameliorator will, in an ameliorative spirit, assess metaphysical concepts.
- A semantic ameliorator will, in an ameliorative spirit, assess semantic concepts.

This normative project contrasts with a descriptive one. The descriptivist aims to describe the concepts we have—to describe our epistemic, moral, metaphysical, semantic, and so on, concepts. One important strand in the history of philosophy

[1] It can also include the activity of trying to implement the proposed improvements (for more on this, see Cappelen and Plunkett's Introduction to this volume).

[2] In Cappelen (2018), I end up not describing conceptual engineering in this way, but the argument in this chapter can be presented independently of those reservations.

Herman Cappelen, *Conceptual Engineering: The Master Argument* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Herman Cappelen.
DOI: 10.1093/oso/9780198801856.003.0007

is a battle (or tension or at least division[3]) between Descriptivists and Revisionists. The distinction will not be simple or clear-cut and the battle lines have been drawn in different ways in different time periods. But in each time period and in all parts of philosophy, we find these two fundamentally conflicting attitudes or goals. For some, success is measured by a true description of, for example, what knowledge, belief, morality, representation, justice, or beauty is. For others, the aim is figuring out how we improve on what we have: how can we improve on (our concepts of) knowledge, justice, belief, beauty, etc.? Those with the former aim tend to find the latter unintelligible (or naive) and those with the latter aim tend to find the former complacent, uninspired, and lazy.

Nietzsche is perhaps the paradigm of a philosopher advocating what he calls 'absolute skepticism' towards our inherited concepts. In *The Will to Power*, he writes:

> Philosophers . . . have trusted in concepts as completely as they have mistrusted the senses: they have not stopped to consider that concepts and words are our inheritance from ages in which thinking was very modest and unclear . . . . What dawns on philosophers last of all: they must no longer accept concepts as a gift, nor merely purify and polish them, but first make and create them, present them and make them convincing. Hitherto one has generally trusted one's concepts as if they were a wonderful dowry from some sort of wonderland: but they are, after all, the inheritance from our most remote, most foolish as well as most intelligent ancestors. . . . What is needed above all is an absolute skepticism toward all inherited concepts
> (Nietzsche 1901/1968: section 409)

Strawson thought about the history of philosophy in part as a battle between revisionists, like Nietzsche, and what he called descriptivists. At the beginning of *Individuals* he distinguishes between *descriptive* and *revisionary* metaphysics:

> Descriptive metaphysics is content to describe the actual structure of our thought about the world, revisionary metaphysics is concerned to produce a better structure. (1959: 9)

One of his characterizations of the revisionists' objection to descriptivists is acute. He imagines the revisionist insisting that metaphysics is

> . . . essentially an instrument of conceptual change, a means of furthering or registering new directions or styles of thought. (1959: 10)

In that brief introduction Strawson also gestures at a way of writing a history of philosophy based on the distinction between descriptivists and revisionists. He says

> [p]erhaps no actual metaphysician has ever been, both in intention and effect, wholly the one thing or the other. But we can distinguish broadly: Descartes, Leibniz, Berkeley are revisionary, Aristotle and Kant descriptive. Hume, the ironist of philosophy, is more difficult to place. He appears now under one aspect, now under another. (1959: 9)

This contrast is salient in many philosophical domains today. There are lively revisionist projects in moral philosophy, theories of truth and logic, feminist and

[3] How best to describe it is discussed further in connection with Objection (7) below.

race theory, and so on.[4] At the same time, there are descriptive projects dominating large swaths of philosophy. The paradigm might be epistemologists' endless efforts to correctly describe what the English word 'knows' means.[5] The same kind of descriptivist aim is found throughout philosophy: just think of the obsession within the philosophy of language with the minutiae of natural language and its semantics (e.g., the now-over-a-century-long debate about how the word 'the' works in English, or the flurry of work on the English word 'if'), or efforts to describe our concepts of 'freedom', or 'self', or 'object'. These are more often than not pursued as purely descriptive enterprises. Success is measured by descriptive adequacy—not by answering the question *what* should *those words mean?* and *what* should *the relevant concepts be?*[6]

II. The Master Argument

The argument I'll be defending in what follows—'The Master Argument'—is in no way original to me. It is a line of thought that can be seen as motivating many revisionist projects.[7] Nonetheless, the argument in full generality is often left implicit. The aim in what follows is to articulate the most *general* (not domain specific) version of the argument and then consider a range of objections to it.

The Master Argument:

1. If W is a word[8] that has a meaning M, then there are many similar meanings, $M_1, M_2, \ldots, M_n$, W could have.
2. We have no good reason to think that the meaning that W ended up with is the best meaning W could have: there will typically be indefinitely many alternative meanings that would be better meanings for W.
3. When we speak, think, and theorize it's important to make sure our words have as good meanings as possible.[9]
4. As a corollary: when doing philosophy, we should try to find good meanings for core philosophical terms and they will typically not be the meanings those words as a matter of fact have.

[4] The literature is large here. A representative selection is: Railton (1989, 1993) for moral philosophy; Scharp (2013) and Eklund (2014) for truth and logic; and Haslanger (2012) and Appiah (1992) for the philosophies of gender and race. See my *Fixing Language* (2018: chapter 2, section 1), for a more complete list.

[5] What Robert Pasnau laments, in his 2013, as the degeneration of epistemology into lexicography.

[6] For a brief history of the tension between revisionist and descriptivists in twentieth-century philosophy, see Cappelen 2018: chapter 2.

[7] See, for example, Chalmers 2011 and Eklund 2014, and the work in general of David Plunkett and Timothy Sundell (2013), and of Haslanger (many of the papers collected in her 2012).

[8] In this chapter I don't individuate words semantically. If you like to use the word 'word' so that it denotes a lexical item that has its meaning essentially, then you can translate from your way of speaking to mine by substituting 'lexical item' for 'word' in this chapter. The metaphysics of words is difficult and relevant, but not addressed in this chapter (see Kaplan 1990 and Cappelen 1999 for some discussion).

[9] Or, as David Plunkett suggested to me, maybe 'good enough meanings' is good enough. Maybe aiming for 'best possible' is too ambitious. I'm open-minded on this issue. This is one of several places where one can articulate related versions of the Master Argument.

5. So no matter what topic a philosopher is concerned with, they should assess and ameliorate the meanings of central terms.[10]

This is a controversial argument that builds in many assumptions. I take the project of outlining a general theory of conceptual engineering to be, in large part, the project of defending this kind of argument. In contrast, most work done on conceptual engineering is domain-specific and done by those working in those domains. For example, you might be interested in questions such as:

- What should our moral concepts be like?
- What should our gender and race concepts be like?
- What should our concepts of, say, CIVILIAN or PERSON or FAMILY or TERRORIST or TRUTH be?

Those specific topics derive their significance from the importance of the specific subject matters addressed. The Master Argument, on the other hand, isn't domain-specific. This, I take it, shows that there is a *general* research project here, that is of interest independently of the case studies.

III. Replies to Seven Objections

In what follows I address seven related objections to the Master Argument:

- Objection (1): Why think that if a word, W, has a meaning M, then there are many similar meanings W could have?
- Objection (2): In what sense can one meaning be better than another?
- Objection (3): Why not think the meanings words have are the best they can be (or at least very, very good)?
- Objection (4): If we change the meaning of an expression, won't that result in massive verbal disputes and changes of topic?
- Objection (5): Aren't meaning assignments normatively neutral, as long as each thing worth meaning is meant by some word or other? Some things are worth meaning, but why does it matter whether a given word means one of those things or something else?
- Objection (6): Why think the importance of the revisionist project undermines the importance of the descriptive project? Why think there's a tension between the two approaches? Aren't they complementary?
- Objection (7): If we are to engage in conceptual engineering, don't we have to assume that meaning assignments are within our control? If they are out of our control, how can we meaningfully engage in conceptual engineering?

In answering these objections, I hope to (a), clarify the nature of conceptual engineering, and (b), outline the central challenges for conceptual engineering as a field.

[10] As will become clear below, amelioration sometimes involves improving the meaning while keeping the lexical item fixed, and sometimes it involves the introduction of a new lexical item with an improved meaning.

Objection (1): Why Think that if a Word, W, Has a Meaning M, Then There Are Many Similar Meanings W Could Have?

I respond to this objection in two steps: First, I motivate the idea of similar meanings, and then the idea that W could have one of these similar meanings.

> *On the idea of similar meanings*: There's little agreement on what meanings are and the Master Argument is neutral on that foundational issue. However, it's hard to think of any account of what meanings are that's incompatible with this claim. A way to illustrate this is to think, in common with many philosophers, of meanings as at least intensions, that is, functions from points of evaluation to extensions.[11] So 'freedom', 'knows', 'justice', 'belief', and all other expressions are associated with an intension and this intension is a function that gives a value (an extension) for each point of evaluation. Now, just change the function a little bit and you have a similar but different meaning (if meanings either are or determine intensions). Suppose, for example, 'knows' picks out a relation between an agent and a content. The literature on the definition of 'knows' has given us literally hundreds of proposals for intensions of 'knows' that are very similar—they differ a tiny bit on how to deal with Gettier cases or lottery cases or some other weird scenarios, but are otherwise very similar. 'Knows' presumably picks out one of these functions, but there are very many similar ones.
>
> *On the idea that words can change from one meaning to another one*: So far so good, but why think that once a meaning has been fixed, it is changeable? Why not think that meaning assignments get fixed and then are stuck eternally? First-pass answer: a word, say 'marriage' *could* mean anything. We could, right now, use it to mean what 'camel' or 'soup' does. There's nothing in that sign that makes those meaning assignments impossible. If it's possible, now, for 'marriage' to mean *camel*, then it could also mean one of the meanings that are similar to its current meaning.
>
> In reply to this one might think that conceptual engineering is a matter of implementing gradual changes, not the kind of change exemplified by the possibility of assigning *camel* as the meaning of 'marriage'. However, we know that there can be such gradual changes and that they happen constantly. Historical or diachronic linguistics is a field devoted entirely to the study of various forms of syntactic and semantic changes over time. These fields are in part the study of how meanings of words evolve gradually (from one meaning to a similar one) over time. The claim that gradual semantic change is impossible is refuted by the findings of these well-established research fields.

Two issues raised by this reply will be addressed later:

a) I just argued that meaning adjustment is possible. I didn't say that it was easy or within our control. I return to the question of how we can (or cannot) be in control of meaning change in reply to Objection (7) below.

[11] For the purposes of this chapter, I will be neutral on what points of evaluation are—they might be worlds, world/time pairs or something more complicated. I don't need to take a stand on those issues in this chapter (though I do have views, for which, see Cappelen and Hawthorne 2009).

b) I've focused on cases where we want to improve the meaning of a particular lexical item. In some cases, we don't care about the lexical item in question, but rather want to introduce a *new* lexical item for the improved or alternative meaning. There's a range of options here and they're discussed in reply to Objection (5) below.

Objection (2): In What Sense Can One Meaning Be Better Than Another?

There's an extensive philosophical tradition for thinking that the concepts or meanings our words express can be defective and can be improved along various dimensions, and so for thinking that some meanings can be better than others. This view can be found throughout the history of philosophy, but in what follows I'll focus on some versions found in the twentieth and twenty-first centuries. Carnap's idea that intellectual work typically involves explication is a paradigm. The core thought was that the meanings we assigned to our words could be defective and for Carnap the central defects were indeterminacy and vagueness. Explication was a process of improvement. For Carnap, improvements were measured relativized to purposes. Here is how Anil Gupta explains the difference between an absolute and purpose-relative improvement:

> An explication aims to respect some central uses of a term but is stipulative on others. The explication may be offered as an absolute improvement of an existing, imperfect concept. Or, it may be offered as a 'good thing to mean' by the term in a specific context for a particular purpose. (Gupta 2015: section 1.5)

I'd like to focus on the idea that an explication is a 'good thing to mean' by the term in a specific context for a particular purpose. Here is an illustration from Gupta:

> The truth-functional conditional provides another illustration of explication. This conditional differs from the ordinary conditional in some essential respects. Nevertheless, the truth-functional conditional can be put forward as an explication of the ordinary conditional *for certain purposes in certain contexts*. Whether the proposal is adequate depends crucially on the purposes and contexts in question. That the two conditionals differ in important, even essential, respects does not automatically disqualify the proposal. (Gupta 2015: section 1.5)

Much the same idea of explication can be found in Quine's *Word and Object* (Quine 1960), where Quine writes about explication:

> We do not claim synonymy. We do not claim to make clear and explicit what the users of the unclear expression had unconsciously in mind all along. We do not expose hidden meanings, as the words 'analysis' and 'explication' would suggest; *we supply lacks*. We fix on the particular functions of the unclear expression that make it worth troubling about, and then devise a substitute, clear and couched in terms to our liking, that fills those functions. Beyond those conditions of partial agreement, dictated by our interests and purposes, any traits of the explicans come under the head of "don't-cares" (§38). Under this head we are free to allow the explicans. (Quine 1960: 258–9)

On this more general understanding of explication, there is no unique correct explication of any term; the improvement is relative to contextually specific purposes. With that in mind, there is no reason why there should be a fixed set of theoretical virtues that are used to measure improvement. In certain contexts, non-theoretical virtues/advantages could make a big difference.

So understood, much work in social and political philosophy can be seen as a continuation of Carnap's proposal. Take, for example, Sally Haslanger's work on gender and race concepts. An important element of that work is a proposal for how gender and race terms can be ameliorated, that is, can be given better meanings. The dimensions of assessment that she has in mind are different from Carnap's, but they are part of the same general project: of improving our concept along *many and diverse dimensions.*

A great deal of philosophy engages in this form of amelioration, often without making a big deal out of it. Consider, for example, Clark and Chalmers' paper 'The Extended Mind' (1998). Clark and Chalmers propose, among other things, that 'A believes that p' be used in a way that makes it true even when p is a proposition that A has access to only with the assistance of various 'external' devices (Clark and Chalmers 1998). In connection with that proposal, they consider various objections of the form: *Well, that's just not how we use 'belief' in English.* In response they say:

> We do not intend to debate what is standard usage; our broader point is that the notion of belief *ought* to be used so that Otto qualifies as having the belief in question. In all *important* respects, Otto's case is similar to a standard case of (non-occurrent) belief [. . .] By using the 'belief' notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and is more useful in explanation.
>
> (Clark and Chalmers 1998: 14)

Clark and Chalmers' goal is not primarily to describe our current concept of belief—they want to *revise* our concept. Note two particularly important features of the proposal. (i) Their revision changes both the extension and the intension of the concept of 'belief'. (ii) They briefly provide a justification for the revision: it is, they say, more useful in explanations. Their new revised concept is also 'deeper' and 'more unified'. So our current notion is defective, as it is not sufficiently unified, not sufficiently deep, and not sufficiently useful in explanations. These kind of ameliorative moves are made in all parts of philosophy. For some salient recent examples think of Haslanger's work on race and gender concepts, Eklund and Scharp on truth, and Railton's proposed revision of core moral terms (see footnote 4 for references). For a wide range of illustrations, see Plunkett and Sundell (MS); Ludlow (2014); and Cappelen (2018: chapter 2). Beyond philosophy it's easy to find lively debates about (what at least looks like) meaning assignments. Consider discussions about what should be in the extension of 'torture', or 'person', or 'marriage', or 'rape'. These are not plausibly construed simply as efforts to find a descriptively adequate account of what these words actually mean. Imagine if a semantically omniscient god told us that 'torture' has a semantic value that doesn't include waterboarding. That's extremely unlikely to stop the debate over whether waterboarding is torture. One way to explain, that is, to construe it as a debate over whether 'torture' *should* have a meaning that makes it include waterboarding in its extension.

So far, I've simply pointed out that there are people who think meanings can be improved. It goes beyond the scope of this chapter to assess each of these proposals.

Suffice it to say that if you find one or more of those views plausible, you should be on board with a version of premise (2) in the Argument.

It is, however, worth considering the denial of the claim that meaning assignments can be assessed. This is the view that there's no normative dimension along which one meaning for W can be better than an alternative meaning. More generally:

Complete Neutrality: Meaning assignments are always normatively neutral.

To assess Complete Neutrality, consider the fact that meaning assignments have a huge influence on *what* we can think about and *how* we can think about those features of the world that we can think about. Both of these have all kinds of effects on humans, both inter- and intra-personally. To hold that those effects cannot be assessed seems implausible for a range of simple reasons: If, for example, Fs are important to a group of agents, then it's good for them to be able to think about Fs and not good if they can't think about Fs. If, however, thinking, theorizing, or discussing Fs is unfortunate for those agents, then nevertheless having an expression that denotes Fs is worse than not having one.

Objection (3): Why Not Think the Meanings Words Have Are the Best They Can Be (or Need to Be)?

I hope Objection (3) sounds silly: It's implausible that a cultural artifact that's generated in a messy, largely incomprehensible way that's outside our control[12] should end up producing something we can't improve on. Everything we humans produce can be improved—why think meaning assignments are ideal (or good enough) straight off? Even if there are degrees of appropriateness, it would be amazing if we got exactly the right degree right off the bat.

A natural thought in this vicinity is nicely captured by Austin. In 'A Plea for Excuses' Austin says that "ordinary language . . . embodies . . . the inherited experience and acumen of many generations of men. . . . If a distinction works well for practical purposes in ordinary life (no mean feat, for even ordinary life is full of hard cases), then there is sure to be something in it, it will not mark nothing" (Austin 1956: 11). Austin might be right: the carvings up that have survived over many generations are likely to 'mark something'. But note that even if this is true, we're not even close to the claim that there's no room for improvement. The Austinian thought gives us reason to think we're not totally wasting our time thinking with, say, the predicates we have, but in no way moves us towards the claim that they can't or shouldn't be improved. This is of course recognized by Austin who goes on to say that '. . . ordinary language is not the last word: in principle it can everywhere be supplemented and improved upon and superseded.' (Austin 1956: 11). The challenge here is to recognize when ordinary language is good enough and when it can be improved upon. This is a deeply normative project, not primarily a descriptive one, and it is continuous with the kind of engineering projects described above. In the passage from *The Will to Power*, quoted above, Nietzsche points out that our concepts are 'the inheritance from our most remote, most foolish as well as most intelligent ancestors'.

[12] Or at least so I argue in reply to Objection (8) below.

Given the impact of the most foolish, it would be naive in the extreme to trust our conceptual dowry in any domain.

Austin connects the survival of what he calls 'a distinction' to the promotion of certain purposes. Again, as Nietzsche points out, these purposes are typically remote and often the purposes of fools. As purposes change (maybe because some of us become less foolish) we need to adjust the way we carve things up. Sometimes this will be the result of moral and political evolutions. In other cases, the changes are driven by needs and purposes that result from technological changes. In yet other cases, the changes are the result of theoretical developments. As purposes change, meanings need to change as well. This is yet another reason to endorse an attitude closer to Nietzsche than to Austin: continuous and radical scepticism towards inherited concepts and distinctions.

Objection (4): If We Change the Meaning of an Expression, Won't That Result in Massive Verbal Disputes and a Change of Topic?

Many philosophers have had concerns roughly of the following form. Suppose 'F' means M. We then ameliorate and in so doing revise the meaning of 'F'. The following will be the result: Pre-ameliorators using 'F' will be talking about something other than those using 'F' post-amelioration. The result of the amelioration will be a change of topic. That, again, can lead to verbal disputes: The pre-ameliorator asserted 'Fs are G' and the post-ameliorators say 'Fs are not G'. It looks like a disagreement, but if there's been a change in meaning of 'F', then there's no disagreement. Moreover, if people pre-amelioration had put massive effort into trying to answer the question: 'Are Fs G?'; it now looks like we've lost track of that question. The question asked by post-ameliorators when they utter 'Are Fs G?' is a different question.

Worries of this sort constantly come up in discussions of revisionary traditions in philosophy. My favorite illustration is Strawson's objection to Carnap's account of explication (in his 1963) but we find the same kinds of concerns in the work of, for example, Haslanger and Railton.

This is a concern I take seriously and much of my (2018) *Fixing Language* is an effort to respond to Objection (4). Here is a summary of the reply:

> We know independently of considerations having to do with conceptual engineering that speakers who use the same sentence, S, with different but relevantly similar semantic contents, can use S to *say the same thing*. So suppose A and B both utter S, but the semantic value of S differs a bit in their two utterances (say, because they occupy slightly different contexts.) We can still, in many contexts, report them by saying that they have both said that S. That's to say, we can use S to say what they have both said. So same-saying can be preserved across differences in semantic content. I suggest we use that as a model for what happens when meanings are ameliorated as described above. The result is a change in, at least, extensions and intensions, but that's consistent with a preservation of same-saying. If it can preserve same-saying, there's a sense in which they both talk about the same subject matter. I call this a preservation of *topics*. So even though the meaning of, say, 'family' can change over time, we can still say that there's

> continuity of topic among those who use ‘family’: they are talking about families. Evidence for this is that two speakers who utter ‘Families are G’, one pre-amelioration and one post-amelioration, can both be described as having said that families are G.

To endorse this you need to buy a collection of claims that that can be summarized as follows:

- First, you have to accept that same-saying can be preserved despite semantic differences. Same-saying data from speakers who use context sensitive terms is an important source of evidence here. An adjective like ‘smart’, for example, will fix a comparison class (or a cut off on a scale) in context. There can be two utterances of ‘Jill is smart’ where the comparison class or cut off differs a bit (i.e., they have different semantic contents), but where it is true to say that both A and B said that Jill is smart. So this is evidence that same-saying can be preserved across semantic difference. I think it’s fair to say that there’s a broad consensus about this and that it constitutes more or less common ground among many of those who think about meaning and communication. (For more on this, see Cappelen and Lepore 2005.)
- Second, you have to accept that this notion of same-saying across semantic difference can be used to establish a notion of topics and topic-preservation that covers what happens when we engage in conceptual engineering. Alternatively, you could take ‘sameness of topic’ as a primitive and use it to explain why we treat speakers as samesayers. Either way, I suggest we treat this cluster of concepts as basic and not aim for a reduction. The core question, for a theory of same-saying or topic-preservation, is what theoretical use they can be put to.

I explore these issues further in section III of *Fixing Language* and my conjecture is that even those who oppose the Austerity Framework that I develop there can endorse this part of it.

Objection (5): Aren’t Meaning Assignments Normatively Neutral, as Long as Each Thing Worth Meaning is Meant by Some Word or Other?

Some things are worth meaning, but it doesn’t matter whether a given word means one of those things or something else. Suppose we’ve established the meaning of a word, W, is defective along some important dimension and we come up with an ameliorative strategy. A natural question is: when implementing this strategy, why keep using W? Since you’re introducing a new (allegedly improved) meaning, why not use a new word to mark the change? This is an important question.[13] In what follows, I’ll articulate it as a choice between Lexical Expansion and Lexical Improvement:

[13] This useful articulation of the objection is due to Alexis Burgess. A discussion with Cian Dorr also helped me get a bit clearer on all of this. For a discussion of related issues and an opposing conclusion, see Dever (Chapter 8, this volume).

Lexical Expansion: Where a new meaning is introduced as the meaning of a new expression.

Lexical Improvement: Where a new meaning replaces the meaning of an already 'in use' lexical item has.

The objection say: We have no reason to prefer Improvement over Expansion. The underlying thesis is Pro-Expansion:

Always-Expand: New meanings should always be attached to new lexical items.

One way to motivate Always-Expand is through thinking about a strategy involved in what Chalmers (2011) calls 'the subscript gambit'. According to Chalmers, most philosophical concepts are surrounded by similar concepts which constitute clusters. For example, there's a bunch of concepts in the vicinity of the concept of freedom. The English word 'free' might pick out one of those, but we should not think that the concept our word happens to pick out is particularly interesting or useful. What we should do instead is explore the entire conceptual neighbourhood, and then introduce a range of new expressions. Chalmers tends to describe these using subscripts, so we have 'freedom_1', 'freedom_2',..., 'freedom_n'. We should keep each of these in our conceptual arsenal, so to speak. They can be useful for different purposes—and having all of them lexicalized enables us to express a wider range of truths.

Chalmers thinks philosophers have spent too much time (thinking they are) fighting over what freedom *really* is by asking the question "Which one of freedom1, freedom_2,..., freedom_n is *really* freedom?" (although they wouldn't phrase it that way, of course). As he sees it, that's a fight over what the semantic value of 'freedom' *simplicter* is, and that is an uninteresting question. Chalmers's thought is similar to the one in the Master Argument: The English word 'free' has ended up with a particular semantic value, but it could easily have had any of a bunch of similar meaning. That it ended up with the particular meaning it has is in part the result of random and intellectually insignificant factors. Chalmers (2011) goes further. He says that obsessing over what that semantic value is can only be motivated by a 'fetishistic' value system.

With that as a background, the motivation for Objection (5) and Always-Expand should be clear: just as it would be irrational to care about which freedom-concept is assigned to 'freedom_1' and which to 'freedom_2', it doesn't matter which freedom concept is assigned to 'freedom'. What matters is that we articulate and lexicalize the range of interesting concepts in this vicinity.

Before turning to my reply, a couple of initial remarks:

- First notice that a proponent of Always-Expand is deeply involved in the core activities of conceptual engineering: Evaluating meanings and reflecting on ameliorative strategies. She is not opposed to conceptual engineering, but is making a particular proposal for how ameliorative strategies should be implemented.
- Choice of lexical items is important. If Always-Expand is true, that's a deep fact about the practice of conceptual engineering. In other words, I don't think this is just a trivial issue about words, because issues about words are hardly ever trivial.

That said, I think Always-Expand is false: Sometimes it is important to preserve the lexical item. Here are four considerations against Always-Expand:

1. SOMETIMES WE CARE ABOUT LEXICAL EFFECTS

The lexical item *itself* can have cognitive and non-cognitive effects on us that we want to preserve. In my *Fixing Language* (2018: chapter 11), I call these 'lexical effects.' Here are some illustrations:

- Brand names: There is a reason why companies spend an enormous amount of resources protecting their brand names. 80–90% of the value of the Coca-Cola company lies in its ownership of the name 'Coca-Cola'.[14] What does that mean? It is of course complicated, but one thing it means is that if Coca Cola had to change the name of its core product, then the value of the company as a whole would decline dramatically. What's important for our purposes is this: in a scenario where the company was not allowed to put the words 'Coca-Cola' on their product, people's propensity for buying and consuming the product would decline. This proves that choice of lexical item matters. A proponent of Always-Expand wouldn't be able to convince Coca-Cola executives that the name of their product is irrelevant.
- Names of children: That words' lexical properties affect people in interesting ways is shown by a study on how political affiliation influences the way parents name their children. As surprising as it might (or, on further reflection, might not) seem, there is a correlation between what a child is named and the political affiliation of the parents who name it. In particular, a study suggested that parents in liberal neighbourhoods are more likely to opt for 'soft' letters in naming their child, such as 'l's and 'm's, while conservative people are more likely to opt for harder sounds like 'k' or 't'. Thus baby Liam is more likely to be the product of liberals while Kurt might follow in his parents' footsteps and become conservative.[15]
- The kind of signalling that is involved in name choices is, again, not about the meaning of the name (or the meaning of sentences containing the name). It is about triggering certain kinds of lexical effects and what the study shows is that parents' choices are guided by lexical effects even when they are not aware of it.
- Finally, some of the debate over the term 'marriage' illustrates the point. The word 'marriage' has a certain effect on people and in the debate over same-sex marriage it was important for proponents that the lexical item 'marriage' was used about their relationship. To see that, note that the following wouldn't suffice to meet the demands of all proponents of same-sex marriage: a proposal to introduce another term—say, 'zwagglebuggle'—that denotes the same rights and obligations as 'marriage' and same-sex couples could say that they were 'zwagglebuggled', but weren't entitled to use the term 'marriage' about their relationship. One reason why some proponents of same-sex marriage would

[14] The economist Aswath Damodaran gives an analysis of the value of the Coca Cola brand name in a blog post at http://aswathdamodaran.blogspot.com/2013/10/the-brand-name-advantage-valuable.html

[15] https://www.livescience.com/37196-politics-baby-names.html

reject this is that the lexical item 'marriage' has important cognitive and non-cognitive effects and those are important in the debate over 'same-sex marriage'. The aim is, at least in part, to change of meaning of 'marriage'.[16,17]

These are not isolated examples. As Chalmers point out:

> Ideal agents might be unaffected by which terms are used for which concepts, but for nonideal agents such as ourselves, the accepted meaning for a key term will make a difference to which concepts are highlighted, which questions can easily be raised, and which associations and inferences are naturally made. (Chalmers 2011: 542)

Many words have a massive effect in social, political, legal, medical, and interpersonal contexts. We have no reason to think theoretical contexts are immune to these kinds of effects. In cases where preservation of lexical effects are important, amelioration will involve changing the meaning of the current word, not introducing a new word. I think lexical effects are ubiquitous and of enormous importance to communication. They are poorly understood and under-investigated (chapter 11 of my *Fixing Language* provides the beginning of a theory).

2. THE ORIGINAL LEXICAL ITEM AS MARKER OF TOPIC CONTINUITY

In reply to Objection (4), I outlined an account of topic continuity across semantic changes. I'll use 'freedom' as an example. Suppose that at time t 'freedom' denotes a certain property, P. Then a semantic change happens, and as a result, at time t_1, 'freedom' denotes a different property, P*. This does not prevent there from being continuity of topic in uses of 'freedom' at t and t_1. Speakers who, at t, utter 'Freedom is G' and speakers who, at t_1, utter 'freedom is G' can *say the same thing.* They can all be talking about freedom and say about it that it is G. The topic—freedom—can be preserved through semantic changes. That is an important kind of continuity in discourse. This continuity is why we can say about speakers at t and t_1 that they are trying to answer the same questions and it's what underpins their agreement. Now suppose we left the word 'freedom' behind and instead introduced indefinitely many new words 'barakuns','hostomas', 'notacabil', etc., for each new meaning. In that case we would lose an important marker of discourse continuity. The connection to previous discourse would disappear. Continuity of lexical item is an important marker of topic continuity.

In response to what I said above, a proponent of Always-Expand could say:

> You say continuity of lexical item is important, but is it essential? It might be no more than a contingent heuristic—useful for restricted agents like humans, but in principle dispensable. Ideally we should just name all the different properties in the freedom neighborhood, and then some of these (or many of them) would constitute topic continuity. An ideal agent wouldn't need lexical continuity to recognize that continuity.

[16] I'm here assuming, for illustrative purposes, that this involves a meaning change. It's an open question whether it does.

[17] This is a point Chalmers agrees with: he is clear that in some—maybe many—cases, words matter (see the discussion of 'torture' in his 2011). And just for the record: Chalmers himself isn't a proponent of Always-Expand.

In reply, I would point out that the communicative features that are important for non-ideal agents matter to us. Humans are very much non-ideal and lexical continuity is what we—with our limited minds and fragile access to semantic and communicative content—often need to track topic continuity. Maybe it's not needed by gods or by massively improved humans but that does not make it less important for us (who are not gods or cognitively enhanced). Here is one way to see why we need it: We care about inter-contextual and inter-conversational continuity. We can individuate conversations so that they last over a long time, take place in different places, and involve a broad range of people who will never know one another. People who will never meet each other can participate in a conversation about, say, democracy and terrorism. Technology has changed the paradigm of a conversation: it's no longer a group of people standing within hearing distance of each other, it is, rather, people transmitting data across the internet to people they might never encounter. Since words don't come with little definitions attached or lines that mark topic continuity, we often have to use lexical continuity as markers of topic continuity. What ties conversations about democracy or terrorism together is often 'democracy' and 'terrorism'. If someone started using 'swugleding' for an ameliorated meaning of 'terrorism', that would most likely fail to connect to the continuous conversations about freedom. There are millions of existing tokens of 'terrorism' and the best way for you to connect to those is to use 'terrorism'. 'Swugleding' likely won't do the work, no matter how carefully introduced. Even 'terrorism*' will most likely fail because hardly anyone will have access to the meaning of the '*'.[18]

3. THE ANCHORING ROLE OF THE ORIGINAL LEXICAL ITEM

To see what I have in mind, consider again the subscript strategy. It start with freedom and then finds properties in the vicinity (or neighbourhood) of it. The subscript strategy presupposes that there is a freedom *cluster*. It presupposes that some properties are within the freedom cluster and other properties are not. For example, the property of being one of my eyes isn't in the freedom cluster (speaking in the subscript lingo: it's not a candidate for being one of 'freedom$_1$', 'freedom$_2$', etc.). That's because it's not one of the properties in the neighbourhood of freedom. The metaphors of 'neighbourhood' and 'clusters' play important roles in the description just given and we need an account of what demarcates the clusters or neighbourhoods. I tacitly introduced such a criterion (in connection with the example of freedom): it has to be appropriately/relevantly related to freedom (without a subscript). We have no other place to start. When theorizing we start with freedom (or, more generally, one of the non-subscripted lexical elements) and then we find properties that are related to it. If that's the general strategy, we need 'freedom' simpliciter as the anchor point for amelioration. It is what topic continuity (being about freedom) is measured relative to. If something like that is right, then an appeal

[18] Note: I am not saying that it is always important to preserve a lexical item. The claim is that lexical preservation often matters.

to freedom simpliciter might be theoretically *indispensable* if you want to preserve topic continuity.[19,20]

4. THE ROLE OF LEXICAL ITEMS IN SOCIAL ONTOLOGY

According to many views of social ontology, language plays an important role in the creation and preservations of social facts.[21] There isn't much agreement on just what that role is, but there is fairly broad agreement that it plays a role. Much here is going to depend on the exact role language plays and which part of language is most important, and settling these issues goes beyond the scope of this chapter. With all those reservations in place, there is an important conditional claim:

> *Possible Connection between Lexical Items and Social Facts*: If lexical items play some role in the creation and preservation of social facts, then changing the meaning of a lexical item might contribute to a change in social reality.

Illustration: Suppose a term like 'family' plays a role in creating and sustaining the social category of families. If so, then improving the meaning of 'family' can be contribute to a change and maybe also an improvement in that part of social reality—it can change/ameliorate the social category of families. Suppose, instead, we introduce a new term, 'scramies' with the improved meaning. That might not have the same effect.

This point is made with a lot of 'might's and 'maybe's. We don't yet know enough about the role of language (and expressions in particular) to make confident claims here. It's an important issue that conceptual engineers should explore further.

Objection (6): Why Think the Importance of the Revisionist Project Undermines the Importance of the Descriptive Project? Why Think There's a Tension between the Two Approaches? Aren't They Complementary?

The Objection: In section I of this chapter I described a *tension* between a descriptive project and a revisionist project. But why aren't they complementary projects? You can do some describing and some amelioration, and these go hand in hand. Moreover, I have described conceptual engineering as a two-step process: first describe deficiencies and then develop ameliorative strategies. That first step is at least in part descriptive,[22] so the revisionist project presupposes the descriptive project. Setting it up as a conflict is misleading.

Before replying to this objection, I should note that it contains an important element of truth. There's no inconsistency between the two projects; they could

[19] None of this is to deny that you could just ignore topic continuity and describe properties independently of what 'clusters' or 'neighbourhoods' they belong to. Nor is it to deny that having created the cluster, we couldn't then shift emphasis to one of the others (i.e., the anchoring doesn't have to be continuous).

[20] See chapter 17 of my (2018) *Fixing Language* for more on this line of thought.

[21] For a paradigm of the kind of view I have in mind, see Searle 1995: chapter 3. See also section 4.6 of Brian Epstein's *Stanford Encyclopedia* entry on Social Ontology (Epstein 2018) and the references in that entry.

[22] That it's a defect is of course a normative judgement, but presumably that judgement relies on a description of what it is like (or what properties it has).

complement each other. Despite this, I think it's accurate to talk in somewhat loose terms about a 'tension' between the two projects. Here is why:

Reply 1: The goals and purposes you have when doing research guides much of what you do. For example, consider someone interested in how cells grow and divide as a purely intrinsically interesting topic. Their research will be guided in very different directions compared to someone who is interested in understanding cancer and is doing the research in order to develop a cure for it. There's obviously no incompatibility between the two projects, but the focus will be immensely influenced by the goal. For a closer-to-home example, consider someone interested simply in the syntax and semantics of 'true' compared to someone interested in 'true' because of the semantic paradoxes. This will give rise to very different research projects and there's a tension in this sense: a lot of what the one is doing will be irrelevant to the other because their research direction and priorities will differ significantly. Or compare the following two: on the one hand, someone interested in the semantics and syntax of terms like 'woman', and on the other, someone like Sally Haslanger who is interested in how such terms ought to be used to classify group for political purposes. Again, there's no incompatibility between the two projects, but the differences in goals will lead to difference in priorities and this again will shape radically different research projects. However, none of this is to say that these differences in goals, priorities and direction imply that the ameliorator shouldn't in part be guided and restrained by descriptive insights.

Reply 2: What is impossible—or at least incompatible with the Master Argument—is to be a 'pure' descriptivist. A descriptivist who claims to have no interest in or need for conceptual engineering shows a lack of understanding of the Master Argument. A corollary of that argument is that all the concepts involved in describing conceptual engineering should themselves be subject to critical assessment. So should the concepts used to describe and execute the descriptive project. We have to assess and improve on the following concepts: 'concept', 'conceptual defect', 'descriptive work', and so on. In other words, the very terminology with which you engage in the descriptive project is itself subject to assessment. So a 'pure' descriptivist would show a lack of understanding for the need to assess and improve on the concepts used to engage in the descriptive project (and even the concepts involved in describing the contrast between the descriptive and the ameliorative project).

Objection (7): If We Are to Engage in Conceptual Engineering, Don't We Have to Assume that Meaning Assignments Are Within Our Control? If They Are Out of Our Control, How Can We Meaningfully Engage in Conceptual Engineering?

The objection: Setting up the objection will require slightly more work than in the previous cases. Here is the basic thought: Conceptual engineering, I claim, involves identifying representational defects and then finding ameliorative strategies. This seems to imply (or presuppose) that successful conceptual engineering requires that meaning assignments are in large part within our control. Why construct ameliorative strategies if we can't implement them?

One reason I take this objection very seriously is that I think meaning assignments are in large part incomprehensible and outside human control. In what follows I first explain why, and then reply to the objection.

Background for the objection: The metasemantic facts are out of control and inscrutable. I focus on what two broad kinds of metasemantic theories tell us about meaning determination: externalist theories and internalist theories. On the standard externalist story, content determining factors include:

- Introductory events that typically happened a long time ago and were performed by people we don't know anything about. The introductory events will be of two broad kinds: demonstrative ("let 'F' denote those kinds of things") or descriptive ("let 'F' denote the things that are G"). Note that if this is right, then in many cases the facts surrounding these introductory events will be unknowable to us. We have no way to access what happened: there will be no written record and, absent a time machine, we can't know what happened. We don't know what was pointed to and we don't know the details of the descriptions used. Moreover, we don't know the motivations for these introductions.
- Externalist theories also appeal to chains of reference transition. In such communicative chains, expressions are 'passed along' from one speaker to another and there is some kind of reference-preserving element in the communicative chain. Sometimes those chains are not reference preserving; in such cases, we get reference shifts.
- Other externalists, such as Timothy Williamson, talk more generally of meaning as supervening on use patterns over time where this connection between use and meaning (or reference) is chaotic in the sense that there's no algorithm that takes us from use patterns to meanings. The connection is too complex—indeed it might be in principle too complex for humans to grasp. We can't know about all the particular uses, and even if we did know about them, the way in which meaning is generated by such use is too complex for the human mind to grasp.

An important feature of these kinds of views is that we are in large part *not* in control of the reference determining process. We're not, for example, in control of what happened far into the past and we're not in control of what happened in the transition periods. In general, no one is in control of the total pattern of use. Nor are we in control of the supervenience relation that takes us from complex use-patterns to meanings.

It's tempting to think that if your metasemantics is more internalistic, meanings would be more within our control and so meaning change would be easier. Burgess and Plunkett, for example, say:

> The textbook externalist thinks that our social and natural environments serve as heavy anchors, so to speak, for the interpretation of our individual thought and talk. The internalist, by contrast, grants us a greater degree of conceptual autonomy. One salient upshot of this disagreement is that effecting conceptual change looks comparatively easy from an internalist perspective. We can revise, eliminate, or replace our concepts without worrying about what the experts are up to, or what happens to be coming out of our taps.
>
> (Burgess and Plunkett 2013: 1096)

I agree with one part of this: from the point of view of an internalist theory, meaning change and revision is possible. That's all that's needed to support the second premise in the Master Argument. However, and this will be relevant later, I disagree with the claim that internalism, as such, puts us more in control of these changes. Internalism is a supervenience claim: the extensions and intensions of expressions supervene on individuals (and then lots of bells and whistles to elaborate on this in various ways, but the bells and whistles don't matter right now). Suppose the meaning of my words supervene in that way on *me*. Note first that this is compatible with the meanings and extensions supervening on features of me that I have no control over. It's also compatible with it being unsettled and unstable what combination of internal features ground reference. So there's just no step from Internalism to control. Moreover, even if meaning supervenes on something internal that I have control over, control doesn't follow: it could supervene on something we could control, but the determination relation from the supervenience base to meanings/extension could still be out of our control. For example, even if there's supervenience on what we want or intend or decide, the supervenience relation doesn't have to make it the case that semantic values are what we intend for them to be, what we want them to be, or what we agree on them to be (for all we know, it could be a total mess or get us to the opposite of what we want, intend, or decide).

In sum: both externalist and internalist theories makes meaning change possible, though both of those theories make it hard to see how this is a process we can be in control of. This is why Objection (7) is pressing.

Reply: In summary form, my reply is that conceptual engineering shares this feature with most normative theorizing. On the view I defend, the tools we think with are often defective, but there's very little we can do about it. We can talk and think about it, but doing so has hardly any effect. Compare that to me talking to you about crime in Baltimore, poverty in Bangladesh, or the Trump presidency. Such talk is unlikely to have any effect on what happens in Baltimore, Bangladesh, or with Trump. The ineffectiveness of talking is an almost universal aspect of large-scale normative reflections. Anyone who spends time thinking and talking about large-scale normative matters should do so without holding out too much hope that their talking and thinking will have significant or predictable effects on the relevant aspect of the world. If you think your views and theories about crime in Baltimore, poverty in Bangladesh, or the Trump presidency will have a significant or predictable effect on either, you're extremely likely to be disappointed (and to end up feeling you've wasted the part of your life that has been devoted to these issues). There are of course small-scale local issues where normative reflections will have a direct effect. If I think my daughter shouldn't have an ice cream, then, at least in a few cases, the result will be that she eats no ice cream. Moving to slightly larger-scale issues—say speed bumps in the street where I live—my opinions, views, and pleadings will have tiny effects, but already these effects will be fairly marginal, unsystematic, and unpredictable (as I've discovered). On the view proposed in Cappelen (2018), changes in extensions and intensions of words are far over on the large-scale and unpredictable side. Much closer to crime in Baltimore than to speed bumps in Sofies Gate.[23] So, in sum, the worry that I've painted too bleak a picture of the prospects of conceptual engineering simply fails to take into account the relevant comparison class. What I say about conceptual engineering shouldn't be surprising and doesn't make the activity of trying to engineer concepts much different from a wide range of other human efforts to think about how things should be.

Conclusion

The Master Argument provides a general argument for the importance of conceptual engineering. It has nothing to say about *particular* deficiencies or ameliorative strategies. As a heuristic it's useful to think of conceptual engineering as having two parts: the general theory and the specific applications. We should expect a two-way interaction: the general theory will inform the specific cases and the specific cases will inform the general theory. It should also be clear from the discussion above that there can be many frameworks for thinking about conceptual engineering. What one takes conceptual engineering to be (when thinking about both the general theory and specific cases) will be shaped in large part by what one takes meanings and concepts to be, what one assumes about metasemantics, and what one takes to be conceptual defects and virtues. One advantage of the Master Argument is that it is neutral on those questions and so can provide a kind of common ground for all those who see conceptual engineering as central to philosophy.

Acknowledgements

Much of the discussion in this chapter draws on Cappelen (2018). This can be seen as an elaboration of an argument in section II of that book, but this is the more refined version of the argument (better than the one I present there). I presented this chapter as a paper at the Philosophy Mountain Workshop in Telluride and at the Arché Conceptual Engineering in St Andrews. I'm grateful to all the participants for helpful discussions and in particular to Alexis Burgess, David Plunkett, Kevin Scharp, and Seth Yalcin for detailed comments.

References

Austin, John. 1956. A Plea for Excuses. *Proceedings of the Aristotelian Society* 57:1–30.

Cappelen, Herman. 1999. Intentions in Words. *Noûs* 33:92–102.

Cappelen, Herman. 2018. *Fixing Language*. Oxford: Oxford University Press.

Cappelen, Herman, and Lepore, Ernest. 2005. *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Wiley-Blackwell.

Cappelen, Herman, and Hawthorne, John. 2009. *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Cappelen, Herman, and David Plunkett. Chapter 1, this volume. Introduction: A Guided Tour of Conceptual Engineering and Conceptual Ethics. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

[23] And as I just said, even the speed bumps turned out to be more or less completely out of my control and I ended up concluding that this particular instance of 'local' activism was a waste of time. The conclusion is not that we shouldn't do it, but rather that if we do it, we should do so without illusions.

Chalmers, David J. 2011. Verbal Disputes. *Philosophical Review* 120 (4):515–66.

Clark, Andy, and Chalmers, David J. 1998. The Extended Mind. *Analysis* 58 (1):7–19.

Dever, Josh. Chapter 8, this volume. Preliminary Scouting Reports from the Outer Limits of Conceptual Engineering. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Eklund, Matti. 2014. Replacing Truth? In Alexis Burgess and Brett Sherman (eds.), *Metasemantics: New Essays on the Foundations of Meaning* (pp. 293–310). Oxford: Oxford University Press.

Epstein, Brian. 2018. Social Ontology. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (summer edn). https://plato.stanford.edu/archives/sum2018/entries/social-ontology/

Gupta, Anil. 2015. Definitions. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (summer edn). http://plato.stanford.edu/archives/sum2015/entries/definitions/

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Kaplan, David. 1990. Words. *Proceedings of the Aristotelian Society*, Supplementary Volumes 64:93–119.

Ludlow, Peter. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford: Oxford University Press.

Nietzsche, Friedrich. 1901/1968. *The Will to Power*. Trans. W. Kaufmann. New York City: Random House.

Pasnau, R. 2013. Epistemology Idealized. *Mind* 122 (488):987–1021.

Plunkett, David, and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Plunkett, David, and Sundell, Timothy. MS. Work on Conceptual Engineering [title to be determined].

Quine, W. V. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Railton, Peter. 1989. Naturalism and Prescriptivity. *Social Philosophy and Policy* 7 (1):151.

Railton, Peter. 1993. Noncognitivism about Rationality: Benefits, Costs, and an Alternative. *Philosophical Issues* 4:36–51.

Scharp, Kevin. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Searle, John. 1995. *The Construction of Social Reality*. New York: Free Press.

Strawson, P. F. (1959). *Individuals*. London: Routledge.

Strawson, P. F. (1963). Carnap's Views on Conceptual Systems versus Natural Languages in Analytic Philosophy. In Paul Arthur Schilpp (ed.), *The Philosophy of Rudolf Carnap* (pp. 503–18). London: Open Court.

8

Preliminary Scouting Reports from the Outer Limits of Conceptual Engineering

Josh Dever

1. How to Argue about the Planets

Brown and Black are arguing about Pluto. Brown holds that Pluto is not a planet, in light of its similarities to paradigm non-planets such as Makemake, Eric, Haumea, and Quaoar. Black, on the other hand, holds that Pluto is a planet, in light of its similarity to paradigm planets such as Mercury, Venus, Earth, and Mars. (This is a tendentious and problematic way of describing the content of their disagreement. Be patient—it's early days, and Brown and Black will increase in philosophical sophistication soon.)

For a time, Brown and Black's dispute just amounts to citing various Plutonian features. Brown points out that the center of mass of the Pluto-Charon rotational system lies outside of Pluto. Black points out that Pluto has assumed a spherical shape due to its own gravitational forces. After a while, though, and after each discovering that the other is sometimes strangely unmoved by the considerations the first has advanced, they start to wonder what it is that they are arguing about. One point on which they quickly reach consensus is that they are *not* arguing about whether Pluto is characterized by the English word "planet". They both recall that a *Star Trek: Enterprise* episode (not surprisingly, a rather bad one) featured a rogue planet Dakala, and that Iain Banks' novel *Matter* is set on the artificially constructed shell planet Sursamen. They conclude that while the English word "planet" can apply to bodies that aren't orbiting a star and to human-made objects, nevertheless that *whatever* it is that they are trying to settle about Pluto, it's not whether *that* word applies to it, since they're both interested in some question that excludes the possibility of being an artificial or rogue object. (Brown and Black clearly still haven't gotten over their philosophical naivete and are moving rather too quickly from usage facts to meaning facts. But this won't matter much for we're heading. The point is that even if they had the right sort of evidence about the meaning of the English word "planet", they would in the face of that evidence still not be particularly interested in settling a question delimited by that meaning.)

Josh Dever, *Preliminary Scouting Reports from the Outer Limits of Conceptual Engineering* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020).

DOI: 10.1093/oso/9780198801856.003.0008

Once they set aside the English word "planet", progress (of a sort) follows quickly. Brown suggests categories of PRIMARY PLANET and DWARF PLANET, with (e.g.) Earth and Mars in the first category and Pluto and Eris in the second. Black suggests categories of SOLITARY PLANET and BELT PLANET, with Earth and Pluto in the first category and Eris and Makemake in the second. At this point Brown and Black become tempted to describe their original argument as a "verbal dispute". Whether that's the right thing to say won't be a central issue here (although it's definitely running along tracks proximate to the path I want to explore), but it is right that Brown and Black agree that Pluto is a dwarf planet and a solitary planet and is not a primary planet or a belt planet.

Brown and Black's argument at this point might turn into an argument about *which concepts to use*. Brown suggests using PRIMARY PLANET and DWARF PLANET, and perhaps reserving the label "planet" for PRIMARY PLANET, and hence rejecting the sentence "Pluto is a planet". Black suggests using SOLITARY PLANET and BELT PLANET, and perhaps reserving the label "planet" for SOLITARY PLANET, and hence endorsing the sentence "Pluto is a planet". There can of course be many reasons to have such an argument, and many reasonable ways for such an argument to proceed. We are finite beings, so we have to pick and choose which theoretical projects to engage in, and we can look for reasons for suspecting that one of the two classificatory schemes proposed by Brown and Black will do a better job of furthering various of the goals that an astronomical theory is pursuing.

But there is also a clear sense of "ought" in which what we ought to be doing in this situation is just assembling all of the truths about PRIMARY PLANETS, DWARF PLANETS, SOLITARY PLANETS, and BELT PLANETS. No need for Brown and Black to disagree any longer. Brown can assemble his total theory of PRIMARY PLANETS and DWARF PLANETS, and Black can assemble his total theory of SOLITARY PLANETS and BELT PLANETS, and then we can dump these two theories together into a single megatheory. Brown, after all, agrees with everything Black has to say about SOLITARY PLANETS and BELT PLANETS, and Black agrees with everything Brown has to say about PRIMARY PLANETS and DWARF PLANETS.

2. Big Theory, Little Theory

We can characterize this point in Brown and Black's discussion in terms of the norms of theory selection. Let T1 be the total theory of PRIMARY PLANETS and DWARF PLANETS and T2 be the total theory of SOLITARY PLANETS and BELT PLANETS. Then T3 is the (logical closure of) the union of T1 and T2. How should theorists choose among T1, T2, and T3? There are two basic orientations toward the project of theorizing that we can distinguish here. We can be 'Big Theory' theorists, who think that the goal of theorizing is to say everything there is to be said. The Big Theory theorist favors T3 over T1 and T2. Why leave out the insights of either T1 or T2 when we could just say it all? Or we can be 'Little Theory' theorists, who think that the goal of theorizing is to say everything *worth saying*. The Little Theory theorist might, for example, favor T1 over T2 and T3 on the grounds that the additional information included in T2 and T3 isn't information worth having (perhaps even is information that gets in the way of understanding the world).

The Big Theory theorist doesn't disagree with the Little Theory theorist that some information matters more than other information. But the Big Theory theorist has a 'just more theory' attitude toward that insight. So the Big Theory Theorist might want to add to T3 the further claim:

> PRIMARY PLANET and DWARF PLANET are more helpful categories for understanding astronomical matters than are SOLITARY PLANET and BELT PLANET.

If the Little Theory theorist is *right* in their practical preference for the T1 information over the T2 information, then that claim is true, and the Big Theory theorist thus wants to endorse it as well. The Little Theory theorist worries, however, that the 'just more theory' approach doesn't help separate the important from the unimportant, because we still don't distinguish the important from the unimportant auxiliary claims about rankings among primary claims. (One version of this worry is that if we are Big Theory theorists, we might include auxiliary claims about which of our primary claims are in a maximally natural vocabulary. But we'll also include auxiliary claims about which of our primary claims are in a maximally schmatural vocabulary, for some grue-like variant of natural, and we're then in danger of not knowing whether it's naturalness or schmaturalness that makes for importance. A bit more on this line of concern in the final section of this chapter.)

Of course, no one is really a Big Theory theorist in the simplistic sense set out above. We don't want our theory genuinely to say *everything* there is to be said. As Putnam (1978) observes, if there's anything that's clear, it's that not absolutely everything is to be endorsed. A decent first stab at *why* not everything should be endorsed is that endorsing everything requires endorsing all the falsehoods as well as all the truths, and we don't want to endorse the falsehoods. So a standard more plausible way of being a Big Theory theorist is to endorse the following norm of theorizing:

- **Alethic Theory Selection**: The sole criterion for theorizing is that all and only truths should be endorsed. (Since everything in all of T1, T2, and T3 is true, all of these claims should be endorsed. Since T3 is the meet of T1 and T2, endorsing everything in all three theories amounts to endorsing T3.)

Contrasting with a norm of alethic theory selection is:

- **Pragmatic Theory Selection**: Theories are selected (at least) in part on the basis of pragmatic factors such as explanatory power, computational efficiency, productivity in guiding future research, promotion of social welfare, and so forth. We might thus favor T1 over T2 and T3 on the grounds that the PRIMARY/DWARF PLANET distinction fits well with the principal classificatory aims of astronomy.

This truth-anchored picture of Big Theory theorizing is enough to get us going, but we'll see later that it will be helpful to have a way of thinking about things that isn't tied specifically to truth. We'll thus consider another way of putting things.

3. Conceptual Maximalism and Global Versus Ideal Language Theorizing

One attraction of being a Big Theory theorist is that there is a kind of winnowing down that we don't need to engage in. When confronted with the conceptual diversity of PRIMARY PLANET, DWARF PLANET, SOLITARY PLANET, and BELT PLANET, we don't need to make a choice. (*Qua* theorist. *Qua* person actually doing some investigating, we may need to decide which investigations to do now. But those decisions can be haphazard, idiosyncratic, and unprincipled.) We can be **conceptual maximalists**, welcoming the use of *all* concepts in our theorizing. But if we are pragmatic theory selectors, we do winnow down, since not *everything* goes into the final theory. The winnowing down then can be *haphazard* or *stratified*. Haphazard winnowing doesn't break down nicely along conceptual lines—we might accept some but not all (true) DWARF PLANET claims (where those claims have various pragmatic virtues) and also some but not all (true) BELT PLANET claims (where those have pragmatic virtues).

Standard pictures of theory selection, however, tend to assume that the winnowing will be stratified—that we will *first* pick out a pragmatically privileged collection of conceptual resources, and then *second* build a theory consisting of all true claims using those resources. On the stratified picture, we decide (on pragmatic grounds) to theorize in terms of DWARF PLANET rather than BELT PLANET, and having so decided, we then build the *full* theory of DWARF PLANET, containing all the DWARF PLANET truths.

Stratified pragmatic theory selection is then a form of **ideal language theorizing**. In ideal language theorizing, there is some privileged language, such that a theory couched in that language is theoretically preferred over a theory couched in another language. The ideal language might be the ordinary language of the theorizer (as in Hirsch (2002)), or a logically perspicuous fragment/regimentation of the ordinary language (as in Quine (1948)), or a non-ordinary language well-suited for metaphysical theorizing (Sider (2012)'s 'Ontologese' or a Lewis (1983)'s appeal to a language of maximally natural concepts).

We'll set aside haphazard pragmatic theory selection and focus on the dispute between alethic theory selection and stratified pragmatic theory selection. If stratified pragmatic theory selection is ideal language theorizing, alethic theory selection is **global language theorizing**. No one language is privileged; our theoretical obligation is to work in all languages (or in the meet of all languages, if there is such a language).

(One version of) Carnap is the distinctive global language theorist of our local tradition, providing a distinctive break with the various ideal language inclinations of Frege, Russell, Wittgenstein, and Quine. There are many Carnaps available in the interpretational space; I'm interested in the Universalist Carnap, who takes the external question to be without theoretical significance. What needs to be done is to answer all the internal questions in all the languages; the external question is just a decision about what portion of what needs to be done we are going to do now.

One advantage of recasting the Big Theory/Little Theory distinction as the global language/ideal language distinction rather than the alethic norm/pragmatic norm distinction is that it lets us relax the specific role of truth in guiding Big Theory theorizing. We can take the target Big Theory to be the theory that lists all claims in all languages and for each claim specifies the semantic status of that claim. If 'true' and 'false' emerge as privileged semantic statuses across all languages, this will be isomorphic to the theory yielded by the alethic norm, but with the global language picture, we can engage in theorizing even if we don't know how to pick out a single status that across all languages marks claims as theoretically privileged. Our attention can turn from the normative questions raised by the alethic/pragmatic split to content-based questions raised by the global/ideal split. (A first look at one of those content-based questions will be our central concern here.)

(We shouldn't, though, be too sanguine about the non-alethic global language version of Big Theory theorizing. On this picture, a theory doesn't consist of a collection of claims, but rather of a collection of claim-status pairs. Given that the Big Theory response to Little Theory insistence that theories encode more than mere information (by, e.g., encoding what is important, fundamental, explanatory, and so on) was the 'just more theory' response that these other things could be included in the theory in the guise of further claims, a retreat from theory-as-claims is in danger of undermining a central motivation of Big Theory theorizing.)

4. Conceptual Engineering, Conceptual Ethics, and Xenolinguistics

I've couched things above in the language of some old disputes about metaontology. That was done in part to help highlight the (never very dimly illuminated) fact that the recent conceptual engineering movement is heavily indebted to those old disputes. To bring this portion of the two discussions fully into alignment, let's make the distinction between *conceptual engineering* and *conceptual ethics*:

- Conceptual engineering is the broadly theoretical task of designing new concepts (or changing our old concepts; I won't worry about the trans-engineering identity condition question here).
- Conceptual ethics is the broadly practical task of determining, post-engineering, which concepts we ought to use. (Note that conceptual ethics, as a species of pragmatic theory selection, requires that the pragmatism be of the stratified rather than the haphazard genus.)

The conceptual maximalist is an eliminativist about the project of conceptual ethics. There is no interesting question about which concepts we ought to use: we should use all of them, and build the maximal theory saying all that there is to be said in the most expansive language. That's a tempting eliminativism, especially for those of us inclined to think that the pragmatic considerations don't go deep enough to guide the kind of intellectual enterprise we take ourselves to be engaged in. (Again, this is eliminativism about conceptual ethics as a general issue in theory selection. That's an eliminativism that's compatible with serious and important questions about what *part* of the overall theoretical project any one person should engage in at any one time. But those questions, says the conceptual maximalist, belong in a different disciplinary box. Our general answer to the question of how to theorize doesn't need to do anything to tell us whether Smith, in particular, should be a category theorist or a Marlowe scholar.)

But there is a price to be paid for the conceptual maximalist's eliminativism about conceptual ethics. If there is no ethical question here, and our theoretical task is to say what there is to be said using *all the concepts*, then we had better have something to say about what the range of possible concepts is. The way I've set things up here, that amounts to the question of what kinds of things count as *possible languages*. The central point I want to push here is that the question of what possible languages there are (equivalently, perhaps, what could be said in any possible language) is one of the deepest, most intractable, and most neglected philosophical issues out there. But, as I'll try to bring out, one of the things that makes the territory difficult here is that there's also a decent chance that the question is utterly trivial.

The conceptual maximalist/global language theorist (I'll assume henceforth that the two labels can be used interchangeably) dodges a specific conceptual engineering bullet that hits the non-maximalist directly: the conceptual maximalist doesn't need to decide which of the PRIMARY PLANET/DWARF PLANET and SOLITARY PLANET/BELT PLANET distinctions is more worthy of enshrining in our theory. But, of course, these aren't the only astronomical concepts available. There's also the PLANET WITH RINGS/PLANET WITHOUT RINGS distinction, and the STELLAR BODY WITH MORE SOLID THAN GASEOUS VOLUME/STELLAR BODY WITH MORE GASEOUS THAN SOLID VOLUME distinction, and many many more. Very quickly we reach the point of observing that a collection of N astronomical bodies immediately gives us 2^N astronomical distinctions. And even that's only the beginning. By adding temporal and modal intensional dimensions, our available distinctions presumably become robustly infinite, and numerous grue-ish categories emerge. The conceptual maximalist isn't *bothered* by the inclusion of grue-ish categories—there's no commitment to everything in the theory being projectable, and as usual the theory itself can include commentary on which tracked categories do project—but there is still a lot of theory to compile.

That sort of predicative plenitudinousness, though, only scratches the surface of possible languages. Minimally, we can make similarly plenitudinous moves at other semantic categories. There are quite a lot of generalized quantifiers, or adverbs, or modal operators, or (((e,t),t),(((e,t),(e,t)),(t,e)))-category expressions, that we can add to our expressive resources and then require the conceptual maximalist to theorize about. Such plenitudinous moves expand the language by adding more inputs to the same underlying semantic infrastructure—a basic term-predicate structure, or its generalization to a full categorical grammar. But a genuinely *global* language needs to entertain other more alien ways that languages could be structured. The theory of the global language theorist needs to include the contents of map-like representations and the contents of non-conceptual perceptual experiences. We need to work out whether there are claim-like (theory-worthy) contents captured using infinitary, non-well-founded, or gunky versions of semantic type theory, using scoreboard update procedures or strategies for scoreboard manipulation or equilibrium points for massively multi-player scoreboard scheming tactics, using relations between fundamental groups of non-Hausdorff topological spaces, using methods of manipulation patterns of social prestige markers, using functions from n-tuples of alien phenomenal states to dispositions to adjust alien quasi-epistemic attitudes, and so on. Linguistics to date has been merely domestic linguistics; examination of the languages we happen to find around us. Global language theorizing calls for an ambitious xenolinguistics, in which we consider what languages there could have been. (In particular, that part of xenosemantics which is concerned not with the question of how xenomeanings are correlated with and derived from xenosyntax, but with what the xenomeanings are.) We haven't even started on the project of xenolinguistics.

Maybe we get lucky, and there's no real need for xenolinguistics. Maybe all of the potential exotica above, if they're capable of being deployed in the representational devices of a genuine language, end up encoding contents that are already made available by the semantic tools made available in English. If so, English is already a maximal language, capable of saying anything that can be said in any language. If not, there is at least local ineffability—claims that we can't express with English. There might be stronger versions of ineffability—claims that can't be expressed in any human language, or can't be thought or grasped by human minds. But whether there is ineffability or not, the conceptual maximalist, in order to say what would even count as the Big Theory, needs to answer:

- **The Boundary Question**: What is the range of, or the characteristic feature of, possible languages and possible things to be said in possible languages?

The Boundary Question isn't a question about what expressions could stand in the expressing relation to what contents, or a question about what it takes for speakers to be speakers of a language. (It's thus not asking about Lewis's *actual language* relation.) It's rather a question about what contents there are to be expressed. At a first draft, it can be taken as the question *what propositions are there?*, or *what does it take for a content to be a proposition?*, although I don't want to build in to the question the assumption that propositions are uniquely the kind of thing a language must express.

If you're a serious conceptual ethicist, you can avoid wrestling with the Boundary Question. If you go in for some conceptual ethics, your pragmatic norms can pick out a privileged ideal language of theorizing. With the language picked out, you've got your domain of theorizing in hand, and you won't need to explore the boundaries of conceptual possibility in order to carry out your theorizing task. In practice, what this comes down to is that your conceptual ethical concerns give you a prior picture of what kind of concepts might fulfill your theorizing goals, and so the conceptual engineering task gets constrained from the beginning. You might not know in advance what language will end up being the ideal one, but you can see *roughly* where the ideal language is located, and not venture too far outside that region in your engineering. DWARF PLANET and BELT PLANET get a look in, but even relatively parochial exotica such as IT PLANETIZES SOLITARILY and PERIHELION-PLANET/APHELION-PLANET (compare INCAR and OUTCAR) aren't the kinds of things we want our engineers to build.

Of course, not needing to answer the Boundary Question doesn't go make the Boundary Question go away. (If you're very lucky, your conceptual ethics might even make the question go away. Suppose your conceptual ethics are that we should adopt whatever concepts maximize overall human flourishing. (Black wins the planetary dispute, e.g., because he avoids saddening schoolchildren everywhere by depriving

Pluto of its status in the pantheon.) Perhaps then a language (a conceptual repertoire) just *is* a tool for maximizing overall human flourishing—in the end, when confronted with some practice and asked whether that practice amounts to a language, there is nothing more to do than to see whether that practice has as an aspect promoting human flourishing.)

And of course, there is some sense in which the conceptual maximalist doesn't have to answer the Boundary Question. There's lots of global language theorizing to do prior to engaging in some xenolinguistics. But an understanding of the full scope of the global language theorizing project, unlike an understanding of the full scope of the ideal language theorizing project, does call for an answer to the Boundary Question.

5. Problematic Languages and Limiting Damage and Exposure

The project for the remainder of this chapter, then, is to make some preliminary forays into answering the Boundary Question, mostly with an eye to demonstrating that easy answers aren't going to work. We'll consider various delimiting criteria stating what the range of possible languages is, and extract some overarching morals about the kinds of difficulties these criteria get into. To give a little extra punch to the Boundary Question, we begin by noting some potentially troubling commitments of conceptual maximalism, commitments that we will then hope that an adequate answer to the Boundary Question will help us avoid. The conceptual maximalist bears a theoretical commitment to build a maximal theory collecting up all of the truths using all of the concepts in all of the possible languages. When we are dealing with Brown and Black, the maximalist route of theorizing using *all* of DWARF PLANET, BELT PLANET, and so on looks pleasingly cosmopolitan. Even in more pragmatically loaded settings, the maximalist approach has an appeal. We might think that on due consideration the normatively weighty concept of RACISM is one of systematic ill-treatment based on perceived racial classification deriving from a history of institutional oppression and power inequities, making the idea of "reverse racism" incoherent. But we can also easily acknowledge the concept of RACISM* that drops the institutional oppression and power inequity requirements, and build a total theory that collects truths about both RACISM and RACISM*, while expecting that few *normative* truths will involve RACISM*. But things aren't always so easy for the conceptual maximalist. Consider the following:

- Will the conceptual maximalist be required to include all of the (insert your favorite racial/gender/religious/etc. slur here—I'll use "freethinker" both so that I'm in the slurred group and because I like reclaiming that eighteenth-century slurring feel to the term) truths in the total theory? We might have hoped that "Jones is a freethinker" and "Freethinkers lack a moral compass" are the sorts of things that the *bigot* is committed to, not the sort of thing that we as ideal theorizers are committed to. But if there is a FREETHINKER concept then the conceptual maximalist must use that concept in theorizing. (The conceptual maximalist can always hope that there are no FREETHINKER truths to include in the total theory, despite the presence of the FREETHINKER concept. Probably "Freethinkers lack a moral compass" can be avoided. "Jones is a freethinker" is harder. And avoiding *any* FREETHINKER truths presumably calls for some logical revisionism.)

- Will the conceptual maximalist be required to include all of the SHERLOCK HOLMES truths in the total theory? The conceptual maximalist is going to mention a lot of strange and unfamiliar objects in the total theory—belt planets, incars, book-like objects that exist only while it is unethical to open them to page 37. But does the conceptual maximalist need to mention Sherlock Holmes in the total theory? We might have thought that "Sherlock Holmes is a detective" is the sort of thing *John Watson* is committed to, not the sort of thing we as ideal theorizer are committed to. But if there is a SHERLOCK HOLMES concept, then the conceptual maximalist must theorize using it. (As before, we can hope there are no SHERLOCK HOLMES truths to be included. As before, realizing that hope creates pressure for logical revision.)
- Will the conceptual maximalist be required to include all of the TONK truths in the total theory? Letting in *any* TONK truths looks dangerous, because given the constitutive inferential rules of TONK, once some TONK claims go in, *all* claims go in. We might have thought that TONK claims were the sort of thing the logical deviant was committed to, but not the sort of thing that we as ideal theorizers should countenance.
- Will the conceptual maximalist be required to include the Eiffel Tower in the total theory? Not (the familiar concept) THE EIFFEL TOWER, and not the physical Eiffel Tower as (say) a constituent of a Russellian singular proposition, but the tower itself as the 'propositional' content of an utterance in a possible language?

In general, there's a lot of potential weird conceptual junk out there, and it would be nice if the conceptual maximalist had a way to produce a respectable junk-free total theory. There are two approaches here. One approach is **Limit the Damage**. On this approach, we countenance the "defective" concepts, but we argue that those concepts, because of their defectiveness, don't manage to feature in any true claims, and thus don't get into the total theory. The other approach is **Limit the Exposure**. On this approach, we find grounds for declining to countenance the "defective" concepts. If there are no languages that use the (putative) concepts FREETHINKER, SHERLOCK HOLMES, and TONK, and no languages in which the Eiffel Tower is the content of a claim, then the conceptual maximalist has nothing to fear from these cases.

Clearly the plausibility of the "no such language" line is increasing as we proceed through the examples. After all, English (we might think) *does* contain the concepts FREETHINKER and SHERLOCK HOLMES, so those two cases at a minimum do represent parts of possible languages. But we shouldn't be too quick to conclude that English is indeed a language. Perhaps it merely *appears* to be a language, but is prevented from being one by its deployment of "defective concepts". (In the same way that we might say that *Sherlock Holmes is a detective* merely *appears* to be a thought.) In the end there's a probably a choice here between a more expansive use of the term "language", on which English definitely counts as a language and on which the conceptual maximalist is committed only to theorizing uses the resources of all languages bearing some good-making feature, and a less expansive use of the term "language" which reserves the term as an honorific for cases in which the good-making feature is present. I doubt it matters which way we go (as befits a conceptual maximalist); I'll talk in the latter way henceforth.

I'm going to seek to **Limit the Exposure** rather than **Limit the Damage**. That's in part because I think the damage is hard to limit, and the Boundary Question needs to be answered even if we decide to limit damage (and answering the Boundary Question is a way to limit the exposure). It's also in part because **Limit the Damage** requires the specifically *alethic* formulation of the non-pragmatic theoretical enterprise, and (as hinted earlier, and as will come out soon) I think it's hard to hold on to the alethic formulation as we investigate the Boundary Question.

6. Answering the Boundary Question

So much for preliminaries. What, then, is the possible range of languages? We will consider two general strategies for answering the Boundary Question. One strategy is to extract an answer from our understanding of semantic theories—we check whether the existing semantic toolkit offers an answer to the question 'what sentence-level contents are available to be expressed" (in, e.g., the way that the standard semantic toolkit offers as an answer to the question 'what available quantifier-level contents are available to be expressed' the category ((e,t),t)). The second strategy is to extract an answer from our *metasemantics*, by examining how our understanding of the theoretical role of contents delimits what entities could play that role.

First Semantic Attempt: Possible Worlds

Let's start with an answer that drops naturally out of one popular framework for semantic theorizing. Propositions, many people say, are sets of possible worlds. One nice thing about this doctrine is that it immediately tells us what the full range of propositions is: the power set of possible worlds. The maximal language, then, is the language that allows expression of each member of the power set. (Note that we're concerned with the language only with respect to what the language expresses, not with what vehicle it uses in doing the expressing. So even if, for Kaplan paradox reasons, we can't come up with enough vehicles to express all the contents, we needn't worry about that limitation for current purposes. We similarly won't be interested in the question of whether anyone could ever entertain all of the resulting contents, or even whether each of the contents is possibly entertainable, again sidestepping Kaplan paradox worries.) The conceptual maximalist is then committed to building a theory that settles each of these propositions (and which, as a result, decides exactly which world is the actual world). PLUTO IS A DWARF PLANET picks out one set of worlds containing the actual world (which thus goes in the total theory); PLUTO IS A SOLITARY PLANET picks out a different set of worlds containing the actual world.

I'll focus on two worries about this possible worlds answer to the Boundary Question—both selected because they are relatively easy versions of worries that will hound more sophisticated approaches as we go. The first concern as one of **Undue Expressive Limitation**. The possible worlds framework is of course notoriously coarse-grained in its expressive capacity. By giving the possible worlds answer to the Boundary Question we are thereby saying, for example, that *no possible language* can have more than one logically true content, or can distinguish between Hesperus and Phosphorus contents. We rule out as *impossible* Fregean languages that have more than one concept of a given object or Russellian languages that distinguish among concepts based on their internal structure. Of course, fans of the possible worlds framework have things to say in response to these coarseness of grain considerations, and I don't mean them to be decisive. But I do think they are weightier when we consider whether possible worlds give us a suitable framework for *all possible languages* than when we simply consider how to analyze our own language.

The second and deeper concern is one of **Passing the Buck**. I'm assuming for now that our possible worlds framework is not a Lewisian "modal realist" framework (we'll touch on the Lewisian alternative below). As a result, we need some story about what possible worlds are. But all of the off-the-shelf stories seem to just raise the same Boundary Question problems again, simply slightly relocated. This is most obvious if we take possible worlds to be maximal modally compossible collections of propositions (but shows up also if we take possible worlds to be a partition of some basic/atomic/fundamental propositions, or to be maximal properties/states of affairs). Consider: is there a possible worlds content JONES IS A FREETHINKER? That depends on whether possible worlds themselves are characterized in part in FREETHINKER terms. We definitely want a possible world to settle who is an atheist and who is not. Do we also want a possible world to settle who is a freethinker and who is not? If we do, we'll find FREETHINKER propositions among our maximalist collection of possible worlds contents; if we don't, we won't. But the question of whether possible worlds settle who is a freethinker looks suspiciously like the question of whether there is a (real, non-defective) FREETHINKER concept or whether a total theory will settle who is a freethinker. Of course, paradox of analysis issues threaten here—these things had better be closely linked, given that we're trying to get an account of what the range of languages is that answers the Boundary Question and hence tells us what our ideal theory will discuss. But the territory has the definite odor of the non-explanatory regress and the unilluminating circle here. Similarly with other cases—do possible worlds settle TONK matters and SHERLOCK HOLMES matters? In picking out a possible world, do we need to specify whether (insert Eiffel Tower here)? In the end, I'm skeptical that we have any better grip on the notion of what a possible world is than on what a possible content is. Onward, then.

Second Semantic Attempt: Truth Conditions

The coarse-graining observed under **Undue Expressive Limitations** above suggests a natural next attempt: let's move to a more fine-grained semantic framework. There are many to choose from, but we'll consider the thought that a theory of meaning for a language takes the form of specifying truth conditions, so that contents can be equated with those truth conditions. This is an especially natural proposal in a setting in which we're tracking the prospects for a purely *alethic* criterion of theory selection. I'm not sure I know what truth conditions are (more on that momentarily), but many people seem to think that the truth conditions *true when 2 and 2 make 4* are different from the truth conditions *true when all groups of order 7 are abelian*, and even that the truth conditions *true when Hesperus is visible* are different from the truth conditions *true when Phosphorus is visible.*

Nevertheless, truth conditional semantics can still be plausibly accused of **Undue Expressive Limitations**. There are many off-the-shelf semantic machineries available to witness the potential expressive limitations. To pick a few:

- Many semantic frameworks use *truth-like* features, such as truth at a world, at a time, at a point of assessment, or at some other index of evaluation; or a Tarskian hierarchy of truth predicates, or separate ascending truth and descending truth.
- Many semantic frameworks *supplement* truth conditions with other content features, as in Potts-style two-dimensional accounts that have a dimension of expressive meaning.
- Some semantic frameworks eschew truth entirely, as in Gibbard/Blackburn-style expressivism.

I don't mean any of these cases for expressive limitation to be decisive. I'm not presupposing a methodology of ecumenicism, on which our account of the range of possible languages needs to accommodate everything any theorist has proposed as a language—it may well be that people have been writing down machinery for things that don't in fact count as languages. But I do think they are at least indicative that the notion of truth isn't enough to give us everything there could be in a language.

More importantly, truth conditional semantics are still subject to **Passing the Buck** worries. We can see this preliminarily by asking whether the following are specifications of truth conditions:

- "Jones is a freethinker" is true iff Jones is a freethinker.
- "Sherlock Holmes is a detective" is true iff Sherlock Holmes is a detective.
- "Trump is president tonk Pence is vice-president" is true iff Trump is president tonk Pence is vice-president.
- (Insert Eiffel Tower here) is true iff (insert Eiffel Tower here).

Hopefully not, for at least some of these, but I don't see anything better to say about why not than the prior observation that some of the putative concepts involved are defective and aren't part of any real language. More generally, the question is whether we have a picture of what kind of property truth is that puts helpful limits on what kinds of things can be truth conditions. (Note that it's not the range of truth *bearers* that is at issue here, but the range of truth conditions born.) Perhaps we can agree that it's in the nature of truth that truth conditions have to be given via (contents of) T-sentences of the form:

- (Truth bearer) is true iff (truth condition)

But that's not helpful if we don't know what the range of (contents of) T-sentences is. On the one hand, we already have reason to think that English may be overly generous on this account. *"Sherlock Holmes is a detective" is true iff Sherlock Holmes is a detective* is a grammatical English sentence of the form of a T-sentence, but that can't be *decisive* on the question of whether SHERLOCK HOLMES claims have real truth conditions. On the other hand, it's also plausible that English is overly restrictive on this account. If there are expressive limitations of English, of course, we should expect there to be truth conditions that can't be specified by English-language T-sentences. But it's also unclear what argument there is that the English syntactic category of sentence matches the range of permissible T-specifications (i.e., whether it's really distinctively T-*sentences* that we are out for).

Disquotational accounts of truth, for example, simply give no answer to the question of why (e.g.)

- "Aristotle" is true iff Aristotle.

Is not a valid instance of disquotation, beyond the simple insistence that the disquotational instances be sentences. (Is the reason that *"Aristotle" is true iff Aristotle* is uninterpretable (as opposed to ungrammatical))? I don't see why it would be, unless we have already decided that:

- "'Aristotle' is true iff Aristotle" is true iff "Aristotle" is true iff Aristotle.

is uninterpretable.

Third Semantic Attempt: Inferential Roles, Syntax, and Carnapian Tolerance

Let's look next at a rather different way of giving a semantic theory: via an inferential role semantics that associates each concept with governing inference principles. Given the Carnapian origins of the kind of conceptual maximalism that has spawned this investigation into the Boundary Question, perhaps we should expect this kind of syntacticized approach to language to bear fruit. Carnap, in his Principle of Tolerance, does seem to be treating it as a solution to the Boundary Question:

> In logic, there are no morals. Everyone is at liberty to build his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments.
>
> (Carnap 1934)

If we think of a language as a set of syntactic rules, then it looks like we have an easy answer to the question of what the possible range of languages is: we simply consider all possible syntactic rules. (Fussy point: we will need to take syntactic expressions to be individuated by their governing rules, so that we aren't at risk of overloading any one syntactic item. DWARF PLANET obeys a rule allowing transition from PLUTO IS A DWARF PLANET to PLUTO ORBITS A POINT OUTSIDE ITS BOUNDARY. PRIMARY PLANET obeys a rule allowing transition from JUPITER IS A PRIMARY PLANET to JUPITER ORBITS A POINT INSIDE ITS BOUNDARY. Those syntactic rule facts suffice to tell us that DWARF PLANET and PRIMARY PLANET are different syntactic items (and would be, even if both had the morphology "planet").)

There is a less tolerant and a more tolerant version of this Carnapian tolerance. On the less tolerant version, we require that a language be a collection of syntactic transformation principles that amount to rules of inference. On the less tolerant version, a "language" that consisted only of a single rule allowing "Jupiter" to be replaced by "Neptune" wouldn't count as a real language, because that rule wasn't allowing us to *genuinely infer* anything. Answering the Boundary Question in the less tolerant framework then requires a prior account of what makes a syntactic transformation rule a genuine rule of inference, and this looks to be just as buck-passing as options we've already considered. (There won't be any progress, e.g., in saying that a syntactic rule is a rule of inference if it is appropriately truth preserving.)

On the more tolerant version, a language can be just any collection of syntactic transformation rules. The worry now is a new one: that the resulting framework is *too* generous, and that no sensible theoretical project can be maximal with respect to it. One version of this worry, of course, is the familiar "tonk" worry. There is a perfectly well-formulated syntactic rule for "tonk", so if our theorizing framework needs to be maximal with respect to syntactic rules and contain every expression for which a rule can be given, it must contain "tonk". But if we theorize using "tonk", then our theory will contain *everything*, and a norm of theorizing that requires us to end up with a theory containing everything can't be giving us a good picture of the theoretical enterprise. (In fact, things are even worse than that. We could also have a term "tunk" which is governed by the rule that it is *impossible* to move from "A tunk B" to "B tunk A". If we try to include both "tonk" and "tunk", then it's not just that our endorsed theory is inconsistent, but that we have inconsistent verdicts on what we should endorse. But it's unclear why rules of forbidding should be less acceptable than rules of permitting.)

There are well-known avenues for dealing with the "tonk" problem, of course—variants of constraining tolerance to some variant on conservative syntactic rules. But these avenues aren't terribly well-suited for maximalist enterprises. It's not hard to craft a pair of expressions such that adding either one to a core language is a conservative extension, but adding both is a non-conservative extension. (From A we infer "A tank B", and from A we infer "A tink B". From "A tank B and A tink B" we infer B, but there is no independent elimination rule for either "tank" or "tink".) Then we're left with no suitable maximal language.

One point that comes out from consideration of these heavily syntacticized options is that the specifically *alethic* formulation of the non-pragmatic norm of theorizing may need to go. This point lurked already in the earlier observation that our language might contain resources whose contents are given in terms of truth. When the applicability of truth gives out before the boundaries of the language, we need some more general picture of what our theorizing enterprise is beholden to—some notion, perhaps, of what is to be said in the language.

7. Answering the Boundary Question with Metasemantics

There are, of course, many more semantic frameworks from which we could attempt to extract an answer to the Boundary Question. But I think the pattern of difficulties we saw in the two cases above is likely to continue. In both cases boundary concerns about language/concept/proposition just get transferred over to the central semantic coin of the specific semantic framework being considered (possible world, truth), and no progress is made. Simply putting more coin on the table seems unlikely to change that difficulty.

Maybe, then, we need to look past our semantic theories to the metasemantics. Perhaps understanding *why* our semantic frameworks are using the semantic coins that they are will give us a better understanding of what the boundaries of those coins need to be. Note that we can distinguish two types of metasemantics. There is *grounding metasemantics*, which answers the question 'In virtue of what do expressions have the meanings that they do?'. Grounding metasemantics isn't obviously helpful for answering the Boundary Question, since the Boundary Question isn't concerned with the expressing relation between syntax and semantics, but just with the *range* of that relation. (It's not impossible that understanding the grounding of the expressing relation will yield insight into its range, of course.) And there is *guiding metasemantics*, which answers the question 'What theoretical role is being fulfilled by semantic values being what they are?'. It's the guiding metasemantics that's potentially helpful in answering the Boundary Question.

We'll consider three kinds of guiding metasemantics:

- Metaphysical guiding metasemantics, holding that the theoretical role of propositions is to stand in a representation relation to privileged 'sentence-shaped chunks of the world' (Rorty 1986).
- Cognitive guiding metasemantics, holding that the theoretical role of propositions is to be the possible contents of thoughts.
- Practical guiding metasemantics, holding that the theoretical role of propositions is to characterize a possible move in a communicative exchange.

First Metasemantic Attempt: Facts and Other Metaphysics Heavy-Weights

Suppose our metaphysics hands us a domain of facts, and contents are then determined by the facts (e.g., contents are just possible facts, so that the truth conditions of contents are those facts. But the details won't matter). Or instead our metaphysics might hand us a domain of states of affairs, or of fundamental entities and features—anything to which the theory of content is then taken to be representationally responsible.

The basic dialectic should be predictable at this point. Looking to the metaphysics for an answer to the Boundary Question threatens to leave us subject to a dilemma between **Undue Expressive Limitations** and **Passing the Buck**. Roughly, if the metaphysics is not providing us with what are stipulatively well-suited to propositional expression, then there are expressive limitation worries as we try to fit the theory of content onto what the metaphysics does give us. And if the metaphysics does give us sentence-shaped chunks of the world, there is a concern that we've just passed the buck to the metaphysics, and don't know any better how to answer a metaphysical analog of the Boundary Question.

Suppose what the metaphysics gives us is Lewisian possible worlds. Then we can avoid buck-passing worries, because we're in a good(ish) position to say what kind of thing a Lewisian possible world is (a maximally connected spatiotemporal region and its contents), and we can just let the world answer for how many and how diverse such regions there are. But if the metaphysics gives us Lewisian possible worlds, the semantics we can build off of it is a possible worlds semantics, and we're back with the **Undue Expressive Limitations** worries we encountered earlier for possible worlds semantics.

Suppose instead that the metaphysics gives us facts. Facts look much better (although not unquestionably perfect) for grounding a suitably expressive theory of contents. But now the question *what possible contents are there* just gets shifted to the question *what possible facts are there?* Is there a fact that Jones is a freethinker? Is there a fact that Sherlock Holmes is a detective? More pressingly, does the combination of, say, Pluto and Neptune constitute a fact? Or do we need to say what the **combination** relation is to answer this question? If so, the need to select among candidate combination relations threatens to be as hard as the question of selecting among candidate boundaries of concepts.

Second Metasemantic Attempt: Concepts as Thinkables

If *what a language is* is a tool communication of our mental states, then the starting point for answering the Boundary Question is thinking about our mental lives. What a proposition is, fundamentally, is the kind of thing that can be the content of one of our beliefs (etc.). To answer the Boundary Question, then, we need to start by figuring out what kinds of beliefs it's possible to have.

But this, of course, threatens to be buck-passing again. As usual, the worry can be posed using any of our stock supply of "defective concepts". But let's cut more directly to the core. Suppose we're presented with a creature and told that the content of one of its beliefs is SOUTH AMERICA or BEING A FAN OF BORGES. Can we give any convincing explanation of why those aren't possible belief contents? We might say that the regulative norm of belief is truth, and that SOUTH AMERICA, lacking truth conditions, can't be subject to such a regulative norm and hence can't be a belief. But this looks like we're just passing the buck one more step back to the theory of truth. If we can't answer the Boundary Question for truth conditions, then we can't (on this approach) answer the Boundary Question for beliefs, and thus can't answer the Boundary Question for contents. Or we might say that beliefs are teleofunctionally given states whose role is to track the facts—but then we've passed the buck back to the metaphysics. Or we might be functionalists about beliefs—but then we've passed the buck onward to the theory of action, and it's hard to see why we get an improved grip on the Boundary Question there.

Third Metasemantic Attempt: Moves in a Language Game

One last attempt. Perhaps the lesson of all of this is that the pragmatist was on the right track all the time—we need to give up on the "purely theoretical" project that inevitably begins by carving out a notion of content on theoretical metasemantic grounds, and then uses that notion of content to set out a domain of theorizing. Instead, we need to get our grip on the very *tools* of theorizing in pragmatic terms. We need to think about what a language *is* by thinking first about what a language *does*. On this view, a proposition is a device for producing certain kinds of effects.

The crucial question will then be: what kind of effects? If we try to set out the kind too narrowly, we're back in the soup above. (Consider: if we say that a proposition is a device for producing the effect of getting the audience to believe something-or-other, in Gricean metasemantic style, we're back in cognitivist metasemantic territory, and unless we can say what the range of believables is, we've made no progress.) If we set out the kind too broadly, we don't effectively carve out a propositional, or even a linguistic, kind. (An utterance of "Aristotle" is going to do *something*, but that isn't enough to give that utterance propositional content.)

I think the best hope here is to appeal to the idea of "making a move in a language game". Consider the way that Dummett argues for a Fregean Context Principle in which sentential meanings have conceptual priority over subsentential meanings:

> If I take some coloured counters, and say, 'Let this one stand for the Government, this one for the Opposition, this one for the Church, this one for the Universities, this one for the Army, this one for the Trade Unions, . . . ', and so on, I shall be understood on the presumption that I am about to make some arrangement of the counters by means of which I intend to represent some relations between these institutions, and assert that they obtain. If I do not go on to make any such arrangement, but simply start talking about something else, my earlier declarations lose their original intelligibility: I cannot, when questioned why I said all that, reply, 'Oh, I just wanted those counters to stand for those things, that's all'; for their standing for those things only amounts to anything if they are to be used to effect some symbolic representation by means of which a thought is expressed. Otherwise, my stipulation of their reference is like my saying, in the course of explaining a card game, 'Ace is high', and it later turns out that the ranking of the cards plays no role in the game; or it is like my saying 'Suppose there is life on Mars', and then failing to draw any consequences from this hypothesis, and, when challenged, saying, 'Oh, I simply wanted you to suppose that'. (Dummett 1973)

But in the end, I don't find the suggestion helpful—I don't have any better grip on what it is to "make a move in a language game" than I do on the target notion of a proposition. I *do*, for example, find it intelligible to say "I just wanted those counters to stand for something". That in itself means that my notion of language game doesn't helpfully distinguish a 'claiming' move from a 'referring' move. That's enough to create worries for this pragmatic route to an answer to the Boundary Question, but it's also indicative of a deeper problem. In Dummett's relatively prosaic game, we can perhaps usefully distinguish claiming from referring moves, even if we regard both as legitimate moves. But as the games get more exotic, it's not at all clear that we have any helpful grip on what distinguishes an interesting *kind* of move.

8. Primitivism and the Star Gambit

Perhaps, of course, what all of this shows is that there's no *illuminating* answer to be given to the Boundary Question. Maybe all we can say is that there is a property of *being a language* or *being a proposition*, but we then have to take that property as primitive, and can't learn more about it by linking it to other properties (truth, fact, belief, action) in the way we've been attempting above. Or maybe it does link to some of these other properties, but in a way that produces a small and unilluminating circle.

It's hard to know what to say in response to "it's a primitive" moves. But I do think there is a specific and a general issue here each of which needs to be grappled with. The specific issue is that, as we've seen, there are particular questions about where the boundary lies that we'd like answers to. Does global language theorizing entail a theoretical commitment to catalog freethinker claims and tonk claims? The primitivist non-response to the Boundary Question leaves us at sea in answering those particular questions. Of course, no one promised we'd get all of our questions answered. Nevertheless, these particular questions do seem like ones where we might have plausibly hoped for some helpful guidance.

The general issue is this: suppose we take the notion of a proposition to be a primitive. Nevertheless, surely in the end there are variant notions that are in some sense *proximate* to the notion of a language/concept/proposition, in the familiar "quantifier variance" manner. So in addition to languages/concepts/propositions, there are languages*/concepts*/propositions*. Then two closely connected questions:

1. In virtue of what are we talking about languages, rather than about languages*?
2. Why should we take the aim of theorizing to be to collect claims from all languages, rather than from all languages*?

In response to the first question, we can hope that there is some kind of metasemantic story that gets us pinned down to LANGUAGE, rather than to LANGUAGE*. As with familiar worries about the metasemantic commitments of epistemicism, we might worry about whether there is going to be a metasemantics that's precise enough to target just one out of a dense cloud of related notions. But even if we think that there is a precise metasemantics (maybe it's naturalness to the rescue), it's unclear whether it really helps. After all, if it's naturalness that targets LANGUAGE, there is presumably also NATURALNESS* targeting LANGUAGE*. Naturalness* might not be the source of our metasemantics, but it might be the source of our metasemantics*. Our metasemantics* doesn't determine what language we speak, but it might determine what language* we speak*.

Similarly in response to the second question. Suppose we get some convincing answer to why we ought to theorize in languages, rather than in languages*. Does this help? Maybe all it shows is that we ought to theorize* in languages*. But then we can suggest that, although that's true, we don't *care* about theorizing*—theorizing* isn't one of our projects. The counter-response, of course, is that although theorizing* isn't one of our projects, and isn't normatively important, it is one of our projects*, and is normatively* important*. The crucial question, I think, is whether the star gambit is revealing a *problem*, or is simply allowing us to say at each stage that we care about what we care about. I don't know the answer to that question.

References

Carnap, Rudolf. 1934. *The Logical Syntax of Language.* London: K. Paul, Trench, Trubner & Co.

Dummett, Michael. 1973. *Frege: Philosophy of Language*. London: Duckworth.

Hirsch, Eli. 2002. Quantifier Variance and Realism. *Philosophical Issues* 12 (1): 51–73.

Lewis, David. 1983. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61 (4):343–77.

Putnam, Hilary. 1978. There Is at Least One A Priori Truth. *Erkenntnis* 13 (1):153–70.

Quine, W.V.O. 1948. On What There Is. *Review of Metaphysics* 2 (5):21–38.

Rorty, Richard. 1986. Pragmatism, Davidson, and Truth. In E. Lepore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (pp. 333–5). Cambridge: Blackwell.

Sider, Ted. 2012. *Writing the Book of the World.* Oxford: Oxford University Press.

9

Descriptive vs. Ameliorative Projects

The Role of Normative Considerations

E. Díaz León

1. Descriptive and Ameliorative Projects

Sally Haslanger (2000, 2006) has distinguished between *descriptive* projects and *ameliorative* projects in philosophy. The main idea is this: philosophers engaged in a descriptive project aim to reveal the *operative* concept, that is, the objective type that our usage of a certain term tracks (if any), whereas philosophers engaged in an ameliorative project aim to reveal the *target* concept, that is, the concept that we should be using, given our purposes and goals in that inquiry. The questions pertaining to ameliorative projects are the following: what is the point of having this concept? Which concept would serve these purposes best?

Haslanger (2000) offered a social constructivist analysis of gender and race, and argued that those analyses are not intended to capture our ordinary concepts of gender and race, but her aim was rather to figure out the target concepts, that is, the concepts of gender and race that would be most useful in order to achieve social justice. These analyses go as follows:[1]

> A group G is a *gender* (in context C) iff$_{\mathrm{df}}$ its members are similarly positioned as along some social dimension (economic, political, legal, social, etc.) (in C) and the members are "marked" as appropriately in this position by observed or imagined bodily features presumed to be evidence of reproductive capacities or function.
>
> A group G is *racialized* (in context C) iff$_{\mathrm{df}}$ its members are similarly positioned as along some social dimension (economic, political, legal, social, etc.) (in C), and the members are "marked" as appropriately in this position by observed or imagined bodily features presumed to be evidence of ancestral links to a certain geographical region. (Haslanger 2003: 8)

However, in more recent work (2005, 2006), she has argued that her social constructivist accounts of gender and race could also be seen as trying to capture the

[1] These formulations are a simplified version of her original, more complex characterization in her (2000) paper.

E. Díaz-León, *Descriptive vs. Ameliorative Projects: The Role of Normative Considerations* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020).

DOI: 10.1093/oso/9780198801856.003.0009

operative concept that we actually associate with our terms 'gender' and 'race'. As she argues, the operative concept could differ from the *manifest* concept (i.e., the concept that people take themselves to be using). Crucially, the characterizations of gender and race that would come to mind if ordinary speakers are asked 'What is race?' or 'What is gender?' might differ from the objective types that our uses of the terms 'gender' and 'race' actually track. In other words: even if the manifest concepts that many people associate with the terms 'gender' and 'race' are not explicitly social constructivist, these terms might refer to socially constructed properties after all.

This important distinction gives rise to the following question: should philosophers of gender and race engage in the descriptive project, or in the ameliorative project? It could be argued that these are two independent projects and that they are both useful. But if they are both useful, we can still ask: useful for what purposes, and under what conditions? In this chapter I would like to discuss this important question. In particular, I want to examine the nature and the prospects of both the descriptive and the ameliorative project regarding philosophy of gender and race. My main questions will be the following: what does the descriptive project consist in, and what steps are required in order to successfully complete it? What are the main criteria in order to evaluate different answers to a descriptive project? Furthermore, what does the ameliorative project consist in? What are the main criteria in order to evaluate different answers to an ameliorative project? Furthermore, are there any important connections between these two projects? If we have found a satisfactory answer to the descriptive project, does that pose constraints to our possible answers to the ameliorative project? And likewise, if we have found a satisfactory answer to the ameliorative project, does that pose constraints to our possible answers to the descriptive project?

Mari Mikkola (2015) argues that whereas philosophers of science have long recognized the role played by contextual factors in our inquiries (i.e., "the political and moral values embedded in the social context of an inquiry" (p. 782)), contemporary metaphysicians often assume that contextual factors do not belong in metaphysics. One of my main aims in this chapter is to argue that moral and political considerations among other contextual factors are relevant in metaphysics, and more in particular, I will show that they can be relevant at different stages of a metaphysical inquiry. I will start by discussing the nature of descriptive projects seeking to reveal the operative concept associated with a term, and I will argue that moral and political considerations are relevant at different stages of this project. In addition, I will distinguish two kinds of ameliorative projects (namely, cases where a term determinately refers to an entity but we ought to change the referent, and cases where a term's referent is indeterminate but it should become determinate, given normative considerations). I will argue that many projects that are usually taken to be purely descriptive are actually ameliorative or contain important ameliorative elements.

2. The Descriptive Project Revisited

Let's start with the descriptive project. Haslanger (2005, 2006) argues that the operative concept associated with gender or race might be a social constructivist concept after all. That is, the objective type that our concept of gender (or race) tracks might be a socially constructed property, like the ones she characterized above. What are the conditions for this proposal to be correct? That is, what would determine that our operative concepts of gender and race actually pick out socially constructed properties, and in particular the socially constructed properties that Haslanger (2000) characterizes?

Haslanger (2005, 2006) appeals to semantic externalism in order to defend this possibility. That is, she argues that our terms do not get their referents fixed in terms of the descriptions, conceptions, and beliefs that we associate with them, but rather in terms of externally individuated factors. Her strategy is to extend the familiar view of semantic externalism about natural kind terms to social kind terms as well. In particular, she characterizes externalism as follows:

> *Objective type externalism*: Terms/concepts pick out an objective type, whether or not we can state conditions for membership in the type, by virtue of the fact that their meaning is determined by ostension of paradigms (or other means of reference-fixing) together with an implicit extension to things of the same type as the paradigms. (2006: 110)

If we apply objective type externalism to gender and race (as objective types), it then follows that our concepts of gender and race get their referents fixed by means of *ostension*, that is, our terms 'gender' and 'race' refer to whatever objective types unify paradigms of gender (and likewise for race). The idea here, I take it, is that competent speakers (or at least the original speakers who introduced the term) would be able to point to paradigm cases of genders or races (e.g., gender groups, or racial groups). The referent will then be determined as follows: 'let "gender" refer to the most objective type that all those instances have in common', and likewise for the term 'race'. A first worry arises: how can we identify the most objective type that some entities share? As I understand Haslanger's view, the property that unifies all paradigmatic instances of a putative social kind will be that property that can better explain the paradigmatic features of the kind. That is, the most objective kind corresponds to the kind that satisfies some sort of explanatory role.

However, it could still be argued that this account leaves room for indeterminacy, since it is not clear that there is always a unique objective type that unifies paradigmatic instances of genders or races, since paradigms of genders have many properties in common, and paradigms of race have many properties in common, so it will be difficult to figure out which unifying property is the referent.[2] The crucial question, then, is the following: how can we identify the most objective property, out of the many properties that are shared by those paradigmatic instances? Appealing to their explanatory role might not be enough to identify a unique property. Or the situation might be even worse for the advocates of social constructionism: it could be argued

[2] As I am formulating the issue, the relevant terms are 'gender' and 'race', where paradigmatic instances that fall under the former would be properties or kinds such as *men* and *women*, and paradigmatic instances that fall under the latter would be racial groups such as *Black*, *White*, *Asian*, and so on. The issue could also be formulated concerning terms for specific genders or racial groups, such as 'woman' or 'Black'. In this case, paradigmatic instances falling under the first term would be individuals such as Hillary Clinton, whereas paradigmatic instances of the second would be individuals such as Barack Obama. The question then would be: which objective type, if any, unifies all paradigmatic instances falling under the term?

that the most objective, explanatory property that unifies our paradigmatic cases of gender (or race) is actually a biological property. Haslanger discusses this indeterminacy worry, and she says: "Sets of paradigms will typically fall within more than one type. To handle this, one may further specify the kind of type (type of liquid, type of artwork), or may (in the default?) count the common type with the highest degree of objectivity" (2006: 110). The suggestion here, as I understand it, is that our concepts of gender and race might involve some *sortal* information (i.e., information about the kind of thing something is) constraining the possible candidate referents, such as 'social kind that those instances have in common', or something along those lines. If this is indeed the sortal information associated with our concepts of gender and race, then the relevant objective type in the vicinity can be a social kind. However, this move seems to beg the question against the advocate of biological realism, who would claim that concepts of gender and/or race do not involve sortal information about their being social kinds. Perhaps the social constructionist could argue that the sortal information associated with terms such as 'gender' or 'race' is not explicitly social, but it does rule out biological properties as possible referents. My point here is that if the social constructivist wants to argue that gender and race concepts involve sortal information that rules out biological properties as referents, they would need to provide independent motivation for this.

Haslanger also suggests that in some cases there might be no sortal information associated with the concept, and then the referent will just be by default the most objective type shared by the paradigms, whatever that is. But then we face our original worry: there might be more than one objective type in the vicinity (all being equally explanatory), or even worse for the social constructivist, the most objective type might turn out to be a biological property after all. In my view, one way of avoiding this worry on behalf of social constructionism would be to allow that concepts of gender and race can be associated with some information (sortal or otherwise) that rules out biological kinds as possible referents, so that social kinds can be the referents after all.

In any case, in order to settle this dispute about what kind of information is associated with our concepts, one would need to engage in something very similar to what Haslanger (2006) calls the *conceptual* or *analytical* project. She characterizes this as the project of revealing our *manifest* concept, that is, the concept that we take ourselves to be using, or would come easily to mind if we are asked. However, I believe we should characterize the conceptual project in terms of the search for the *application conditions* of our concepts (which is a fallible inquiry, and which may require a lot of reflection about actual and possible cases, and even empirical research).[3]

[3] Haslanger (2005, 2006) seems to conflate these two different characterizations of the *manifest* concept. On one reading, the manifest concept corresponds to the application conditions of the concept. On this interpretation, the manifest concept corresponds to the conditions that something should satisfy in order to fall under the term. For example, the manifest concept associated with 'water' would correspond to something like 'the actual watery stuff', whereas the operative concept would correspond to H_2O, since this is the property that satisfies the manifest concept in the actual world. On another reading of 'manifest' concept, this corresponds to the concept we take ourselves to be using with a certain term, which may or may not correspond to the way the term is actually used in that linguistic community or practice.

But if we agree that the conceptual project is necessary in order to avoid indeterminacy, then we would have to deny one of Haslanger (2006)'s central claims, namely, that the descriptive project does not require the conceptual project. In my view, this only requires a friendly modification of Haslanger's approach, as follows. It could be argued that when it comes to the descriptive project, we need to engage in a *two-step* process: first, we need to find out the core information that competent speakers associate with the concept (where this might be accessible relatively *a priori*, although it might involve a lot of careful considerations of actual and possible cases, and even experimental semantics); and, second, we need to find out what objective kinds, if any, satisfy those descriptions in the actual world (where this can be found out only empirically).[4] However, Haslanger's main claim still applies: when it comes to figuring out the operative concepts of gender and race, that is, what objective properties our concepts of gender and race actually track, there is at least a central part of this inquiry that will be *a posteriori* or empirical, and social constructionism is still a live option.

In order to illustrate this two-step conception of the descriptive project, it will be useful to consider an example. Joshua Glasgow (2009) has argued that ordinary speakers associate the concept of race with the belief that racial groups are characterized in terms of certain visible traits that most members of a racial group have in common. He argues that ordinary speakers wouldn't be willing to give that belief up without replacing the ordinary concept of race with another concept. Glasgow (2009) then uses this claim about the application conditions of the concept of race in order to argue that the candidate meanings for 'race' that have been proposed by social constructivists and biological relists about race cannot actually be the referents of our ordinary concepts of race. For instance, he argues that reproductively isolated biological populations cannot be the referent of our ordinary concept of race, given that those biological populations do not satisfy some of the central features that are part of our ordinary concept of race (namely, that racial groups correspond to specific visible traits, whereas reproductively isolated populations do not need to have visible traits in common). Likewise, he argues, social structures of the sort advocated by Haslanger do not satisfy those central features of our ordinary concept either (because members of socially constructed groups do not have to have visible traits in common), and therefore social constructions cannot be the referent of 'race' either. This line of argument assumes that there is some information associated with an ordinary concept such that a candidate referent must satisfy it in order to be the referent. On the contrary, Haslanger (2005, 2006) explicitly denies this assumption (and suggests that the referent of 'race' is the most objective property by default). However, as I have suggested, I think that advocates of social constructionism about the operative concept of gender or race should not reject the first step of the descriptive project. In particular, this two-step approach can help the social constructivist to respond to Glasgow's objections to social constructivism, as follows.

For instance, Haslanger (2006) considers the example of a school that uses the term 'parent' in school memos but actually means 'primary caregiver' by it. Here I am focusing on the former characterization. See Díaz-León (2012) for further discussion of this distinction.

[4] Here I draw on Jackson (1998); Chalmers and Jackson (2001); and Thomasson (2007, 2008).

It could be argued that given the information associated with 'race' and the paradigm cases of racial groups that ordinary speakers would point to, there are no biological properties that unify all those cases, because, for instance, there are no biologically significant properties unifying all individuals that we would call 'white', and that the only properties that unify these paradigm cases are social properties. But in order to formulate this argument, we need to rely on the premise that there is some information associated with the term 'race' that rules out biological properties as being the referent. In addition, as I explained above, the conceptual project could also help with the worry that if we do not appeal to any associated information at all, then the referents of 'gender' and 'race' might be indeterminate, for it could be argued that the only feasible way to avoid this indeterminacy would be to appeal to some central information associated with our concepts that puts some constraints on what shared properties can be the referents. On the other hand, if we renounced to the first step of the two-step descriptive project, as Haslanger does, then the prospects of finding a good response to Glasgow's mismatch objections to social constructionism about race look very dim.[5]

Once this is clarified, we can see that the social constructivist could in principle argue that social structures might turn out to be the most objective property that unifies paradigm cases (assuming that they satisfy the information associated with the concept, if any). But how could we establish that a socially constructed property is actually the most objective, explanatory property that unifies paradigm instances of gender or race? As I said, we can appeal to some central descriptions associated with our concepts so as to rule out any candidate referents that do not satisfy some of those central, hard to give up descriptions. But this could still leave room for disagreement regarding which properties can better explain what unifies those paradigm cases.

One proposal could be to appeal to the kind of metaphysical project that Ted Sider (2011) has recently developed (drawing on Lewis 1983). In a nutshell, his main idea is that when it comes to figuring out the meaning of a certain relevant term, we should always choose the most joint-carving candidate (among those candidate meanings that satisfy the inferential role associated with the term, if any). This view assumes that there are some descriptions of reality that are more joint-carving than others, where a concept or description is more joint-carving that another when it expresses a property or entity that is more fundamental than that expressed by the other. For instance, describing objects in terms of their being green or blue is more joint-carving that describing them in terms of their being *grue*. In this way, it could be argued that if we are wondering which objective type corresponds to our use of a concept C (as the descriptive project aims to reveal), we should find out which is the most joint-carving concept, out of the several candidate meanings under consideration.

[5] The social constructionist has a possible alternative line of reasoning here: she might argue that there are no biologically significant properties in the vicinity of 'race' and 'gender' whatsoever, so that that the most objective properties in the vicinity can be social properties after all. This is a very controversial claim, and in my view social constructionists do not need to appeal to this possibility. As I have explained, another option would be to argue that our ordinary concepts of gender and race are associated with some central information that rules out any possible biological properties in the vicinity (if any) from being the referents.

For instance, we might wonder which is the most fundamental property, out of the several properties that are shared by paradigm instances falling under C. In my view, this proposal has some interesting commonalities with Haslanger's externalist account defended in her (2005, 2006). In the next section, I will develop this idea and I will explain how we can understand the notion of joint-carvingness in a way that can help us to make sense of the descriptive project of seeking the operative concept of gender and race. In section 5 I will argue that moral and political considerations can also be relevant with respect to the descriptive project, according to this framework.

3. Objectivity and Joint-Carvingness

Sider (2011) famously begins his book *Writing the Book of the World* by claiming that metaphysics is mainly concerned with the fundamental structure of the world. However, Elizabeth Barnes (2014) has rightly pointed out that disputes about the nature of gender and race are not disputes involving terms that are perfectly joint-carving. Indeed, many interesting philosophical disputes, such as debates about the nature of the mind, consciousness, concepts, meaning, species, composites, and artifacts, to name just a few, are arguably not about fundamental matters, and the corresponding terms are arguably not perfectly joint-carving. Therefore, if metaphysics is only concerned with the fundamental, we couldn't have proper metaphysical debates about gender and race (or many other non-fundamental phenomena). It seems plausible to conclude that we should revise our conception of metaphysics so as to allow debates about non-fundamental matters. However, the first paragraph of his book notwithstanding, Sider's characterization of metaphysics does not rule out the possibility of genuine metaphysical debates about non-fundamental matters. In this section, I will explain how his framework allows the possibility of debates about non-fundamental matters, and in particular I will argue that semantic externalists could understand the descriptive project of figuring out the referent of 'gender' and 'race' in terms of this framework.

So how can we have debates involving the notion of joint-carvingness, in order to talk about non-fundamental matters? Sider (2011), following Lewis (1983), appeals to the notion of terms that are not perfectly joint-carving but can still be *reasonably well* joint-carving, in the sense that they are more joint-carving than some alternatives.[6] That is, when it comes to a certain level of inquiry, say, chemistry, or psychology, or sociology, we can always ask which terms are the most joint-carving with respect to that level. For many levels, all the terms that are posited by theories at that level are going to be less than perfectly joint-carving, but we can still compare them according to how well they carve nature at the joints. In this way, as Sider explains, the notion of fundamentality or joint-carvingness that he appeals to is *absolute* (i.e., an entity is either perfectly fundamental or not, and a term is either perfectly joint-carving or not), but at the same time we can introduce a comparative notion of less-than-perfect joint-carvingness, where two terms can be less than perfectly joint-carving but one

[6] See Sider (2011: 77–8).

can be more joint-carving than the other (2011: 128–9). In what follows, I will say that a term that is not perfectly joint-carving but is more joint-carving than the (relevant) alternatives is *relatively* joint-carving. For example, terms that are relatively joint-carving with respect to sociology will arguably not be so with respect to chemistry, or physics. But we still can find out which terms are the most joint-carving at a certain non-fundamental level of inquiry, such as psychology or sociology. How can we figure this out? According to Sider (as I understand his view), this has to do with which terms or concepts are the most explanatorily useful, given the purposes and goals of explanations in that level. That is, to use Haslanger's terminology, when we are considering some terms in, say, sociology, and we are trying to figure out the operative concepts associated with them, a useful method here would be to figure out which kinds and properties (that are shared by the relevant paradigms) have the most explanatory power within sociology. In this way, we can make sense of the idea that in order to find out the operative concept, we need to find out which concepts are the most (relatively) joint-carving in that context, even if they are not perfectly joint-carving. So, of course, there are many debates in metaphysics that are not about perfectly joint-carving matters, but they can still be about relatively joint-carving matters.

One possible worry is the following: this framework appeals to different levels, such as physics, chemistry, biology, psychology, sociology, and so on, which are supposed to be ordered in a hierarchical way, according to how fundamental they are. And in addition, I am trying to make sense of the idea that there are some terms or concepts pertaining to a certain level that are more fundamental or more joint-carving than others with respect to that very same level. How can we make sense of this framework? It seems that we would need two different notions of fundamentality: on the one hand, we need a notion in order to account for the idea that physics is more fundamental than chemistry, which in turn is more fundamental than biology, and so on; and, on the other hand, we need a separate notion in order to make sense of the idea that we can compare different concepts posited at a certain level, according to how relatively fundamental or joint-carving they are, that is to say, according to how explanatorily useful they are. But it is not clear at all that we can make sense of this complex framework just in terms of the elusive notion of relatively fundamental or relatively joint-carving, in the sense of relative explanatory power I have suggested. For if we understand relative fundamentality in terms of different levels, then how can we compare the relative fundamentality of properties within the same level?

In my view, a tentative solution would be the following. We can use a version of the layered model of reality that Jaegwon Kim (2002) has explored in detail, following Oppenheim and Putnam (1958). According to this model, we can make sense of different layers or levels of reality in terms of the part-whole relation, that is, in terms of a mereological conception of reality. This model is primarily an account of different levels of reality in terms of different sums of entities. The model goes roughly as follows: at the most fundamental level n we have the most basic particles, which (we are assuming) cannot be decomposed any further, that is, they are not composed of further entities. At the following level, $n+1$, we have entities that are entirely decomposable into entities at level n. At level $n+2$, we have entities that are entirely decomposable into entities at level n and $n+1$, and so on. For instance, at a certain level m (where $m>2$) we will have molecules, which are decomposable into entities of lower levels. But the proper entities of level n and level $n+1$ are not decomposable into molecules. Likewise, at a certain level r (where $r>m$) we will have cells, which are decomposable into entities of the lower level m, but the proper entities of that lower level m are not decomposable into cells. This is what gives rise to the hierarchy of levels. As we have seen, this model makes use of the idea of a part-whole relation in order to account for the different levels of reality.[7]

Once we assume this mereological conception of levels, we can make sense of the idea of concepts that are more fundamental or more joint-carving than others at the same level of reality. First, we can talk about properties that appear only at certain higher levels, such as being a cell, or being a human being, or being conscious. We can then introduce new concepts in order to pick out these properties, and we can then talk about concepts that are posited at certain levels (when they refer to properties that appear in those levels). The crucial idea is that we can compare different concepts of a certain non-fundamental level (i.e., a level other than n) according to their explanatory power, that is, according to how explanatorily useful they are, with respect to the aims and purposes of our inquiries at that level.

Our next question is: what kinds of answers can we get when this method is applied? There are two main possible answers for a given descriptive project: either the term has a unique referent, that is, there is a candidate meaning that is the most objective or explanatorily useful, or the term is such that there are several candidate meanings that are equally joint-carving (with respect to that level). As Sider (2011) explains, candidate meanings are possible referents in the vicinity that satisfy the descriptions or inferential role associated with the concept, if any. If the paradigms have more than one property in common, where all of them satisfy the descriptions associated with the term (if any) and are equally objective or explanatorily useful, then we can say that there are several candidate meanings that are equally joint-carving. On the other hand, if there is a unique candidate meaning that is the most explanatory, then this gives rise in turn to two possibilities again: either this unique referent (the operative concept) corresponds to the referent the term should have (i.e., the target concept), or not. When the operative and the target concept come apart, this corresponds to the standard conception of the ameliorative project: a certain term actually refers to a certain property but given moral and political considerations, we ought to revise the meaning. In the next section I will argue that normative considerations are also relevant for the other outcomes, in addition to

[7] As Kim (2002) argues, this mereological model of levels does not correspond to the standard division of levels in terms of physics, chemistry, biology, and so on. For it is not clear that biology only posits entities of a certain mereological order. It seems that the contrary is true: biology can talk about micro-physical entities such as bacteria, about medium-sized objects such as human beings, and about larger entities such as entire populations, for example species. For this reason, the mereological model of levels and the standard hierarchy of levels of study do not perfectly match. I am hoping this is not a big problem for my account here. What is crucial is that we have a way of distinguishing between different levels of inquiry vs. different degrees of explanatory power. We could appeal to the standard division of levels, in a somewhat unprincipled way, or we could appeal to this mereological model of levels, which is a bit more motivated, but less standard. Either way, I hope this does not affect the usefulness of my approach here.

this standard conception of the ameliorative project. In particular, I will argue (i) that in order to figure out the operative concept associated with a term in the first place, we need to appeal to normative considerations; and (ii) that in cases where a term has several candidate meanings that are equally joint-carving, and therefore (on Sider's view) the referent is indeterminate, we should also appeal to normative considerations in order to check whether one of the candidate meanings ought to be the referent rather than the others. Therefore, normative considerations are not only the business of ameliorative projects as traditionally conceived, but also the business of descriptive projects.

4. Reference, Explanatory Power, and Normative Considerations

As we have seen, according to Haslanger (2005, 2006), it could turn out that the operative concepts of gender and race correspond to social constructivist analyses of gender and race (as characterized in section 1). This will depend on whether these socially constructed properties are the most objective types unifying paradigms of genders and races, that is to say, on whether those socially constructed properties are the most explanatorily useful. But how can we settle this question? In her (2000) paper, Haslanger argued that those social constructivist analyses satisfy some important explanatory roles. Charlotte Witt (2011) has also argued that social constructivist accounts of gender can provide some crucial explanations of interest to feminist theories. However, Katharine Jenkins (2016) has recently argued that Haslanger's account of gender fails to do justice to some important explanatory requirements, namely, to capture the gender identity of trans women who self-identify as women and therefore as being the same gender as cisgendered women. Jenkins argues, very convincingly in my view, that an account of gender that fails to do justice to this requirement cannot be satisfactory. In particular, she argues that we need two notions of gender, namely, a notion of gender as a social class (in terms of Haslangerian social structures) and a notion of gender identity in terms of self-identification, in order to have a satisfactory account of gender. But a worry arises here: it is not clear whether we can capture this important insight purely in terms of the externalist framework that Haslanger (2005, 2006) put forward. As I suggested above, a good way of making sense of the externalist framework is to argue that our terms fix their referent in an externalistically determined way, that is, by means of causal relations to the most explanatorily useful kinds in the vicinity, and so on. (This is basically a form of *reference magnetism*: our terms latch on to the most objective kinds in the vicinity, where these are understood as the ones that satisfy certain explanatory roles, enter in law-like generalizations, and so on.[8]) And it is not obvious that these considerations will always be sufficient in order to identify a unique property corresponding to our notion of gender or race.

As I mentioned above, there are two cases that I want to consider. First, it is at least conceivable that the explanatory, empirical reasons of the sort advocated by Sider's

[8] See Sider (2011: section 3.2), for further elaboration.

framework turn out to establish that our operative concept here corresponds to the social constructivist analysis proposed by Haslanger. This is, after all, the option suggested by Haslanger (2005, 2006), and as she already anticipated in her (2000), social constructivist analyses of gender do satisfy an important explanatory role. If that were the case, that is, if the objective type that our current usage of the concept 'gender' actually tracks corresponds to Haslangerian social structures, and if Jenkins (2016) is right when she argues that Haslanger's account of gender does not include all trans women under the corresponding group for 'woman', then our current usage of the term is arguably morally and politically objectionable. Many feminist theorists have argued (including Bettcher 2009; Saul 2012; Barnes 2014; Díaz-León 2016; and Jenkins 2016) that it is morally problematic to use a concept of 'woman' according to which there are some trans women that do not count as women (at least in most contexts).[9] Therefore, it could be argued that if the operative concept associated with 'gender' corresponds to Haslangerian social structures (or some other kinds that exclude trans women), then we have strong moral and political reasons to change the meaning of our term 'gender' so as to make it more inclusive. This is the business of the ameliorative project, or to use the label recently coined by Burgess and Plunkett (2013a,b), the business of *conceptual ethics*.

But I also want to argue that normative considerations of the sort that are relevant in ameliorative projects or conceptual ethics more in general could be relevant with respect to the other semantic possibilities too. Let's go back to the original question: we were wondering what the operative concept of gender is, and whether it corresponds to Haslangerian social structures, or Jenkins' notion of gender identity, or something else.[10] One possibility might be that when we appeal to the sort of explanatory reasons that Haslanger and Sider advocate, they just do not suffice to identify a unique kind as the most objective, explanatorily useful kind in the vicinity. For example, it might be that different kinds are useful for different purposes, and there is no clear way of choosing among those different purposes, given purely explanatory, empirical reasons. That is, in this case we have a term such that all the relevant candidate meanings are equally joint-carving. (As I said above, the relevant candidate meanings here are those properties that are shared by the paradigms and satisfy the descriptions or the inferential role associated with the term, if any.) What should we do in cases of indeterminacy of this sort? In my view, the best option is to appeal to the relevant moral and political reasons in the vicinity in order to decide which candidate meaning *should* be the meaning of that term (if any). This is part of the business of ameliorative projects or conceptual ethics, but it is different from the previous option in one crucial respect: instead of replacing the (determinate) meaning of a certain term with a different meaning, we are recommending that a

[9] Giving a full defence of this claim is outside the scope of this chapter, and I take it to be obvious anyway. But some of the reasons are the following: claiming that trans women are not women helps to promote and perpetuate stigma and discrimination against trans women, which results in great harm, exclusion, and even violence. A notion of gender that fails to take this into account is contributing to this harm and therefore is morally problematic.

[10] See Mikkola (2011), Sveinsdóttir (2011), and Witt (2011), for alternative accounts of gender; and Saul (2012), Bettcher (2013), and Díaz-León (2016) for alternative accounts of the meaning of 'woman'.

certain term with several candidate meanings (where it is indeterminate which candidate meaning it picks out) comes to have a unique determinate meaning, out of those alternatives.

We are now in a position to see that normative considerations of the sort that conceptual ethics emphasizes are relevant at the different stages of a metaphysical inquiry of the form 'What is X?', or 'Is X real?'. In order to build my case for this claim, I will explain again what the different stages of a metaphysical project of this sort amount to, and what kind of normative considerations can be relevant for each stage.

First of all, we have the *conceptual* project of finding out what central, hard to give up information ordinary speakers associate with the term 'X'. In my view, normative considerations such as pragmatic constraints, and even moral and political considerations, could be relevant at this earlier stage too. For example, in some cases we have concepts that are just associated to very thin descriptions such as 'whatever turns out to be the most objective kind shared by those paradigms'. In cases like this, it will make a big difference which paradigms we choose. It could be argued that in cases of some contested terms such as 'woman' or 'white', we have several choices about which paradigms we should focus on, and there are moral and political considerations that are relevant here. For instance, Bettcher (2013) argues that there are different communities that use the term 'woman' in different ways, and in particular she identifies a dominant conception of women that excludes some trans women (which is perhaps a more widespread usage), and a resistant conception that includes all trans women (which is perhaps only taken up by a minority of speakers). It could be argued that there are moral and political considerations that would recommend focusing on the community of speakers that endorse the resistant conception instead of the dominant conception. If we focus on the resistant conception, then arguably our class of paradigms can include trans women. As Bettcher puts it, "it is inappropriate to dismiss alternative ways in which those terms are actually used in trans subcultures; such usage needs to be taken into consideration as part of the analysis" (2013: 235). These considerations concern how to understand the very descriptions that we associate with the term (e.g., "the most objective property that is shared by *these* paradigms"), so they belong to the *conceptual* stage of figuring out the application conditions of the concept. We can then conclude that moral and political considerations can be relevant with respect to the first step of the descriptive project, that is to say, the project of finding out the application conditions of the term.

Second, there is the *empirical* stage of figuring out what objective type actually satisfies those application conditions that we have established in the first stage. In order to find out what is the most objective type in the vicinity (among those that satisfy the application conditions), we will need to figure out what is the most explanatorily useful kind shared by the paradigms. In my view, in order to be able to compare different properties in virtue of how explanatorily useful they are, we need to decide first what are the main goals and purposes of our inquiry at that level. Only then can we figure out which explanations are more useful, *with respect to the relevant goals and purposes in that context*. In the remainder of this section I will explain my main argument for this claim.

As we saw above, in order to make sense of disputes about non-fundamental matters we have to find out which candidate meanings are more joint-carving, with respect to a certain non-fundamental level. But explanations at different non-fundamental levels could have different aims and purposes. First, different levels have different *explananda*, for instance, biology aims to explain the behavior of biological populations whereas sociology aims to explain the behavior of social entities such as social groups or institutions. Actually, it could be argued that these different explananda might concern entities at the same mereological level, say, groups of human populations, and if so, the only way of distinguishing the explanations would be in terms of the purposes of the explanation, rather than the level (which is the same). For instance, we could argue that what counts as explanatorily useful with respect to explaining and predicting the behavior of reproductively isolated biological populations does not correspond to what counts as explanatorily useful with respect to explaining and predicting the behavior of social groups (even if these happen to be exactly the same groups). That is, we can imagine that we are focusing on a certain class of properties, and we ask: which property is the most fundamental one, out of these? I want to argue that we cannot answer this question independently of the aims and purposes of our inquiry. If we are focusing on the aims and purposes of biological explanations, one of those properties might turn out to be the most explanatorily useful one, whereas if we are focusing on the aims and purposes of sociology, a different property (of the same set) might turn out to be the most explanatorily useful one with respect to that inquiry. This gives us reasons to deny Sider's claim that "the world has a distinguished structure, a privileged description.... There is an objectively correct way to 'write the book of the world'" (2011: vii).

Once we understand joint-carvingness and explanatory power in this pragmatic way, what considerations can be relevant in order to find out what concepts are the most explanatorily useful, with respect to a certain inquiry? In my view, both theoretical and pragmatic factors can be relevant here. We have to figure out which properties are the most explanatorily useful with respect to that inquiry, and we can make sense of this question only when we clarify what the aims and purposes of our explanations in that inquiry are. As Philip Kitcher (2007) aptly puts it: "there were lots of different ways in which the world of living things can be divided up, according to the things human beings find salient and according to the purposes they have" (p. 300).

David Ludwig (2015) has argued that there are *non-epistemic* values that are relevant with respect to the truth-value of scientific statements such as (1): 'There are two different tiger species in the San Diego zoo'. Here I want to extend his argument, in order to argue that there are non-epistemic values that are relevant with respect to the descriptive project of finding the referent of terms such as 'gender' and 'race'. Ludwig's argument has two main premises: (i) the truth-value of many scientific statements depends on which ontological framework we choose; and (ii) choices regarding ontological framework depend in part on non-epistemic values. Therefore, it follows that the truth-value of many scientific statements depends on non-epistemic values. For instance, the truth-value of (1) above depends on which concept of species we choose, and this depends in turn on the explanatory interests of scientists.[11]

I want to extend Ludwig's line of argument in order to apply it to our account of descriptive projects in philosophy. On my view, questions about what is the referent of a term will depend, first, on the application conditions for the term that we choose, and, second, on what turns out to be the most explanatorily useful property in the vicinity, where whether a property is more explanatorily useful than another depends in part on the explanatory interests of the inquirers. In this way, we can argue that the answer to descriptive projects in philosophy of the form 'Is X real?' (or what is equivalent, as Thomasson (2008) argues, 'Does 'X' refer to anything in the actual world?'), and 'What is X?' (or what I am assuming to be equivalent, 'What is the referent of 'X'?), depends in part on non-epistemic values, including the explanatory interests that are relevant and appropriate for each project. Furthermore, we can argue that it is possible to assess the aptness of different explanatory goals according to different considerations, including theoretical, pragmatic, moral, and political factors (following Kitcher 2001). That is to say, moral and political factors can also be relevant at this stage, in two different ways: (a) in order to decide what are the most relevant aims and goals of our inquiry (i.e., some aims can be more politically useful than others); and (b) which explanations are the most useful with respect to some given goals (i.e., some explanations might be useful with respect to some criteria but not others, where these criteria are also morally and politically assessable). Therefore, we can conclude that normative considerations, including moral and political considerations, can be relevant with respect to a descriptive project in philosophy of the form 'What is X?' or 'Is X real?'. In particular, they are relevant in order to fix the criteria of explanatorily usefulness that we apply for comparing the different candidate meanings regarding how explanatorily useful they are.

In order to illustrate this line of argument, I will consider one example. We can consider the concept of *gender*, as the term was introduced in feminist theory in order to capture the sex/gender distinction. (See Mikkola 2008 for a historical survey.) Here I want to focus on the question: what is the operative concept associated with 'gender' in the context of feminist theory? I want to argue that in order to figure out the operative concept of 'gender', we need to make explicit the goals and purposes of explanations within feminist theory. Arguably, one of the main aims of feminist theory is to describe, explain, and resist the oppression of women. In my view, these aims can be used in order to compare different proposals about what is the most objective type that is shared by paradigms of 'gender'. This is one way in which normative considerations (including the sort of moral and political factors that motivate feminist theory) can be relevant in a descriptive project. We can compare this project with the different (descriptive) project of figuring out a notion of biological sex in biology. Arguably, one of the main aims of this inquiry is to explain sexual reproduction (although they might be others). In my view, the chosen goals will make a difference regarding what are the central explanatory aims and purposes

[11] Ludwig's excellent discussion draws on the arguments of many philosophers of science, including Dupré (1993); Anderson (1995); and Kitcher (2001).

with respect to which we should assess the explanatory usefulness of different candidate meanings for 'male' and 'female'. It could also be argued that the paradigms that we should pick up are different in biology than in the case of the term 'gender' within feminist theory, and this would suffice to yield different referents. But even if we focus on similar paradigms, the shared properties that turn out to be the most objective ones with respect to 'gender' and 'sex', that is, the most explanatorily useful ones, could turn out to be different in each case, since the main aims and goals of our explanations in each area are different, and therefore what counts as explanatorily useful can be different.[12] But, to emphasize, these projects are descriptive projects if anything is, and therefore my point is that normative considerations are not only relevant with regards to ameliorative projects about what 'gender' should mean, but also regarding descriptive projects about what 'gender' actually means (in the context of feminist theory, say).

To recap: As I said above, there are two possible outcomes of the descriptive project: either there is a unique candidate meaning that is the most joint-carving (and therefore this will turn out to be the referent, due to reference magnetism), or there are several candidates that are equally explanatory, given our theoretical aims and constraints, and then it will be indeterminate what the referent is. In this second case, if there are any normative considerations that might be relevant here, including prudential, moral, and political reasons, then this would give us good reasons to choose one referent over the others, out of those equally objective candidate meanings, and in my view this means that we ought to revise the meaning so that the term gets to have that unique referent, instead of an indeterminate meaning.[13]

Then, we should reformulate the nature of descriptive projects seeking the operative concept as follows. The corresponding overall question will be the following: what is the most objective type (that satisfies the application conditions), that is, what is the most explanatorily useful kind, given all the relevant considerations, including theoretical, instrumental, moral, and political considerations? More generally, the aim here is to find out what is the most explanatorily useful kind in the vicinity, given what the relevant aims and purposes of our inquiry are, what the relevant paradigms are, and so on. And if it turns out that there are several candidate meanings satisfying

[12] Haslanger (2016) makes a similar case regarding the notion of sex. One of my aims here is to show that this line of argument can be generalized in order to provide a characterization of descriptive projects in philosophy that allows room for normative considerations, in a way that is compatible both with a radical externalist framework like Haslanger's, and with a two-step descriptive project of the sort I have proposed.

[13] There are at least two ways in which we could understand this indeterminacy. It could be the case that the referent of the term is genuinely indeterminate, in the sense in which vague terms such as 'bald' can have indeterminate referents. It could be argued that there are some possible precisifications of 'bald' that are more politically useful than others, and therefore this would give us some reasons for changing the meaning of 'bald' so that it comes to determinately refer to this precisification. Alternatively, we could say that the term is context-sensitive, such as 'tall'. Arguably, 'tall' means something like 'being taller than a certain threshold that is salient in this context'. In this case, it is not the case that the term has several candidate meanings such that it is indeterminate what the referent is at any given context. Rather, the term has a unique referent at each context where a threshold is made salient, but this threshold can change from context to context, and therefore the term can have different referents at different contexts. In Díaz-León (2016), I argued that moral and political considerations are relevant in order to determine which standards of similarity are more salient in each context, with respect to context-shifting terms.

the descriptions or information associated with the term that are all equally joint-carving (given the theoretical and practical criteria governing our inquiry), then we should appeal to any additional relevant moral and political considerations to check whether any of the candidate meanings can better satisfy these normative criteria.

Finally, it might be the case that even if there is a clear objective kind that is the most explanatorily useful corresponding to some term, according to the descriptive project, there might be additional moral and political considerations that trump this, so that it is the case that we *ought* to change the meaning of that term, all things considered. This is the most familiar version of an ameliorative project. But as I have argued, this is not the only stage where normative considerations are relevant. As we have seen, normative considerations are also relevant in order to find out what the actual referent of a certain term is, or what the referent should be, given a situation of indeterminacy.

5. Conclusion

We can conclude that, at the end of the day, there is no sharp distinction between debates that are properly descriptive, and debates that are ameliorative, since normative considerations are relevant at many different stages of both projects. In my view, it is more useful to see the distinction between descriptive and ameliorative projects as different stages of an overarching project, where the relevant overall question is the following: what is the most useful way of using 'X', or what should 'X' mean, given all relevant theoretical, practical, moral, and political considerations?

Acknowledgements

Research on this project has been funded by grants RYC-2012–10900 and FFI2015-66372-P (Spanish Government) and H2020-MSCA-ITN-2015–675415 (European Union). I have presented ancestors of this material at the Universities of Barcelona, Essen, Leeds, Leuven, Oslo, Sheffield and St Andrews, and the Eastern APA in Baltimore. I am very grateful to the audiences on all those occasions for very useful feedback. Extra thanks are due to Brice Bantegnie, Delia Belleri, Ingo Brigandt, Alexis Burgess, Herman Cappelen, Aurélien Darbellay, Sally Haslanger, Dan López de Sa, David Ludwig, David Plunkett, Kevin Scharp, Tim Sundell, and Amie Thomasson for very helpful comments.

References

Anderson, E. 1995. Knowledge, Human Interests, and Objectivity in Feminist Epistemology. *Philosophical Topics* 23 (2):27–58.

Barnes, E. 2014. Going Beyond the Fundamental: Feminism in Contemporary Metaphysics. *Proceedings of the Aristotelian Society* 104 (3):335–51.

Bettcher, T. 2009. Trans Identities and First-Person Authority. In L. Shrage (ed.), *You've Changed: Sex Reassignment and Personal Identity* (pp. 98–120). Oxford: Oxford University Press.

Bettcher, T. 2013. Trans Women and the Meaning of "Woman". In N. Power, R. Halwani, and A. Soble (eds.), *The Philosophy of Sex* (pp. 233–49). London: Rowman & Littlefield.

Burgess, A., and Plunkett, D. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, A., and Plunkett, D. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Chalmers, D., and Jackson, F. 2001. Conceptual Analysis and Reductive Explanation. *The Philosophical Review* 110 (3):315–61.

Díaz-León, E. 2012. Social Kinds, Conceptual Analysis, and the Operative Concept: A Reply to Haslanger. *Humana.Mente: Journal of Philosophical Studies* 22:57–74.

Díaz-León, E. 2016. *Woman* as a Politically Significant Term: A Solution to the Puzzle. *Hypatia* 31 (2):245–58.

Dupré, J. 1993. *The Disorder of Things*. Oxford: Oxford University Press.

Glasgow, J. 2009. *A Theory of Race*. London: Routledge.

Haslanger, S. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Nous* 34 (1):31–55.

Haslanger, S. 2003. Future Genders? Future Races? *Philosophic Exchange* 34 (1):4–27.

Haslanger, S. 2005. What Are We Talking About? The Semantics and Politics of Social Kinds. *Hypatia* 20 (4):10–26.

Haslanger, S. 2006. What Good Are Our Intuitions? Philosophical Analysis and Social Kinds. *Proceedings of the Aristotelian Society*, Sup. Vol. 80 (1):89–118.

Haslanger, S. 2016. Theorizing with a Purpose: The Many Kinds of Sex. In C. Kendig (ed.), *Natural Kinds and Classification in Scientific Practice*. London: Routledge.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jenkins, K. 2016. Amelioration and Inclusion: Gender Identity and the Concept of *Woman*. *Ethics* 126 (2):394–421.

Kim, J. 2002. The Layered Model: Metaphysical Considerations. *Philosophical Explorations* 5 (1):2–20.

Kitcher, P. 2001. *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Kitcher, P. 2007. Does 'Race' Have a Future? *Philosophy & Public Affairs* 35 (4):293–317.

Lewis, D. 1983. New Work For a Theory of Universals. *Australasian Journal of Philosophy* 61:343–77.

Ludwig, D. 2015. Ontological Choices and Value-Free Ideal. *Erkenntnis*. DOI: 10.1007/s10670-015-9793-3.

Mikkola, M. 2008. Feminist Perspectives on Sex and Gender. In E.N. Zalta (ed.), *Stanford Encyclopaedia of Philosophy* (summer edn). http://plato.stanford.edu/archives/sum2008/entries/feminism-gender/.

Mikkola, M. 2011. Ontological Commitments, Sex and Gender. In C. Witt (ed.), *Feminist Metaphysics* (pp. 67–83). New York: Springer.

Mikkola, M. 2015. Doing Ontology and Doing Justice: What Feminist Philosophy Can Teach Us About Meta-Metaphysics. *Inquiry* 58:780–805.

Oppenheim, P., and Putnam, H. 1958. Unity of Science as a Working Hypothesis. *Minnesota Studies in the Philosophy of Science*, vol. 2. Minneapolis: University of Minnesota Press.

Saul, J. 2012. Politically Significant Terms and Philosophy of Language: Methodological Issues. In S. Crasnow and A. Superson (eds.), *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy* (pp. 195–216). Oxford: Oxford University Press.

Sider, T. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Sveinsdóttir, A. K. 2011. The Metaphysics of Sex and Gender. In C. Witt (ed.), *Feminist Metaphysics* (pp. 47–65). New York: Springer.

Thomasson, A. 2007. *Ordinary Objects*. Oxford: Oxford University Press.

Thomasson, A. 2008. Existence Questions. *Philosophical Studies* 141:63–78.

Witt, C. 2011. *The Metaphysics of Gender*. Oxford: Oxford University Press.

10

Variance Theses in Ontology and Metaethics

Matti Eklund

Conceptual engineering concerns questions about what concepts we should employ, for various purposes. I would like to place conceptual engineering in a more general theoretical setting. Much of analytic philosophy has been concerned with analysis of concepts we do have, and with investigation into the properties and relations they ascribe. But these concepts, and these properties and relations, are of course just some among all the concepts, properties, and relations there are. There are our ordinary concepts TRUTH, EXISTENCE, KNOWLEDGE, and FREEDOM, and the properties and relations they ascribe. But there are also other possible concepts, including other possible concepts that, while different from our actual concepts, are like them in certain respects—and there are the properties and relations they ascribe. One theoretical project is that of mapping out what kinds of possible concepts there are, and what properties and relations these concepts ascribe. A related project is that of comparing these concepts, properties, and relations along different dimensions of evaluation. These projects are arguably of greater philosophical significance than the one of getting clear on ordinary concepts and the properties and relations they ascribe. What is so philosophically significant about the concepts we happen to have, and about the properties and relations we happen to have concepts of? Engagement in the two projects just described can in turn issue in practical recommendations, and then we are doing conceptual engineering. But conceptual engineering is just one special case of the broader project of investigating what possible concepts there are, and the properties and relations they ascribe.

In this chapter I will discuss some aspects of this broader project. I will begin by discussing a relatively worked example: the so-called quantifier variance thesis discussed in metaontology, according to which there are different existence concepts and none is privileged. I will then make the (obvious) point that the quantifier variance thesis is just a special case of a more general kind of claim. A variance thesis is, generally stated, a thesis to the effect that there is a multitude of different concepts of some particular kind and none of them is privileged. (Naturally, there is much to unpack here.) One can put forward variance theses in other areas. One other area where one can put forward this kind of thesis is in metaethics. I have investigated this case at length elsewhere, and will here primarily focus on two aspects of it.

Matti Eklund, *Variance Theses in Ontology and Metaethics* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Matti Eklund.
DOI: 10.1093/oso/9780198801856.003.0010

I will discuss interesting similarities between the metaontology case and the metaethics case, thus attempting to illustrate the benefits of taking a general approach to the issue of variance theses. Then, lastly, I will discuss what general theoretical obstacles there may be to evaluating variance theses: the metaethics case illustrates some such obstacles, and a natural question to ask is to what extent general lessons can be drawn.

1. Ontology

Whereas plain ontology is concerned with what exists, metaontology concerns the nature of ontological questions. Much metaontological discussion concerns the status of ontology as an enterprise: is ontology an enterprise in good standing, or is it somehow misbegotten? Much of the current interest in metaontology is due to Eli Hirsch's writings.[1] Hirsch's writings have introduced two different (but not always carefully distinguished) kinds of theses into the literature:

Quantifier variance: there are different existence concepts, and none of them is privileged over all others.

Verbalism about ontology: (many) ontological disputes are purely verbal, due to the disputants meaning different things by 'there exists' and cognates.[2]

The theses are different. Even if there are different existence concepts and none is privileged, it can be that the disputants in ontological disputes tend to use 'there exists' with the same meaning. And it can be that while would-be disputants in ontological disputes use 'there exists' with different meanings, one existence concept is privileged over other existence concepts.

These two theses have then been used for criticism of the enterprise of ontology. In the case of quantifier variance the idea is: if quantifier variance is true, what is so interesting about questions about what exists (in the ordinary sense of 'exists')—given as there are other, equally good existence concepts? In the case of verbalism the idea is: if verbalism is true, then the disputes ontologists engage in are merely verbal; the disputants simply talk past each other.

Let me make a few remarks on the less straightforward of these theses, the quantifier variance thesis. As stated, the thesis of quantifier variance immediately invites the question: what is it for something to be an existence concept? This is a hard question, and one that actually has not been much discussed in the literature, but an answer commonly gestured toward is: having the right sort of *inferential role*; more specifically, being governed by the same inference rules as the ordinary existential quantifier. The thesis, famously associated with Quine, that so-called existential quantification expresses existence is presupposed as background. Hence the name 'quantifier variance' for the thesis. (I will keep dropping the 'existential' and

[1] See, for example, the essays collected in Hirsch (2011).

[2] Note the cautious 'introduced into the literature'. It is doubtful that Hirsch has ever subscribed to quantifier variance as I go on to explain it. And although Hirsch does defend a form of verbalism, the exact verbalist thesis he defends is carefully circumscribed. For some discussion of these matters, see Eklund (2011).

speak of quantifiers and quantifier meanings even though it is specifically existential quantification we will be concerned with.) If instead one thought of so-called existential quantification as not having ontological import and of existence as being expressed by a predicate, the relevant variance thesis for ontology would be that there are different existence predicates, and none of these predicates is privileged. I have formulated the quantifier variance thesis in terms of concepts, but the thesis could equally well be stated in terms of possible existence meanings, and occasionally I will talk that way.

It may be worth comparing a variance thesis which doesn't give rise to this kind of questions:

> *Liberalized variance*: there are different languages within which to state our overall theories of the world, not all employ the ordinary concept EXISTENCE, and no such language is privileged over all others.

Questions about what it is for something to be an existence concept are irrelevant to the liberalized variance thesis. Still the liberalized variance thesis promises to pack the same metaphilosophical punch as the original quantifier variance thesis. If liberalized variance is true, then what is so significant about questions about what strictly *exists* (in the ordinary sense of 'exists')? There are other concepts we equally well could have used to state our overall theory of the world.

The variance theses—both the quantifier variance thesis and the liberalized version—speak of concepts and languages, devices for representing the world. Some may wish to protest that what they are concerned with is *existence*, not the concept EXISTENCE, or the word 'exists', and hence variance theses are irrelevant. But such a protest is beside the point. Even if what we as a matter of fact are concerned with existence and not its representations, one can ask, for example: why focus on existence—that thing which as it happens is ascribed by our concept EXISTENCE—and not what is ascribed by some alternative existence concept?

The quantifier variance thesis speaks of different existence concepts. This invites questions about how concepts are individuated. But the thesis gets its bite from the associated claim that some purported entities may 'exist' in one sense of 'exist' but not another: that the different concepts can differ in extension. So as far as the talk of concepts in a variance thesis are concerned, we can think of concepts as being different exactly when they have different extensions.

Another question regarding quantifier variance concerns what 'privileged' comes to. In some way to be privileged is to be better than the competition. But there are lots of different dimensions along which to evaluate concepts. One concept may be better relative to one aim, another better relative to another aim. One may be more practical to use, another may be more explanatory, a third may in principle have certain aesthetic qualities that the others lack, and so on. A common view in the metaontology literature is that the relevant dimension of evaluation is something like joint-carvingness, naturalness, fundamentality, . . .[3] While there are differences between these notions I will treat them under the same heading: *metaphysical eliteness*, or

[3] See, for example, Lewis (1983, 1984); Schaffer (2009); and Sider (2011).

eliteness for short. A common idea is that what is elite is what makes for objective similarity, and what is fundamentally explanatory. There is much to say about this notion of eliteness (as well as objections to address regarding this notion). But for present purposes I will treat the notion as perfectly in order, and none of what I will go on to say will depend on finer details regarding what eliteness comes to.

Saying that the relevant dimension of evaluation for the purposes of the variance theses in metaontology is eliteness is not to say that eliteness is the only possible relevant dimension of evaluation for existence concepts. It could be, for example, that the most elite existence concept is impracticable to use for creatures with minds like ours, and then for various practical purposes other existence concepts will be preferable. A different dimension along which existence concepts could be evaluated is then: suitability for these particular practical purposes. In principle, one could consider different quantifier variance theses corresponding to different dimensions along which concepts may be evaluated. But a working assumption has been that to focus on eliteness is to focus on something that is directly germane to standard concerns of ontologists.

A variance thesis seems to me to be of deeper significance than verbalism, as far as criticism of the enterprise of ontology is concerned. Here is one argument for why verbalism is not of principled importance regarding the enterprise of ontology.[4] It is compatible with verbalism that there is a privileged existence concept. But then suppose, for example, that ontological disputes tend to be purely verbal, as verbalism says, but there is such a privileged existence concept. Then ontology could simply be recast as an enterprise of asking what exists in the sense of *that* existence concept. It could be insisted in response that even if there is a unique, privileged existence concept, it is simply impossible to have a non-verbal dispute over what exists in the sense of that. The idea would be that even if there is a privileged existence concept, when two theorists have a philosophical dispute over, as they would put it, ‘Fs exist’, it is simply *inevitable* that they do not use ‘exist’ to express the same existence concept. I take this to be clearly absurd. (Note that to say that this is absurd is not to deny that many ontological disputes as they actually are conducted are merely verbal.)

That verbalism lacks principled significance regarding the enterprise of ontology does not mean that a variance thesis does better in this regard. Why cannot the practicing ontologist respond to the thesis of quantifier variance by saying: I don’t care if existence (in the ordinary sense) is not privileged—I will continue focusing on it as before anyway? If someone consistently holds on to this stance, there may be no arguments that could rationally sway her. To each her own. The supposed point of quantifier variance rather comes in through the assumption that the significance that ontologists attach to their enterprise is due to their thinking of questions of existence as deep, in a way properly spelled out in terms of existence being privileged.[5]

[4] For a different but related argument to the same effect, see Eklund (2016).

[5] Sider (2011: 62) holds that it is a correctness condition on belief that it track the world’s “structure” (what is elite). Given Sider’s view, to not care about getting structure right is akin to not caring about getting at the truth.

A variance thesis compares existence concepts with each other and concerns whether one existence concept is privileged over all others. One may think a better question to ask concerns not about how privileged some existence concept is compared to alternative existence concepts, but how privileged it is full stop. So long as it is privileged enough it, or what it ascribes, is a worthy object of investigation. While the question has not been the focus in metaontology it is worth keeping it in mind, also for when, later, parallel issues in metaethics are considered.[6]

Largely in response to what he sees as the threat posed by the thesis of quantifier variance, Ted Sider has developed and defended what he calls *ontological realism*: the thesis that there is a privileged existential quantifier meaning.[7] This is in opposition to the quantifier variance claim that no existence concept is privileged. It is worth stressing that Sider's ontological realism is not simply the thesis that ontological sentences—sentences about what exists—have objective truth-values, and that some atomic ontological sentences are true. That much is fully compatible with quantifier variance. What sets Sider apart from the friend of quantifier variance is Sider's claim that one existential quantifier meaning is *privileged*. Given that eliteness is the relevant dimension of evaluation, the claim is that one meaning is more elite than the others.

It has become more and more common in recent years to hold that it is not sufficient for "realism" about discourse D that D-statements are capable of objective truth and some D-statements are true—realism requires something more *beefed-up*. Apart from Sider on ontological realism one might compare Kit Fine on realism, Crispin Wright on marks of realism, and the discussion in metaethics of creeping minimalism. Kit Fine (2001) operates with a primitive notion of what is real, such that it can be that F without it being the case that in reality, F. Statements of a discourse can then be true without them holding in reality. Crispin Wright (1992) distinguishes between different marks of realism. Many of the marks are held to gauge objectivity of a discourse. But one of them, what Wright calls wide cosmological role, has, in brief, to do with the explanatory power of the truth of a class of statements. If statements of discourse D are true but lack wide cosmological role, then discourse D is not fully realist. The problem of creeping minimalism in metaethics (see James Dreier 2004) has to do with the fact that even the non-cognitivist can, as it is often put, earn the right to speak of normative statements as true and mind-independently so: and this raises the issue of what the would-be realist can say to distinguish herself from this non-cognitivist. In each of these cases, there are pressures to say that realism demands something more than mind-independent truth. In this way, these views are like Sider's ontological realism. But one central thing that sets Sider's ontological realism apart is that the fate of Sider's ontological realism is explicitly not bound up with how *actual* ontological discourse works. It is no part of

[6] Another complication regarding quantifier variance is the following. How threatening quantifier variance is to the enterprise of ontology may depend on details not captured by the above statement of the thesis. Quantifier variance as stated could be true for the reason that there are exactly two existence concepts that are maximally, and equally, elite (and they are, moreover, nearly coextensive); or it can be true for the reason that there is a wide variety of significantly different best existence concepts.

[7] See Sider (2009) and (2011: chapter 9).

Sider's ontological realism that the *actual* existential quantifier has the privileged meaning: the extra demand that Sider imposes is only that some possible quantifier has a privileged meaning. Whether or not one accepts Sider's view on what makes for privilege, something seems right about the focus on possible instead of actual discourse. It could, for example, be that non-cognitivism, or an error theory according to which no atomic statements of the discourse are true, is correct regarding actual normative discourse but one could engage in some possible normative discourse which is cognitivist and where some atomic statements are true—and, generally, satisfies any demands a realist may wish to impose. This could be sufficient for the realist's demands. The important thing for the realist about the normative is that there are genuine normative aspects of reality; not that our actual languages or conceptual scheme contains the means to pick them out.

Someone like Fine or Wright could certainly take on board this aspect of Sider's view. For example, Fine could say that what matters for normative realism is not whether actual normative sentences express truths which hold in reality, but instead whether some normative truths hold in reality. A would-be normative realist concerned with creeping minimalism could be more concerned that some possible normative discourse has features by virtue of which it cannot be understood as non-cognitivist than that actual normative discourse does so. What I wish to emphasize is just that the fact that the point that the details about actual language do not matter is something that is stressed by Sider but is not a theme in other discussions of beefed-up realism.[8]

2. Variance

Hirsch focuses on ontology. But the same themes can rather obviously crop up elsewhere:

X-variance: there are different X concepts, and none of them is privileged.
Verbalism about X: (many) X disputes are purely verbal.

For the same reasons as given earlier, I believe that a variance thesis is of greater potential significance than a verbalism thesis. Let me then focus on variance theses.

Many philosophical debates—over knowledge, free will, meaning, . . . —concern (or, as I will turn to shortly, are conducted *as if* they concern), our actual concepts, or, better, the properties and relations they stand for.

The subject matter, say knowledge, is taken as given and theorists concerned to, by for example eliciting judgments about cases, figure out the nature of knowledge. It is not asked whether there may be some other epistemic relation—call it knowledge*—such that knowledge* is of greater epistemic significance than knowledge. (To take a tired example: it is as if natural scientists took the notion of weight for granted and sought to illuminate what weight is, without considering whether there are other

[8] Sider was of course not the first to defend a realist thesis while formulating his preferred view not in terms of ordinary notions but in terms of reformed counterparts. For example, in metaethics, Railton (1986) does the same, relating to Brandt's (1979) earlier talk of reforming definitions. What is new in Sider is the explicit focus on this way of conceiving of realist theses.

notions, like various notions of mass, which are capable of doing a better job as far as physical theory is concerned.)

In each case one can wonder whether our actual concept is the best concept in the relevant class. Take again knowledge. The post-Gettier literature has seen many different suggested analyses of the concept KNOWLEDGE. Whether or not these analyses succeed as such, they characterize various possible knowledge concepts. It is even possible to hold that even if, as per Gettier, knowledge is not justified true belief, the best knowledge concept is a justified true belief concept.

In response to what I have just urged, it may be objected (and here the *as if* from above comes in) that philosophers who are concerned with, say, knowledge are not in fact concerned with the ordinary folk concept and what it ascribes. If one considers how epistemological inquiry is actually conducted, one will find that philosophers are already concerned with improving concepts, and are using a somewhat technical knowledge concept, one perceived as meeting theoretical needs better than the folk concept KNOWLEDGE does. I actually think there is a lot to this objection. But even if what the objection alleges regarding the current state of philosophy is correct, two points deserve stressing. One is that the substantive point remains, regarding the justification for not merely focusing on our actual concepts and what they ascribe—it is just that the advice given may already be followed. Second, even if in fact philosophers tend to use somewhat technical concepts, the strategy of doing so has not been always followed in a self-conscious way, and it has not often been explicitly considered.

I believe the above points about how we ought to consider alternatives to our concepts are all rather intuitive. But there are obviously questions to be raised, analogous to questions raised regarding quantifier variance.

What, in general, makes something an X concept? When introducing quantifier variance, I made some remarks about what counts as an existence concept. But the issue now arises in a more general setting. If one understands 'X concept' as 'concept of X', one can take this to be a concept that ascribes or refers to X. But that idea is a non-starter: for we wish to be able to regard non-coextensive concepts as X concepts. For example, the actual concept KNOWLEDGE and the concept JUSTIFIED TRUE BELIEF can both be knowledge concepts even if some cases of justified true belief are not cases of knowledge. What one might wish to say is something of the form: an X concept, in the relevant sense, is a concept that could play the X role. The thought would be that even though our concept KNOWLEDGE is not coextensive with the concept JUSTIFIED TRUE BELIEF, the latter concept of justified true belief could play the knowledge role—it could be used for epistemic evaluation in the way the actual concept KNOWLEDGE actually is. In the discussion of existence above, it was said that what makes something an existence concept is its having the right inferential role. This can be made to fit the present mold: having the right inferential role suits a concept to do what the actual existential quantifier does for us. Note that matters here are delicate; not to say slippery. The specific claim in the case of existence concepts, that being governed by some specific inference rules is what makes something an existence concept, arguably does not generalize. It is not plausible that sharing of such structural, inferential features is necessary and sufficient for being a knowledge concept.

I am talking about 'the' X role, and I will for the most part talk about 'the' role a concept has. But obviously, if the role talk is acceptable in the first place, a given concept can be used to play different roles. Taking this into account would complicate some of my formulations. I will instead continue to, naively, speak about 'the' X role.

It is important to be clear on exactly what the talk of roles amounts to in the present context. It is essential to the present notion of role that different, non-coextensive concepts may play the same role. It is if this assumption is met that we can ask significant questions about which one of these concepts best plays this role. Compare, by way of contrast, the notion of role that is employed in David Lewis's philosophy of language. On Lewis's descriptivist view, descriptions associated with a term amount to a reference-fixing theory associated with the term, and the term refers to whatever best satisfies this associated theory. This is often expressed as: the theory specifies a *role* for something to play, and the term refers to whatever plays this role or comes closest to doing so.[9] Assuming this conception of reference-determination works for 'knows', it is a given that 'knows' ascribes the relation that best plays the knowledge role, in this sense.

What is the knowledge role, in the present sense? At a first stab, one may appeal to the use of knowledge in epistemic evaluation. But by itself, this is rather unhelpful. There are different kinds of epistemic evaluation. The uses of the concept KNOWLEDGE are different from the uses of JUSTIFICATION, RELIABILITY, ... To play the knowledge role is to play the specific role knowledge plays in epistemic evaluation. But this slogan is problematic in two different ways. First, it sounds rather uninformative. (Concept X plays the X role—duh.) Second, uninformative though it may be, on one natural way of understanding the slogan it may yield unwanted results. Compare our actual concept KNOWLEDGE with the concept JUSTIFIED TRUE BELIEF. Assuming that a lesson from the Gettier cases is that for knowledge that P some anti-luck condition not satisfied by mere justified true belief would have to be met, one may think: so the specific role of knowledge in epistemic evaluation is to rule out epistemic luck of the relevant kind. So justified true belief does not play the knowledge role; not even poorly. This is an unwanted result, not because it is a given that the concept JUSTIFIED TRUE BELIEF must count as a knowledge concept but because the reasoning seems to generalize. For many broadly knowledge-like concepts not coextensive with the actual concept KNOWLEDGE one could construct similar arguments that these concepts are not apt to play the knowledge role: given the differences in extension between such a concept C and the concept KNOWLEDGE, C does not play the same role as KNOWLEDGE. Or so the reasoning runs. For the talk of the knowledge role to do the work it is supposed to do in the context, roles must somehow be individuated differently.

In the case of quantifier variance, I noted that one can sidestep questions about what it is to be an existence concept by appealing to a notion of liberalized variance. One can similarly seek to sidestep questions about what it is to be a knowledge concept by appealing to a corresponding notion of liberalized variance—a notion of

[9] See, for example, section 2 of Schwarz (2015) for a nice exposition of Lewis's views on this.

liberalized epistemic variance. The idea would be that one can sidestep questions about what it is for a concept to be a knowledge concept by appealing instead to languages that contain expressions expressing different epistemic concepts but are equally good, in whatever dimension is relevant for epistemic purposes. This avoids the problems regarding how to identify knowledge concepts, but immediately highlights a separate question about the various X variance theses: how is the 'equally good' to be understood? In the metaontological case this is, as above noted, often cashed in terms of eliteness. But even assuming that this is the best way to cash it there, that does not mean that this is always the best way to cash talk of concepts being equally good. A philosopher with her metaphysics hat on can certainly embark on the project of evaluating all sorts of concepts for eliteness, and nothing I wish to say here is meant to suggest that this would not be a worthwhile project. But even if it is a worthwhile project, it is not clear what relevance it has for epistemology, or philosophy of action, or.... To elaborate: One way that variance-type issues can arise for the epistemologist is that one can reasonably worry that concepts different from our actual ones are better for epistemic purposes. Goodness for these epistemic purposes need not line up with eliteness. There are separate questions about what concepts are best for these epistemic purposes.

Another illustration of how philosophically central concepts may be evaluated along different dimensions is provided by the case of personal identity. One can approach the issue of personal identity with the metaphysician's hat on and wonder which person-like entities exist, which person-like entities are the most fundamental or joint-carving, and so on. Or one can approach the issue with the practical philosopher's hat on, wondering about how praise and blame should be distributed, and how our prudential concerns should be structured. The investigations may line up. Maybe the person-like entities that are metaphysically privileged are the ones that are relevant to the practical philosopher's questions. Maybe the metaphysically privileged person-like entities are bodies and it is also the case that if person A at t does something blameworthy then it is the person with A's body at t* who ought to be blamed for this. But the investigations may also come apart. Maybe a physical criterion of personal identity is correct for the metaphysician's purposes while a psychological criterion of personal identity is correct for the practical philosopher's purposes. More radically, it may be that as far as the practical philosopher's purposes are concerned, the focus on identity is misplaced. Parfit's (1971, 1984) arguments regarding, for example, fission are naturally seen as having this upshot. What "matters" in personal identity is the holding of a psychological relation which does not have the logical characteristics to be an identity relation.

Focusing on variance theses is of a piece with seeing philosophy as *conceptual engineering*. Those focusing on conceptual engineering think that rather than resting content with what concepts we actually have, we should think about how these concepts can be improved or replaced. To focus on variance is to shift attention from our actual X concept to what possible X concepts there are and how they are to be ranked along some dimension.

But there are differences, at least differences of emphasis, between conceptual engineering and focus on variance theses. The conceptual engineering project is held to have some practical import: recommendations are made regarding which concepts to use. One can in principle be skeptical of that project—"is it really the business of philosophy to reform language and thought?"—and still think that there are reasonable variance questions to ask about betterness along some dimension. Using 'conceptual ethics' as a label for the enterprise of evaluating concepts, studying variance theses is part of conceptual ethics, even if it is not directly geared to proposals for language reform or conceptual reform.

That said, one should not exaggerate the differences between the overtly practical, activist project of conceptual engineering and the on the face of it more theoretical project of evaluating concepts along different dimensions. Friends of conceptual engineering tend to stress that they are not concerned to reform how we ordinarily think and talk, but only suggest replacement *for particular purposes*.[10] And if one says a concept is better than another along one dimension, one thereby says that the concept is better to use for some associated purpose.

Still, even though one should not exaggerate the difference, there is a difference between the variance-related project of mapping what possible concepts there are and how they are related, and the engineering project of making particular recommendations regarding concept use. (A well-known point from ethics serves to highlight and dramatize the difference: One can think that consequentialist normative concepts are the ones that get at the features that really matter normatively, while at the same time—and on consequentialist grounds—thinking that it would be bad if agents making decisions about how to act deployed consequentialist concepts.)

3. Thin Normative Concepts

One place where one can ask questions similar to those that have come up in metaontology is in the case of normative concepts:

Normative variance: there are different rightness, goodness, . . . concept, and none of them is privileged.

Verbalism about normative discourse: (many) normative disputes are purely verbal.

Verbalism about normative discourse does tend to come up as a topic in philosophical discussions, but typically in negative arguments. If a theory of how the reference of normative terms is determined leads to verbalism about normative discourse, that is seen as reason to give up the theory. This is for example a theme in the lively debate over moral disagreement.[11]

As before, and for the same reasons, I think the variance thesis is the more significant one. So regardless of the plausibility, or not, of verbalism I will set it aside and focus on variance.

[10] So, for example, in Scharp (2013), the main theme of which is that the ordinary concept TRUTH ought to be replaced, Scharp keeps reminding the reader that the replacement is only for certain purposes. For everyday use, the ordinary concept works just fine.

[11] See, for example, the Moral Twin Earth argument due to Horgan and Timmons (1992 and 2009).

I have elsewhere discussed at some length various questions relating to normative variance (although I have primarily discussed these matters in different terms).[12] In this section, I will briefly rehearse some main points.

The same questions arise regarding normative variance as regarding other variance theses. What makes something a rightness concept (goodness concept, etc.)? What is it for a rightness concept (goodness concept, etc.) to be privileged? I will soon pause on these crucial questions. But before focusing primarily on these questions, let me first pause on the significance of the normative variance thesis. The discussion of this matter will *inter alia* shed some light on the questions mentioned.

Consider the following scenario (which may or may not be possible):[13]

> *Tragic* There is a linguistic community—the Tragic—speaking a language much like English, except for the following differences (and whatever differences are directly entailed). While their words 'good', 'right', and 'ought' (in their "thinnest" uses) are associated with the same normative roles as our words 'good', 'right', and 'ought' (in their "thinnest" uses) are associated with, their words aren't coextensive with our 'good', 'right', and 'ought'. So even if they are exactly right about what is 'good' and 'right' and what 'ought' to be done, in their sense, and they seek to promote and to do what is 'good' and 'right' and what 'ought' to be done in their sense, they do not seek to promote what is good and right and what ought to be done. Moreover, what their 'good', 'right', and 'ought' are true of are things that really ought not to be valued: their normative language is in that way off.

I am not claiming that the Tragic scenario is in fact possible. But there are prominent views on normative language on which it clearly is. For example, views on the reference of normative terms on which the reference is determined causally as much as the reference of natural kind terms is usually held to be, or views on which widely held beliefs linking these normative terms to the descriptive play a reference-fixing role. Their use of their normative terms may be causally linked in the relevant way to properties that ought not to be valued, or their widely held beliefs link the terms to properties that ought not to be valued.

Now, if Tragic is possible, then it would appear that we could in principle be in the same kind of situation. If the reference of normative terms can be determined in either of the ways characterized, the same can go for the reference of our normative terms. But then our positive (/negative) terms too can be causally linked to properties that do not warrant the positive (/negative) evaluation associated with the term. It is harder to *state* the problem as it arises in our own case. For we use our own normative terms when attempting to state what the supposed problem is regarding our normative terms. (Compare a loose analogy: radical indeterminacy arguments such as those presented by Kripke and Quine are typically not first presented as concerning our language now: it seems clear that, for example, 'rabbit' refers to rabbits. The standard strategy is to argue that a language qualitatively like ours is radically indeterminate, and then note that our language cannot be different in that regard.)

[12] Eklund (2017).

[13] Again, see Horgan and Timmons, for example (1992 and 2009).

Let me now state the issue just brought up in more general terms. Consider a scenario like those encountered in the Moral Twin Earth literature:[14]

> *Alternative* There is a linguistic community speaking a language much like English, except for the following differences (and whatever differences are directly entailed). While their words 'good', 'right', and 'ought' (in their "thinnest" uses) are associated with the same normative roles as our words 'good', 'right', and 'ought' (in their "thinnest" uses) are associated with, their words aren't coextensive with our 'good', 'right', and 'ought'. So even if they are exactly right about what is 'good' and 'right' and what 'ought' to be done, in their sense, and they seek to promote and to do what is 'good' and 'right' and what 'ought' to be done in their sense, they do not seek to promote what is good and right and what ought to be done.[15]

Typically when scenarios like this are brought up, they are used to evaluate the plausibility of the predictions of different theories of reference-determination. More specifically, we are supposed to have the intuition that the words really are coextensive, and the scenario described not possible—and theories in conflict with this are to be rejected. Thus employed, the scenarios are used to gauge how our actual words and concepts are used. In the context of considering normative variance, a different question arises: *Suppose, provisionally, that scenarios like Alternative are possible. Then what?*

One natural thought is that there then is a question of which of the rightness concepts one ought to employ. Some things are right; others are right*. There is a question of whether to act in accordance with what is ascribed by one concept or what is ascribed by the other.

But there are complications regarding how to understand this supposed further question. When trying to state it just now I used '*ought* to employ'. But if there are different rightness concepts, there are different ought-concepts. And the question of which rightness concept we ought to employ is different from the question of which rightness concept we ought* to employ. If we use one ought-concept we ask one question; if we use another, we ask a different question. Neither of these questions seems to be the one we wanted to ask when we wondered which rightness concept was objectively privileged. For they both are stated *using* some normative vocabulary or other, and what is in question is the propriety of using these pieces of normative vocabulary. Relatedly, one may suspect that it is rather trivial that we ought to care about what is right but ought* to care about what is right*. These two sets of facts—about what we ought to care about and ought* to care about—don't immediately bring us any closer to the practical question of how to structure our concerns.

The same sort of problem would seem to arise regardless of which normative vocabulary we would use when trying to state the supposed further question. And if instead we tried to state the further question using only descriptive, non-normative vocabulary, our attempts to state the supposed further question would misfire in

[14] Again, see Horgan and Timmons, for example (1992 and 2009).

[15] From Eklund (2017: chapter 2).

another way: we didn't just wish to know which descriptive concepts the various rightness concepts fall under but which one, so to speak, *really ought* to guide action.

Noting the difficulties in stating the supposed further question, one might wish to deny that there is a further issue there. There is what is right, what is right*, etc., *and that is that*. It is right to do what is right, right* to do what is right*, etc. Maybe this in the end is the correct view. But I think many of us are intuitively inclined to take normativity to be *objective*, in such a way as to find this view repugnant. (Suppose, e.g., to dramatize things, that among the actions that are right* are some that we find deeply abhorrent.)

The notion of the objectivity of normativity alluded to in the last paragraph is important, but elusive. Even someone who holds that there is what is right, what is right*, etc., and that is that *can* hold on to the objectivity of normativity in the sense that she can hold that it is an objective matter what is right, an objective matter what is right*, etc. Facts about what is right, right*, etc., are normative facts, it may be said, and it is an objective matter whether they obtain or not. The sense in which the objectivity of normativity is jettisoned on this view is that there does not seem to be a fact of the matter as to whether to go with what is right or what is right* (or . . .) in one's choices about how to act.

We can put the above reflections in the form of a dilemma, what I will call the *alternative concepts dilemma*. Either there is a further question of the kind indicated or there is not. The former alternative seems problematic, for the supposed further question would be unstatable. The latter alternative seems problematic for the mere 'that is that' does not capture our sense that the normative is objective.

Return now to the questions about how to understand normative variance. One question was: what is it for a concept to be a rightness concept (etc.)? A natural reply is to appeal to a concept's normative role—its role in action-guiding and deliberation, perhaps its relation to reactive attitudes, etc. Saying that there are different rightness concepts then amounts to saying: there are different concepts associated with the same normative role in this sense but different in other ways, so that they are not coextensive. And the view that there are different rightness concepts in this sense is what gives rise to the alternative concepts dilemma. These remarks on normative role are obviously vague and sketchy: but one can still see that if a concept's reference is determined by what its use is causally related to or by what descriptions the concept is associated with, its reference is not determined by normative role alone.

When it comes to what makes a rightness concept privileged, the problems in cashing this out were in effect displayed through the discussion of the alternative concepts dilemma. If we explain this using normative terms the problem is that there are different normative terms we could use. If we do it using descriptive terms we don't seem to address the right thing.

Let me elaborate on the last point. As earlier stressed, there are many dimensions along which concepts may be evaluated and compared. For some purposes, for example when one is concerned with metaphysics for metaphysics' sake, one may wish to compare rightness concepts in terms of eliteness. If one tries to do so in the present case, one tries to ask the further question in descriptive terms. When it is insisted that we are asking the wrong question about privilege if we attempt to pose this question in descriptive terms, what is claimed is only that if we are approaching the issue of variance from the perspective of normative theorizing and we are primarily concerned with questions about how to think and act. The metaphysician's question of which rightness concept best carves the world at the metaphysical joints is not immediately such a question.

When discussing the quantifier variance thesis and its supposed deflationary consequences for ontology, I mentioned that one may consider it sufficient for ontology to be in good standing if some existence concept is privileged: it does not matter if other existence concepts also are privileged. Even if this is a reasonable view in the case of ontology, its counterpart in the present case would be misguided. Assuming one sees the considerations that I have brought up in the present question, by appeal to the scenarios *Tragic* and *Alternative*, as serious, it does not help at all if my rightness concept is privileged, if some alternative rightness concept is equally privileged.

One natural way to attempt to avoid the alternative concepts dilemma is to deny that there are these different rightness concepts to begin with: Alternative is not possible. One can insist that normative role *determines reference* so that if two concepts are associated with the same normative role they are guaranteed to have the same reference (and more generally, same intension). On a fine-grained way of individuating concepts one can perhaps insist that there still are different rightness concepts, it is only that they are all coreferential. But since the concepts necessarily apply to the same things, there is no momentous question regarding which one to employ. The dilemma is avoided, for if there are not these different rightness concepts, questions about what to say about one being privileged do not arise.

A variance thesis is a conjunction of two claims: one to the effect that there is a multitude of concepts of such-and-such a kind, and one to the effect that no concept in this multitude is privileged. The present suggestion amounts to denying the first conjunct of the relevant normative variance thesis. The move has an analogue in the original metaontology case: it can be insisted that there is no multitude of existence concepts, for example on the ground that any two concepts governed by the standard inference rules associated with existential quantification must be coreferential.[16]

Of course, good questions can be asked about whether normative role can indeed determine reference in such a way that the alternative concepts dilemma can be avoided in the way suggested. But my aim here is not to evaluate this suggestion. I am only concerned with the conditional claim that if normative role determines reference, then problems like the ones brought up in connection with the alternative concepts dilemma can be avoided.

This conditional claim itself can reasonably be resisted. If normative role does determine reference, there can be different possible normative concepts associated with *slightly different normative roles*, and the same issues as before can be brought up by appeal to such possible concepts. Compare again ontology. Even if, among the different existential quantifier meanings there are, where the condition for being an existential quantifier meaning is that of satisfying the classical inference rules, some unique meaning is privileged, there can be other quantifier*ish* meanings, satisfying

[16] For relevant discussion, see, for example, Williamson (1988); McGee (2006); and Turner (2010).

some slightly different rules, and no quantifierish meaning is privileged over all the others.) Compare too how in the ontology case one could without obvious loss consider a liberalized variance thesis instead of the quantifier variance theses focusing specifically on existence concepts. Analogously one can in principle, in the normative case, focus on a liberalized normative variance thesis which does not focus specifically on rightness concepts (or goodness concepts, or ought concepts, or reason concepts), but instead concerns different normative languages as wholes.

Moreover, just to make things really confusing (sorry!): just as one can reasonably think that there are alternative notions of rightness, goodness, etc., and questions about whether any particular ones among these notions are privileged, one can reasonably think that there are alternative notions of reference and questions about whether any notion of reference is privileged. With this complication in mind: which notion of reference should we employ in the thesis that normative role determines reference?

4. General Lessons

I have talked about issues related to normative variance, and compared the issues that come up in this case with parallel issues that come up in the parallel metaontological debate. Let me now ask: are there more general lessons regarding variance and conceptual engineering to be learned here?

Consider first the possibility that there genuinely are different rightness concepts. Then, as stressed, there arises the question of whether one is privileged, and, more fundamentally, what privilege amounts to in the relevant case. In the case of rightness concepts, there were problems regarding what privilege might amount to. There seemed to be no way of getting at the supposed further question.[17]

One limitation of variance inquiry is presented by the type of case we may be faced with here: there is no way to make sense of the relevant question of privilege.

Let an *ultimate* concept be an X concept such that there are other X concepts and the question of which X concept is privileged cannot be asked in suitably independent terms. The possibility just described is that thin normative concepts are ultimate concepts.

Another possibility in the case of rightness concepts is that there are not, in the relevant sense, different rightness concepts: normative role determines reference, so any concept with the normative role associated with rightness has the same reference. I mentioned that it can be and has been argued to be so also when it comes to existence: the inferential role associated with being an existence concept is such that no two non-coextensive concepts can be associated with this role.

[17] *Intuitively*, this is related to rightness being in some sense *basic*. When it comes to other, less basic normative concepts, like paradigmatic thick concepts such as COURAGEOUS, LEWD, RUDE, . . . , one can reasonably think that, say, a courageousness concept being privileged simply is a matter of it being the courageousness concept one *ought* to evaluate people and actions in terms of. However, once one has problematized 'ought' and raised to salience the possibility that there are different possible ought-concepts, matters look more complicated here too.

Let a *fixed* concept be an X concept such that there are no other X concepts. The possibility just described is then that the concept rightness is a fixed concept.

Ultimate concepts and fixed concepts can appear to present complications for evaluation of variance theses, and for the project of conceptual engineering. When C is ultimate we cannot get a handle on the relevant question of privilege when it comes to C: that is part of the characterization of what it is for a concept to be ultimate. And when C is ultimate, some questions we wish to ask about whether to replace C cannot really be asked. When it comes to fixed concepts, the problem is that when X is a fixed concept, there just are no other X concepts to replace X with.

However, I do not think that fixed concepts in fact do present serious theoretical problems for inquiry into variance theses. Even if there is only a unique X concept there can be concepts in various ways similar to X concepts, and one can still ask whether the X concept is privileged, in whichever respect is relevant, over these other, similar concepts. And if we have a situation where there not only is a unique X concept but moreover there is only one concept in question that can serve the purpose at hand, so that there is no competition, this is not so much an obstacle to inquiry into variance theses as a result regarding what sort of variance there can and cannot be.

Ultimate concepts are a different matter. If a concept C is ultimate but not fixed, then there are alternative concepts that can be used, but the question of whether C or some alternative concept is privileged cannot be asked in suitably independent terms. This is a real limitation to variance inquiry.

In the discussion of thin normative concepts we saw that there is a real threat that thin normative concepts are—in the terminology now introduced—ultimate. But even if this is so in the case of thin normative concepts, is it plausible that there are other instances of this phenomenon, and instances that are not immediately bound up with the problems having to do with thin normative concepts?

Compare a toy example. (With possible similarities to actual debates. But I want to discuss a simple made-up case, in abstraction from various complexities.) Suppose that one group of metaphysicians are fundamentally concerned with what is REAL and another group is fundamentally concerned with what is REAL*. The first group of metaphysicians think that in some deep sense there is nothing more to reality than what is REAL, and the other group think the same about what is REAL*. (Lots here is sketchy: what 'deep sense'? and what does the 'reality' in small letters mean—REALITY, REALITY*, or something else? But I believe the sketchiness does not actually matter to the questions I am about to bring up.) Then the groups attempt to ask the question: is what is REAL or what is REAL* privileged for the purposes of metaphysical theorizing? There is a sense that we may be dealing with something ultimate: that when attempting to ask the relevant question of privilege one group will ask whether it is what is REAL or what is REAL* that is REAL, and the other group will ask which it is that is REAL*. Both questions are trivial; neither gets at an interesting underlying question of privilege.

It may be retorted that there is a way to get a suitably independent handle on whether REAL or REAL* is privileged: one need simply consider whether overall theories of the world which employ one concept is more theoretically virtuous—simpler, more explanatory,... —than the other. This is no different from choice of ideology in theory construction quite generally. Of course, it could turn out that one cannot in fact use this method to choose ideology: REAL and REAL* score equally high. But again this would just be an instance of a familiar phenomenon: underdetermination of theory by data.

But what if the REAL-users say: So what if a theory instead employing REAL* would be in these ways more theoretically virtuous? That theory, since it does not speak of what is REAL, does not state what the world is REALLY like, and it is a theory that states this we should aim for. If the REAL-users respond this way, REAL functions for them as an ultimate concept.

Acknowledgements

Many thanks to audiences at Cornell University, Trinity College Dublin, the Moral Sciences Club at the University of Cambridge, a workshop at University of Leeds, and two Conceptual engineering workshops, at University of Oslo and in Zermatt, for useful feedback. Special thanks to Tim Button, Andy Egan (who was my commentator in Zermatt), and David Plunkett for helpful comments.

References

Brandt, Richard. 1979. A Theory of the Good and the Right. Oxford: Clarendon Press.

Dreier, James. 2004. Meta-ethics and the Problem of Creeping Minimalism. *Philosophical Perspectives* 18:23–44.

Eklund, Matti. 2011. Review of Eli Hirsch, Quantifier Variance and Realism. *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/news/24764-quantifier-variance-and-realism-essays-in-metaontology/

Eklund, Matti. 2016. Carnap's Legacy for the Contemporary Metaontological Debate. In Stephan Blatti and Sandra Lapointe (eds.), *Ontology after Carnap*. Oxford: Oxford University Press.

Eklund, Matti. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.

Fine, Kit. 2001. The Question of Realism., *Philosophers' Imprint* 1.

Hirsch, Eli. 2011. *Quantifier Variance and Realism*. New York: Oxford University Press.

Horgan, Terence, and Timmons, Mark. 1992. Troubles for New Wave Moral Semantics: The 'Open Question Argument' Revived. *Philosophical Papers* 21:153–75.

Horgan, Terence, and Timmons, Mark. 2009. Analytical Moral Functionalism Meets Moral Twin Earth. In Ian Ravenscroft (ed.), *Minds, Ethics and Conditionals* (pp. 221–37). Oxford: Oxford University Press.

Lewis, David. 1983. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61:343–77.

Lewis, David. 1984. Putnam's Paradox. *Australasian Journal of Philosophy* 62:221—36.

McGee, Vann. 2006. A Rule for Everything. In Agustín Rayo and Gabriel Uzquiano (eds.), *Absolute Generality* (pp. 179–202). New York: Oxford University Press.

Parfit, Derek. 1971. Personal Identity. *Philosophical Review* 80:3–27.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Railton, Peter. 1986. Moral Realism. *Philosophical Review* 95:163–207.

Schaffer, Jonathan. 2009. On What Grounds What. In David Chalmers, David Manley, and Ryan Wasserman (eds.), *Metametaphysics* (pp. 347–83). Oxford: Oxford University Press.

Scharp, Kevin. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Schwarz, Wolfgang. 2015. Analytic Functionalism. In Barry Loewer and Jonathan Schaffer (eds.), *A Companion to David Lewis* (pp. 504–18). Oxford: Blackwell.

Sider, Theodore. 2009. Ontological Realism. In David Chalmers, David Manley, and Ryan Wasserman (eds.), *Metametaphysics* (pp. 384–423). Oxford: Oxford University Press.

Sider, Theodore. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Turner, Jason. 2010. Ontological Pluralism. *Journal of Philosophy* 107:5–34.

Williamson, Timothy. 1988. Equivocation and Existence. *Proceedings of the Aristotelian Society* 88:109–27.

Wright, Crispin. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

11

Neutralism and Conceptual Engineering

Patrick Greenough

1. Paradox and Conceptual Engineering with Concepts

Conceptual Engineering *with concepts* is the view that philosophical problems arise because our concepts are defective in some way. Resolving such problems involves suitably revising our concepts or replacing them with new and better surrogates.[1] When it comes to paradoxes, the source of such puzzles is taken to stem from *inconsistent concepts*. Roughly, these are concepts which are governed via conflicting rules—rules which, if sufficiently pressed, give incompatible instructions as to when to apply the concept. Philosophical progress on paradox consists in revising such concepts such that they are no longer inconsistent, or in replacing such concepts with consistent concepts. Once we do so, our most intractable paradoxes will disappear, or so goes the thought.

2. Paradox and Conceptual Engineering without Concepts

Conceptual Engineering, despite its name, need not (and perhaps should not) invoke concepts. Conceptual Engineering *without concepts* is the view that philosophical problems arise because our words have defective *meanings*. Resolving such problems involves suitably improving the meanings of our words or replacing these words with new and better surrogates.[2] With respect to paradoxes, the source of such puzzles is taken to stem from *inconsistent words*. Roughly, these are words which are governed

[1] Candidate Conceptual Engineers include: Tarski (1944); Carnap (1950); Deleuze and Guattari (1991/1994); Schiffer (1996, 2003, 2004); Scharp (2007, 2013); Burgess and Plunkett (2013a,b); Burgess (2014); Eklund (2015); Thomasson (2016); Díaz-León (2017); Simion (2017). Haslanger (2000, 2005, 2006, 2012) is usually taken to be a paradigm Conceptual Engineer when really she is merely interested in having a better theory about *women*—she is not trying to revise the meaning of "woman". Let me just say that again: Haslanger is not a Conceptual Engineer. If you think she is then everybody counts as a Conceptual Engineer, and the debate has been trivialized.

[2] For Cappelen (2018), the Conceptual Engineer should aim to improve our representational devices (words and thoughts). (See also Sider 2014; Leslie 2017.) On the Cappelen view, (inconsistent) concepts drop out of the picture; inconsistent words do not. Call such a view: *Meaning Engineering*. Conceptual

Patrick Greenough, *Neutralism and Conceptual Engineering* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Patrick Greenough.
DOI: 10.1093/oso/9780198801856.003.0011

via conflicting rules—rules which, if sufficiently pressed, give incompatible instructions as to when to apply the word. Philosophical progress on paradox consists in suitably revising the meanings (and use) of inconsistent words so as to make these words have a consistent meaning, or in replacing an inconsistent word with a word which is not governed by conflicting rules of use. Once we do so, our most intractable paradoxes will disappear, or so goes the thought.

3. Happy-Face Treatments

Conceptual Engineers often invoke a distinction between *happy-face* and *unhappy-face* solutions to paradoxes.[3] Happy-face solutions involve identifying and rejecting some false or invalid principle ("the culprit") used in the generation of some paradox (and also explaining why we were initially taken in by this culprit).[4] These treatments are thoroughly *specific*: they isolate a single, basic culprit—no further more specific principle is to blame.

4. Unhappy-Face Treatments

Unhappy-face solutions, meanwhile, are thoroughly *non-specific*: they merely establish the collective guilt attaching to the group of principles which together produce the paradox. When unhappy-face treatments succeed in suitably revising or replacing one or more defective concepts at work in some paradox then that yields a *weak* unhappy-face solution. Sometimes, a conceptual revision or replacement is out of the question—the cure may be worse than the disease.[5] The best that can be hoped for is a kind of palliative conceptual care—a *strong* unhappy-face treatment.[6] According to the standard version of this taxonomy, Conceptual Engineering can only occur via weak unhappy-face solutions and not via happy-face treatments.

5. What's the News?

> *News Item One*: A new taxonomy is needed which allows for both Happy-Face Conceptual Engineering, and two forms of Unhappy-Face Conceptual Engineering. The first is *The Indeterminate Concept* View, whereby it is indeterminate, and so unknowable, which principle is the culprit.[7] The second is *The Indiscriminable*

Ethics (Burgess and Plunkett 2013a,b; Plunkett and Sundell 2013; Plunkett 2015, 2016), which is broader in scope than Conceptual Engineering, can also take place with or without concepts.

[3] This taxonomy is due to Schiffer (1996, 2003, 2004) and has been co-opted by other Conceptual Engineers such as Scharp (2013). Cook (2013) and Cuonzo (2014) also use it in their accounts of paradox.

[4] A paradox may involve more than one false or invalid principle. I will ignore this point in what follows. The apt terminology of "culprit" is taken from Eklund (2002a).

[5] See Chihara (1979) on Tarski.

[6] Schiffer (2003) thinks that the problem of free-will requires a strong unhappy-face response, while Skepticism (Schiffer 2004) and the sorites paradox (Schiffer 2003) require weak unhappy-face treatments. In what follows, I will focus on weak unhappy-face treatments.

[7] Schiffer (1996, 2003, 2004) and Eklund (2002b) are the two most overt defenders of this view.

Concept View, whereby our limited powers of conceptual discrimination make it infeasible to identify the culprit.

News Item Two: Happy-face treatments (whether effective or not) are extremely rare—they represent a kind of limit case.

News Item Three: Unhappy-face treatments (whether effective or not) are also rather rare—they also represent a kind of limit case.

News Item Four: Between these limit cases are treatments which are neither maximally specific nor maximally unspecific. These intermediate treatments can nonetheless be *specific enough* to effectively treat a paradox. For example, a solution may be specific enough to tell us that the conjunction of two premises of the paradox is false, but not specific enough to tell us which one of these premises is false—they remain suitably silent, and so neutral, on this issue.

News Item Five: Such intermediate treatments become more thoroughly neutral when they reject some principle at work in a paradox from a theory-neutral perspective. With respect to various forms of skeptical paradoxes, for example, we can give effective remedies which merely use relatively lightweight and fairly uncontentious theoretical claims about knowledge (and evidence). The upshot is *Neutralism*—the view that philosophical progress can take place when (and sometimes only when) a thoroughly neutral, non-specific theory, treatment, or method is adopted.[8]

News Item Six: Neutralism is available to the Conceptual Engineer and the Semantic Engineer.[9]

6. The Test-Bed of Philosophical Theory

Philosophical puzzles and paradoxes yield a kind of test-bed for philosophical theory. Adapting some remarks of Russell, we may say that:

> A logical [or philosophical] theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about [philosophical questions], to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science.[10]

That is, if your favored philosophical theory of X cannot address the relevant puzzles and paradoxes which centrally involve X, then your theory is akin to a scientific theory which cannot accommodate the experimental data. This is important because, at one extreme, there are prominent Conceptual Engineers (and philosophers more

[8] For intimations of Neutralism, see Greenough (2002, 2003). See Greenough (2019) for an application of Neutralism to the Observational Sorites Paradox. See Greenough (MSb) for the developed view.

[9] I should say from the outset that I don't have much sympathy for Conceptual Engineering. In Greenough (MSa), I argue that prescriptive philosophy does not consist in revising or replacing our concepts and/or the meanings of our words. Rather, we merely need to revise or replace our ideas, beliefs, theories, and conceptions about the things picked out by those words. *Idea Engineering* is all we really need. The main goal here is not to criticize or defend Conceptual Engineering, but rather show that Conceptual Engineers can (and should) help themselves to *Neutralism*.

[10] Russell (1905: 484–5).

generally) who think that puzzles and paradoxes do not (or should not) play any central role in philosophy. At the other extreme, there are some philosophers who think that paradoxes exhaust the main business of philosophy.[11] Russell's point is that an intermediate position is called for whereby paradoxes play a core role.

7. The Standard Account of Paradox

Standardly conceived, a paradox is an argument that proceeds via seemingly valid reasoning from seemingly true premises to a seemingly false conclusion.[12] Relatedly, a paradox is an argument that proceeds via plausible reasoning from plausible premises to an implausible conclusion.

8. A Flaw

Despite being pretty widespread, the standard account has an immediate flaw. Once a (typical) subject notes that the premises of some argument are part of a putative paradox then she may (but need not) reasonably retract her original judgment that the premises are plausible, the reasoning plausible, and the conclusion implausible—she may reasonably sit on the fence until she has worked out what is going wrong. Still, the argument remains a paradox all the while—contrary to what is predicted by the standard account. Equally, once a subject has been exposed to some promising solution for long enough then she may no longer find, for example, that a certain premise in the proof is plausible (or seems true). Again, the argument remains a paradox all the while—contrary to the standard account.[13]

9. The Standard Account Tweaked

Properly understood, a paradox is an argument that proceeds via reasoning which seems *initially* to be valid, from premises which seem *initially* to be true, to a conclusion which seems *initially* to be false. Relatedly, a paradox is an argument that proceeds via *initially* plausible reasoning from *initially* plausible premises to an *initially* implausible conclusion.[14]

10. Treating Paradox: The Standard Account

It is also part of the standard view that a good treatment of a paradox must do at least two things:

[11] Sorensen (2003).

[12] See Mackie (1972); cf. Sainsbury (2009).

[13] Equally, it is sometimes said that putative paradoxes for which we have a solution are not paradoxes proper. That's not a good way to think about paradoxes. We don't speak of diseases for which we have a cure as not being diseases proper.

[14] Cf. Schiffer's useful formulation whereby "a paradox is a set of apparently mutually incompatible propositions each one of which enjoys some non-negligible [or better: high] degree of plausibility *when considered on its own*" (Schiffer 2003, my emphasis).

(1) Provide good reason to: reject some basic premise in the paradoxical proof; or, reject some basic rule of inference; or, reject some basic presupposition(s) of the proof; or, give good reason to "bite the bullet" and endorse the conclusion. (Here the culprit is basic in the sense that there is no more specific culprit to be found.)

(2) Explain why we were so initially susceptible to the paradox—despite the faults isolated in (1). That is, explain how and why we found the premises, rules of inference, or presuppositions, so *initially* plausible. Or, if biting the bullet, we must explain why the conclusion struck us as so *initially* implausible despite being true/acceptable after all.

It is clear that this standard view is an account of how to give a happy-face treatment.

11. Treating Paradox: A Third Desideratum

Something important is missing from the standard account. Suppose we have some particularly stubborn, intractable paradox. It's one thing to provide an explanation as to why we were initially seduced by this paradox; it's potentially quite another thing to explain why this paradox has proved to be so tricky to treat. We thus need to distinguish two properties of paradoxical arguments: *contagiousness* (the easy-to-catch property), which is covered by desideratum (2) above, and *intractability* (the hard-to-cure property). These may come apart both ways. Just because it is easy to be initially seduced by some paradox does not entail that some resolution will be hard to find—perhaps our faulty thinking is perfectly natural but easy to correct once noticed. Equally, perhaps it takes a while for some paradox to get a grip; but, once it does so, it proves very difficult to dislodge. Given this, a complete response to some intractable paradox must also answer the following questions:

(3) Why has this long-standing paradox proved to be so intractable? More generally: Why are intractable paradoxes intractable?

This third desideratum will come into greater relief below.

12. Inconsistent Concepts

To make sense of philosophical puzzles and paradox, Conceptual Engineers (who deploy concepts) place the notion of an *inconsistent concept* centre-stage. On the most prevalent conception, inconsistent concepts are concepts whose conceptual principles cannot all be true.[15] Given classical logic, it follows that one or more conceptual principles for the concept is/are false.[16] A cartoon example is the concept of *blair* whose conceptual principles include: x is a blair if x is a chair; x is not a blair if x is blue. Thus, these conceptual principles entail something contingently false: there are no blue chairs. Other inconsistent concepts have conceptual principles which entail something necessarily false. Take the invented concept *tallster* which has the two conceptual principles: x is a tallster if x is taller than 2m in height; x is not a

[15] See Eklund (2002a); Scharp (2013).

[16] The most prominent forms of Conceptual Engineering (with concepts) retain classical logic.

tallster if x is less than 2.1m in height. According to this concept, someone who is 2.05m in height is both a tallster and not a tallster. Since this cannot be true then at least one of these conceptual principles must be false (given classical logic).

13. Concepts and Conceptual Principles

What are conceptual principles? Say that concepts are constituted by (or fixed by) their conceptual principles, where a principle is a conceptual principle for a concept C if and only if S's understanding C entails that S bears relation X to this principle. The epistemic version of this view says that X is the knowledge relation; the justificationist version says that X is the justified belief relation; the doxastic version says that X is the belief relation; the dispositional view says that X is the disposed to believe relation.[17] We will mostly be concerned with the latter dispositional view.

14. Why are Certain Paradoxes so Contagious and/or Intractable?

Why do we get so caught up in paradoxes? Candidate diagnoses include: easily confusing one principle with another, oversight, ignorance, intellectual prejudice, over-generalization, and the hasty use of false theory. The Conceptual Engineer, meanwhile, is able to offer a very different kind of diagnosis, at least for the most obstinate paradoxes: our mastery of these concepts explains why the paradox is both initially seductive (contagious) and hard to treat (intractable). Such mastery disposes us to accept the conceptual principles for the concept in question—even when one of these conceptual principles is false.

15. Can the Conceptual Engineer Embrace Happy-Face Treatments?

The happy-face/unhappy-face taxonomy, as standardly presented, is incomplete and misleading: it entails that the Conceptual Engineer cannot avail themselves of happy-face solutions. Here the standard thought is something like: "A happy-face approach to paradox is just the traditional, purely descriptive approach to paradox. Conceptual Engineering, meanwhile, calls for a prescriptive approach. Hence, Conceptual Engineering can only involve unhappy-face solutions". That runs together two independent axes of paradox: the descriptive/prescriptive axis with the happy-face/unhappy-face axis. Conceptual defects can be non-specific (a collective defect) or specific (an individual defect); either way, Conceptual Engineering can be used to fix the problem. Happy-Face Conceptual Engineering is thus an eminently live option.

[17] Eklund (2002a) defends the dispositional view (though Eklund is no Conceptual Engineer).

16. Happy-Face Conceptual Engineering

We can summarize the components of *Happy-Face Conceptual Engineering* as follows:

> *Component One*: *Isolate the basic culprit*. Give sufficient reason to think some basic premise or rule of inference or presupposition invoked in a paradox is false or invalid; or, give sufficient reason to "bite the bullet" and endorse the conclusion. Thus, the derivation of the unacceptable conclusion is blocked (or the conclusion turns out to be acceptable). If no more specific culprit is to blame then you have found the basic culprit.
>
> *Component Two*: *Explain contagiousness*. Mastery of the concepts deployed in the proof disposes us to accept all the conceptual principles (including the culprit) which feature as premises or rules of inference or presuppositions in the proof: our conceptual competence pulls us into the paradox.[18]
>
> *Component Three*: *Explain intractability*. Any (initially) promising solution to a paradox entails that we must give up on some particular conceptual principle deployed in the proof. Since our competence with the relevant concept strongly disposes us to accept such a principle then that makes all promising solutions hard to swallow—even when we have succeeded in isolating the culprit. Thus, the paradox is tricky to treat.
>
> *Component Four*: *Revise or replace*.[19] To prevent the paradox from returning we need to either suitably revise one of the concepts deployed in the proof so that this concept is no longer inconsistent.[20] Or, if engaging in conceptual replacement, we need to ensure that the surrogate concept does not give rise to a related paradox.[21]

[18] The terminology of "pull" is from Eklund (2002a).

[19] See Greenough (MSa) for an evaluation as to whether, and in what way, this really is an essential feature of Conceptual Engineering.

[20] See Richard (2018) for a view which permits concepts/meanings to evolve in a species-like way.

[21] Has Happy-Face Conceptual Engineering been embraced by contemporary Conceptual Engineers (who accept concepts)? As it turns out, no Conceptual Engineer (that I know) explicitly endorses the view. Eklund (2002a) seems to endorse the first three components of Happy-Face Conceptual Engineering, but not the fourth. On Eklund's view, inconsistent concepts are not in need of revision or replacement. Eklund thus does not subscribe to Conceptual Engineering. Eklund (2015, 2017) is much more amenable to Conceptual Engineering (for moral concepts) but not because such concepts are inconsistent. Spicer (2008) and Weiner (2009) take the concept of knowledge to be inconsistent, but neither recommend revision or replacement. Fassio and McKenna (2015), meanwhile, sponsor a mild kind of revisionism for the concept of knowledge. Scharp (2013), meanwhile, comes close to endorsing Happy-Face Conceptual Engineering, but he does not accept that mastery of a concept requires that a subject be (initially) disposed to accept the conceptual principles for that concept. That is because Scharp (2013) accepts the arguments given in Williamson (2006) that competence with a concept does not require that a subject be disposed to accept any candidate conceptual principles for that concept. So, Scharp, does not endorse the second and third components of Conceptual Engineering. Rather, Scharp thinks that a subject who is competent with some concept is merely *entitled to believe* the conceptual principles for that concept. (The notion of entitlement deployed is taken from Burge (1993).) Scharp's view has an immediate cost: it cannot straightforwardly explain the contagiousness and intractability properties of typical paradoxes via the notion of conceptual competence. One answer is to hold that philosophical paradoxes are meant to be formulated so as to apply to some typical or normal or idealized subject. Perhaps a case can be made that such a subject, if competent, is disposed to accept the conceptual principles for that concept. Such a fix would enable Scharp to embrace Happy-Face Conceptual Engineering.

Let's now turn to unhappy-face treatments and see if they are needed in addition to, or in place of, happy-face treatments.

17. The Indeterminate Concept View: The Non-Specific Version

Recall that unhappy-face treatments merely establish a kind of collective guilt attaching to the group of principles which, taken together, produce some paradox: not only can we not isolate a single, basic culprit, we cannot even exclude any principle at work in the paradox from suspicion of guilt. Why is this so? On what may be termed *The Indeterminate Concept View*, it is indeterminate which principle in the set of principles which gives rise to the paradox is false—and it is indeterminate which conceptual principles in this set are true. Given the standard view of indeterminacy, under which indeterminacy precludes knowledge, we cannot, as a matter of metaphysical necessity, know that some principle in the paradoxical proof is false; nor indeed can we isolate any true principles.[22]

18. The Indeterminate Concept View: The Gappy Version

What model of indeterminacy could make sense of the view that inconsistent concepts have conceptual principles which are indeterminate in truth-value? One immediate proposal is that all the conceptual principles for some inconsistent concept (used to derive a contradiction in some paradox) are neither true nor false. This provides an immediate explanation as to why we can't know which principle is false: propositions which lack truth-values cannot be known. Some paradoxes are such that they involve no guilty (false) conceptual principles, and they are such that they involve no innocent (true) conceptual principles either.

19. The Excess Baggage Objection

Truth-value gaps are not something that every Conceptual Engineer will be happy to take on board as an essential piece of kit from the outset—Conceptual Engineering was not supposed to be some kind of niche doctrine. Furthermore, one key motivation to introduce inconsistent concepts into a theory of paradox in the first place was that it enables us to preserve classical logic and classical semantics. On a gappy way of understanding indeterminate concepts, that attractive feature is lost. Call that *The Excess Baggage Objection*.

[22] Schiffer (1996: 330) notes that even omniscient beings cannot know which element of an inconsistent ("glitchy") concept is false.

20. The Overkill Objection

On the gappy version of the Indeterminate Concept View, all the conceptual principles deployed in some paradoxical proof are neither true nor false—and so not true. It is just this feature that blocks the paradox.[23] Yet, that's a kind of overkill because the untruth of just one of these principles would be enough to block the derivation. Call that *The Overkill Problem*.

21. The Symmetry Argument

At the root of the Indeterminate Concept View is some kind of symmetry argument: when a set of conceptual principles is used in some paradoxical derivation then we should treat these principles as relevantly symmetrical—they are all equally guilty, as it were, in the derivation of the contradiction. In the simple case, where an inconsistent concept has just two such conceptual principles, then the grounds for accepting one principle (somehow) cancel out the grounds for accepting the other. But that does not mean that both principles are false—just that these grounds are not sufficiently strong to make either principle true. So, both these principles are neither true nor false.

22. The Indeterminate Concept View: The Classical Version

As it turns out, such symmetry considerations need not threaten bivalence. An alternative model of indeterminacy allows that some propositions can be either true or false but nothing grounds the truth-value that they have.[24] This alternative version of the Indeterminate Concept View entails that inconsistent concepts will have at least one false conceptual principle. However, since indeterminacy precludes knowledge, we will never be able to find out just which one is false.[25] That goes some way to addressing the Excess Baggage Worry because bivalence may be retained on an Indeterminate Concept View. It also addresses the Overkill Problem because only one principle in the set of conceptual principles used in the paradoxical derivation is false (the rest are true).

[23] Schiffer is pretty quiet about this feature of his view. Perhaps because a further worry soon emerges: if all the conceptual principles for some inconsistent concept are neither true nor false, and so absolutely unknowable, then, to use his own words back at him, "I think we would have heard about it by now." On that score, Schiffer (1996) is very keen to show that contextualism about "knows" is committed to an implausible error theory whereby alert competent subjects fail to see that "knows" is context-sensitive. Given that worry, however, how come alert, competent subjects fail to see that the conceptual principles for some ordinary concept are indeterminate in truth-value? Ironically, Schiffer also seems committed to an implausible error theory. See Greenough and Kindermann (2017) for the idea that everybody has an error-theory of some sort.

[24] See Greenough (2008) which builds on Sorensen (2001).

[25] Eklund (2002b endorses a version of the Indeterminate Concept View when he says (of the liar paradox): "it is *likely* that it is indeterminate just where the liar reasoning goes wrong. But still, somewhere there is an untrue assumption or invalid step" (p. 323, my emphasis).

23. Unhappy-Face Conceptual Engineering via the Indeterminate Concept View

Unhappy-Face Conceptual Engineering (via the Indeterminate Concept View) can thus be summarized as follows:

> *Component One*: *Explain why a happy-face solution is not available*. A happy-face treatment is not available because all the conceptual principles used in the paradox are equi-culpable. As such, they are each indeterminate in truth-value: they are either gappy (on the truth-value gap version), or they are either true or false but it is indeterminate which (on the classical version). Either way, the derivation is blocked.
>
> *Component Two*: *Explain contagiousness*. Mastery of all the concepts deployed in the proof disposes us to accept all the conceptual principles which feature as premises or rules of inference or presuppositions in the proof. This explains why the paradox was so contagious from the outset: our very understanding of the words used in the proof pulls us to accept a set of incompatible propositions.
>
> *Component Three*: *Explain intractability*. The paradox is *absolutely* intractable because we have been looking for a happy-face solution when it is metaphysically impossible to find one. (See Component One.)
>
> *Component Four*: *Revise or replace*. To prevent the paradox from returning we need to either suitably revise our concepts so that the conceptual principles deployed in the proof no longer, when taken together, entail a contradiction. Or, if engaging in conceptual replacement, we need to ensure that the surrogate concepts do not themselves give rise to a related paradox.

24. Absolute versus Relative Intractability

Should a Conceptual Engineer allow for both Happy-Face Conceptual Engineering and Unhappy-Face Conceptual Engineering (via the Indeterminate Concept View)? One reason to do so would because there are two basic types of paradox: those where it is feasible to find a culprit; and those where it is not (metaphysically) possible to isolate any guilty or innocent principles. The former paradoxes exhibit *relative* intractability whereby what blocks the route to uncovering the culprit is some contingent feature of us, our language, our methods, our concepts, our conceptual competence, and so on. Resolving such paradoxes may require some Happy-Face Conceptual Engineering or some more descriptive resolution of paradox. The latter paradoxes are *absolutely* intractable. Are there any such paradoxes?

25. Paradoxes as Stress-Tests

Paradoxes are akin to stress-testing a complex machine—where the aim is to uncover faults in the design (rather than reveal manufacturing faults in the particular machine being tested). Such a test may reveal that, if sufficiently pressed, the machine malfunctions. That malfunction may be due to the design of a single component—the other components are simply enabling features of the malfunction rather than contributory causes of the defect.[26] In other stress-tests, it may make little sense to speak of a single, faulty component. If an internal combustion engine misfires at low revs that may be due to a confluence of factors involving several features of the design—there will be a kind of collective culpability at work. In such cases, there is no single (best) remedy, but rather various ways in which one or more features of the engine can be altered in order to address the problem.[27] This analogy suggests that we should be very open to the possibility of Unhappy-Face Conceptual Engineering.

26. The Master Argument

It's one thing to be open to the possibility of Unhappy-Face Conceptual Engineering, quite another to think that some, many, or even most of our most stubborn philosophical paradoxes require an unhappy-face treatment. The challenge here is that Unhappy-Face Conceptual Engineering is merely a fall-back approach—one to be adopted after a happy-face approach has been thoroughly exhausted. It turns out that advocates of the Indeterminate Concept View do indeed think that happy-face treatments have had their day with respect to most (and perhaps nearly all) of our most stubborn paradoxes. Their master argument goes something like this: we've looked long and hard for the culprits in our most stubborn paradoxes; we've not been able to find them; what best explains this is that such paradoxes are absolutely intractable—it is indeterminate just which principle is the culprit in some paradox.[28]

27. The Imperialism Objection

It's far too hasty to assume that all the main work has been done as regards finding some (suitably) specific treatment to some long-standing paradox. Perhaps philosophy is merely in its infancy (as I am inclined to think). New philosophical theories continue to spring up. Old theories are still getting reworked. To think that *right now*, and *only* right now, in the twenty-first century, are we sufficiently enlightened so as to recognize that various long-standing paradoxes require us to posit indeterminacy to make sense of their intractability is unduly imperialistic. Call that *The Imperialism Objection*.

[26] In another design of machine, that component may work perfectly well. Likewise, whether a concept is defective is application/environment dependent.

[27] More typically, attribution of fault will be a matter of degree.

[28] Schiffer sponsors just this kind of argument when he says: "That no classical philosophical problem, including the sorites, yet has a happy-face solution is attested to by the fact that we are still debating each one of them" (2003: chapter 5). And: "Philosophers have been debating the problem of free will for centuries, and they are still debating it, with philosophers lined up behind each of the solutions in logical space. If the problem of free will had a happy-face solution, I think we would have heard about it by now" (2004: 179). (I suspect that something like this argument also underlies Eklund's (2002b) advocation of the Indeterminate Concept View.)

28. The Indiscriminable Concept View

Furthermore, a complete taxonomy of paradoxes should countenance an additional species of paradox whereby while it is metaphysically possible to locate the culprit in some long-standing, stubborn paradox, it is simply not feasible to do so. Here the thought is that the conceptual principles at work in some (stubborn) paradox are indeed relevantly symmetrical—but only in the sense that we are unable to discriminate the false/invalid conceptual principle from the true/valid ones. That is, one of these principles is false alright but they are similar enough to each other in their role in our thought and understanding such that our limited powers of (conceptual) discrimination are unable to discern which principle is false. And so many, or indeed most, stubborn paradoxes are not absolutely intractable—an omniscient being, or perhaps even a superior being who is hard-wired differently from ourselves, would be able to uncover the culprit. Call that *The Indiscriminable Concept View*.[29] We can now summarize a third kind of Conceptual Engineering.

29. Unhappy-Face Conceptual Engineering via The Indiscriminable Concept View

Component One: *Explain why a happy-face solution is not available*. A happy-face treatment is not available because the conceptual principles used in the paradox are sufficiently symmetrical such that we are unable to discriminate the true conceptual principles deployed from the false one. Still, the paradox is blocked because we know that one of the conceptual principles deployed is false—it is just not feasible (for us) to work out which one.

Component Two: *Explain contagiousness*. Same as for the Indeterminate Concept View.

Component Three: *Explain intractability*. The paradox is intractable because we have been looking for a happy-face solution when it is not feasible to give one. (See Component One.)

Component Four: *Revise or replace*. To prevent the paradox from returning we need to either suitably revise our concepts so that the conceptual principles deployed in the proof no longer, when taken together, entail a contradiction. Or, if engaging in conceptual replacement, we need to ensure that the surrogate concepts do not themselves give rise to a related paradox.

30. The Imperialism Objection Again

Why think that our most stubborn paradoxes require treatment via the version of Conceptual Engineering just given? The Indiscriminable Concept View is also

[29] There is really a family of Indiscriminable Concept Views depending on just how feasible it is to discover the culprit. One prominent member of this family is Mysterianism, the view that creatures like us will never be able to find the culprit in some paradox (cf. McGinn 1993: 31).

motivated via (a version of) the master argument given above: we've looked long and hard for the culprits in our most stubborn paradoxes; we've not been able to find them; what best explains this is our limited powers of conceptual discrimination—these make it infeasible (for us) to uncover the culprit. But this argument also suffers from a form of The Imperialism Objection given above: To think that *right now*, and *only* right now, in the twenty-first century, have we achieved sufficient philosophical enlightenment to realize that it is not feasible for creatures like us to discover the culprit in some stubborn paradox is unduly imperialistic.

31. Which Form of Conceptual Engineering Wins Out?

Given the above discussion, does Happy-Face Conceptual Engineering win out? That's a bit too hasty. Instead, the Conceptual Engineer is better off acknowledging that they are not in a good enough position to say whether some long-standing, stubborn paradox calls for Happy-Face Conceptual Engineering or calls for Unhappy-Face Conceptual Engineering. In the meantime, they are free to propose both kinds of treatment. Only when they have collected various treatments of both types will an answer begin to emerge as to whether Happy-Face Conceptual Engineering is called for.

32. A Happy-Face Treatment?

As it turns out, the discussion so far presents the Conceptual Engineer with a false choice: choose Happy-Face or choose Unhappy-Face Conceptual Engineering (or choose to pursue both kinds of approach). Most treatments of paradox which purport to be happy-face treatments are not in fact happy-face at all. While these treatments promise to isolate a specific, basic culprit, they merely turn out to have isolated a group of culprits which cannot all be true. For example. There is a whole raft of responses to the liar paradox which reject Tarski's T-schema for truth, namely the schema: a sentence S is true if and only if *p* (where S says that *p*). That may initially seem like a good candidate for being a happy-face solution: a culprit has been identified and rejected; the paradox is thus blocked. Not so. Tarski's T-schema is a biconditional. You only need to reject one direction of the biconditional to block the relevant form of the liar paradox. Those solutions which reject the T-schema but do not tell us which direction of the T-schema fails cannot count as happy-face solutions because they have not put their finger on a single, basic culprit.[30] Rather, they have put their finger on two principles, at least one of which must be untrue.[31]

[30] Horwich (1990), Eklund (2002a), and Scharp (2013) all propose solutions in which the T-schema is rejected. However, they do not tell us which direction of the T-schema is to be rejected. And since classical logic is respected on each of these proposals then one (or both) of the directions of the T-schema is false/invalid.

[31] This is why Scharp (2007, 2013) is not really a Happy-Face Conceptual Engineer (for truth) otherwise there would be no reason, as he thinks, to replace concept of truth with two surrogate concepts *ascending truth* and *descending truth*. See Greenough (2017) for relevant discussion.

33. Happy-Face Treatments Represent a Limit Case

This is not an isolated case. Happy-face treatments (whether involving Conceptual Engineering or not) represent a kind of limit case. Absolute specificity is extremely hard to come by. Some candidate culprit Z will typically be entailed by (or theoretically motivated by) some conjunction of two or more principles (A & B & . . .) which are each weaker than Z. Philosophical treatments of paradox which reject Z will often not be specific enough, as they stand, to say which of these principles A, B, . . . , is to be rejected. The only cases where a happy-face solution is in prospect will be when there is a single culprit which is *basic*—where the guilty party is not grounded in, or theoretically motivated by, a conjunction of two or more principles which are relevantly more fundamental or explanatory. The devotees of happy-face solutions have yet to give us sufficient confidence that paradoxes always bottom out in a single, basic culprit, let alone give us the confidence to declare that it is always feasible to find one.

34. Unhappy-Face Treatments are also a Limit Case

Unhappy-face approaches, as mentioned above, are maximally unspecific. Such treatments are also rather uncommon. Some process of elimination will typically always take place whereby certain principles used in some paradoxical proof will (justifiably) not fall under suspicion. So, while unhappy-face treatments represent a perfectly achievable limit case, no self-proclaimed Unhappy-Face Conceptual Engineer really practices what they preach and offers up such treatments. Rather, a more specific set of allegedly culpable principles will typically be selected for suspicion of guilt.[32]

35. Intermediate Treatments

The discussion in the last two sections now suggests that there is a scale of intermediate treatments of paradox between the limit cases of Happy-Face and Unhappy-Face Conceptual Engineering. Such intermediate treatments are neither fully happy-face nor fully unhappy-face—neither thoroughly specific in isolating a single (basic) culprit, nor thoroughly unspecific in being content to simply apportion collective blame to all the principles at work in some paradox. Furthermore, as we shall see, absolute specificity is not required in order to successfully treat a paradox; rather, a treatment which is *specific enough* will do.

36. Fully Neutral Treatments: A First Pass

Once we allow for intermediate treatments which are non-specific, but nonetheless specific enough to effectively combat a paradox, we can also make room for treatments which are *fully neutral*. These are treatments which don't simply stay suitably silent (and therefore neutral) on certain theoretical questions posed by the paradox. These treatments reject some culprit at work in the proof *from a theory-neutral*

[32] For example, in his (1996), Schiffer does not put any logical concepts under suspicion of guilt in discussing what to say about skepticism.

perspective, a perspective which endeavours to be as neutral as possible on (relevant) points of theory. An exemplar paradox, together with a neutralist solution, will help get clearer on the proposal.

37. Sameness Skepticism

Let the Bad Case be a case where it appears to some subject that they have two legs; they believe that they have two legs; and yet their belief is false—because they are the victim of an evil genie who is deceiving them. In such a Bad Case, the subject fails to know that they have two legs. Let the Good Case be a case where it appears to the subject that they have two legs; they believe that they have two legs; and the subject is not being deceived by an evil genie—so their belief is true and would ordinarily be taken to be knowledge.[33] The *Sameness Skeptic* says that the Good case and the Bad Case are relevantly the same when it comes to knowing: the subject is no better off (and no worse off), with respect to knowing, in the Good Case than they are in the Bad Case. The key skeptical thought here is that: all the subject has to go on, when it comes to knowing, is the evidence of their senses—how things appear—and such appearances are the same in both cases. However, since the subject fails to know in the Bad Case, and since the subject is no better off (and no worse off), with respect to knowing, in the Good Case, then they also fail to know in the Good Case. Upshot: the subject in the Good Case cannot know that they have two legs.[34] This form of skepticism might seem to be a thoroughly troublesome challenge—indeed a paradox because from initially plausible premises, via initially plausible reasoning, we have derived an initially implausible conclusion. Even so, when we properly regiment the symptoms of Sameness Skepticism, it becomes rather easy to see *that* the paradoxical proof fails, but not so easy to see *where* the proof fails.[35]

38. Sameness Skepticism Regimented

Suitably regimented, the symptoms of *Sameness Skepticism* are:

(1) *Ex hypothesi*, the Good Case and the Bad Case are phenomenally alike (with respect to the proposition that p): it appears to the subject, in both Good Case and Bad Case, that p.[36]

[33] The terminology is taken from Williamson (2000).

[34] Sameness Skepticism is not a form of Cartesian Skepticism because it doesn't invoke the claim that the Good Case is (phenomenally) indiscriminable from the Bad Case. (Nor does it invoke any kind of closure principle for evidence or knowledge.) It is much closer to what has come to be known as *Underdetermination Skepticism* (see Brueckner 1994; Vogel 2004; Pritchard 2005).

[35] The common mistake when presenting skeptical arguments is to specify them in compressed form—as if that somehow captures the essence of what is going on. That's bad symptomatology. Rather, we should specify them so that every stage of the disease is properly on display. In the case of Sameness Skepticism, once we do so a Neutralist treatment comes into view. Once in view, there is simply no call for some exotic treatment or curious cure; no need for Contextualism; and certainly no need to succumb (as some philosophers do) to Skepticism.

[36] More generally, let the subject in the Good Case be a phenomenal duplicate of the subject in the Bad Case.

(2) If the Good Case and the Bad Case are phenomenally alike then they are evidentially alike (with respect to *p*): the evidence had by the subject in the Good Case for *p* is the same evidence the subject has in the Bad Case for *p*. (After all, the thought goes, the only evidence we have to go on is the evidence from our senses—from how things appear to us via looking, tasting, smelling, and so forth—and matters appear the same in both cases.)

(3) If the subjects in both Good Case and Bad Case are evidentially alike (with respect to *p*) then they are alike with respect to knowing that *p*: the strength of the subject's position with respect to knowing that *p* is the same in both Good Case and Bad Case. (After all, the thought goes, our evidence is what determines how good our position is with respect to knowing.)

(4) If the strength of the subject's position (with respect to knowing that *p*) is the same in both Good Case and Bad Case then the subject in the Good Case knows that *p* if and only if the subject in the Bad Case knows that *p*. (How good our position is with respect to knowing determines whether or not we know.)

(5) *Ex hypothesi*, the subject cannot know, in the Bad Case, that *p*.

(6) Therefore, given (1) to (4), the subject cannot know, in the Good Case, that *p*.

Premises (1) and (5) are just part of the set-up; the conclusion (6) is highly implausible; and, (2) to (4) unpack the initially plausible sounding claim that "all the subject has to go on, when it comes to knowing, is the evidence of their senses—how things appear". Paradox!

39. A Neutralist Treatment of Sameness Skepticism

The conjunction of premise (2) and (3) entails the following skeptical claim SC:

> (SC) If the Good Case and the Bad Case are phenomenally alike (with respect to *p*) then the strength of the subject's position with respect to knowing that *p* is the same in both Good Case and Bad Case.

But SC is just false—and everybody can agree on that. It is part of our ordinary conception of knowledge that phenomenal alikeness does not entail that two subjects are in the same position with respect to knowing that *p*. A subject with a true belief that *p* is in a better, or at least different, position with respect to knowing that *p* than a subject who has a false belief that *p*. That's because having a true belief that *p* is a necessary condition on knowing that *p*. Meeting that condition puts you in a better (or at least different) position with respect to knowing that *p* than someone who has a false belief that *p*.[37]

What's crucial here is that a failure of SC is neutral between competing theories of knowledge. This means that the paradoxical derivation fails either at step (2), or at step (3).[38] However, it is a far from straightforward matter to see *which* of these two

[37] Note that (4) is not under dispute.

[38] Rejecting both would be overkill. See The Overkill Objection above.

principles fails. That's because we need to deploy much more controversial, specific philosophical theory to blame (2) but not (3).[39] Likewise, we need to deploy much more controversial, specific theory to blame (3) but not (2).[40] Fortunately, we don't need to be so specific in order to effectively treat a paradox. Effective treatments just need to be *specific enough*. Additionally, we don't need to adopt a controversial theory of knowledge to address Sameness Skepticism—we can reject SC from a theoretically neutral, non-controversial, position. Let me now try to bolster these latter claims.

40. Back to the Engineering Metaphor

Recall the engineering metaphor invoked above. An engineer will typically be able to bring to bear sufficient theory to isolate that a design defect is present, say, in the ignition system. Let's say it is this defect that is causing the engine to misfire at low revs. On that basis, they may be able to fix the fault via some fairly non-specific theorizing, but without having a sufficiently specific theory to say exactly what it is about the design of the ignition system that is causing the defect to emerge. That still represents a perfectly respectable, suitably specific resolution of the trouble—because they have enough theory to fix the fault. It would be entirely misplaced to say: "Wait! We *always* need to find out *exactly* what is causing the fault." That is a demand too far. Likewise, we can defeat the Sameness Skeptic but without being able to say which of (2) and (3) is invalid. It would also misplaced to say: "Wait! We cannot defeat the Sameness Skeptic until we know just which of (2) and (3) is false." We have sufficient understanding of knowledge to provide a dialectically satisfying resolution of the paradox.

41. The Primary Goal of Treating Paradox

This suggests there is a primary and secondary goal at work in resolving paradoxes. When the engineer is faced with a faulty ignition system, her primary goal is to fix the fault. When the doctor is faced with a disease, her primary goal is to cure the patient. When the philosopher is faced with a paradox, her primary goal is to prevent the paradox from taking hold (and, if it has taken hold, to release the grip that the paradox has upon us). To do that, in each case, some theory is needed. The mechanical engineer uses fluid dynamics, metallurgy, and more; the doctor uses human biology, biochemistry, pharmacology, and more; the philosopher uses the theory of knowledge, the theory of truth, and more. As we have just seen, this theory need not be that specific (or that deep)—it just needs to be specific enough (and deep enough). It is not part of the primary dialectical goal to have the last word, or even a very specific word, on the concept of, for example, knowledge.

[39] Williamson (2000: chapter 8), for example, blames the phenomenal conception of evidence.

[40] Those who accept the phenomenal conception of evidence will take just this route.

42. The Secondary Goal of Treating Paradox

The secondary goal of treating some paradox is to improve our philosophical theory—to give insights into the deeper nature of, for example, knowledge by answering all of the theoretical questions posed by the paradox. In particular, we want to know whether or not all the premises of some paradox are true (and why these premises have the truth-value that they have). Recall Russell's remark above that philosophical puzzles serve the same role that experiments serve in science. Equally, recall also that above we conceived of paradoxes as yielding theoretical stress-tests. In meeting the primary goal we act like a doctor (or engineer); in meeting the secondary goal, we act like a human biologist (or metallurgist or chemist or physicist).

43. Don't Conflate the Primary and Secondary Goals

The primary and secondary goals may often march in step, but they may come apart. In giving a neutralist treatment of some paradox, the secondary goal will typically not be fully met. In particular, we will not be able to say whether or not all the premises of the paradox are true—that's just what happened with respect to premises (2) and (3) of Sameness Skepticism. An effective treatment of this paradox need not take a stand on the phenomenal conception of evidence. We should not conflate the Primary and Secondary goals of treating paradox: if we demand that all theoretical questions posed by some paradox be answered then that sets an unreasonably high bar for a treatment to be effective.[41]

44. Three Axes of Neutralism

It's worth stressing that there is more to a neutralist treatment than being non-specific but specific enough. There are two more axes of neutralism at work in the treatment of Sameness Skepticism. Suppose that a paradox involves a principle A and a principle B, which, taken together entail some principle Z. Suppose that a treatment involves giving an independent reason to reject Z. Given this, there are three axes of Neutralism to consider. Firstly, we have:

> *Axis One*: Be neutral as to which of A, B, is false (where A, B are both weaker than Z).[42]

This Axis will be an essential feature of typical neutralist treatments. Secondly, we have:

[41] Another way of thinking about the distinction in hand is via two different sorts of opponents: the paradox-peddler—the skeptic, the misologist, the absurdist, the irrationalist, the sophist, the pyrrhonist, the gadfly—versus the bemused theorist who is beset by paradox. The primary goal is to defeat the paradox peddler via some specific or non-specific treatment; the secondary goal is to provide the bemused theorist with better, more complete theory.

[42] On one refinement of Neutralism, principles A and B should both not only be epistemically possible for the subject who is endeavouring to resolve the paradox, but these principles must each have some strong (*prima facie*) evidence. See Greenough (MSb) for such refinements. (Thanks to Tim Sundell here.)

> *Axis Two*: Reject Z using neutral theory.

Not all non-specific treatments will invoke this Axis because on some non-specific treatments it may be a controversial matter that Z fails. Those remedies that do invoke this axis have the attractive feature that the non-specific treatment on offer issues from a perspective which is available to all (sensible) theorists concerning knowledge and evidence. Finally, we also have:

> *Axis Three*: Stay neutral as to the following options: (i) one and only one of A, B is false, and it is feasible to find out which; (ii) one and only one of A, B is false, and it is not feasible to find out which; (iii) it is indeterminate whether A/B is false and so it is metaphysically impossible to find out which.

Axis Three in effect entails that Neutralism stays silent on the issue as to whether or not a more specific treatment is feasible: the jury is out on whether we can move beyond neutral treatments to our central philosophical paradoxes to more specific treatments. (And so we remain neutral on the veracity of the Indeterminate and Indiscriminable Concepts views.)

45. Neutralism and Intractability

Why have our central paradoxes proved so hard to treat via (relatively) specific treatments? Because of the third axis, Neutralism cannot co-opt the accounts of intractability given by the Indeterminate/Indiscriminable Concept Views. Those accounts make it impossible/infeasible, respectively, to discover specific treatments—but Neutralism is neutral on that issue. What about the account of intractability offered by Happy-Face Conceptual Engineers? That is certainly available. So, Sameness Skepticism has proved to be tricky to treat because our conceptual competence with the concept of evidence and the concept of knowledge seduces us into accepting premises (2) and (3). Any promising solution which rejects one of these premises just (initially) feels wrong because of such competence. A broader question is simply: why have our central paradoxes proved so hard to treat *simpliciter*? The neutralist answer is (in part) that we simply have overlooked the possibility of neutral treatments—treatments which have all three axes of neutrality. We have been too focused on specific, controversial treatments and have overlooked the possibility of non-specific, theory-neutral treatments.[43]

46. Neutralism and Minimal Adequacy

Neutralism allows for a modest kind of pluralism because neutral and non-neutral treatments can happily co-exist. There may be some more specific, controversial theory of knowledge which improves upon a neutralist treatment by better satisfying both the primary and secondary goals of an effective treatment. Not all non-neutralist proposals should be taken seriously however. Go back to Sameness Skepticism. Neutralism rejects SC, where SC follows from the conjunction of

[43] Relatedly, we have conflated the primary and secondary goals of treating paradox.

premises (2) and (3). Only those specific responses which both reject SC and go on to reject either (2) or reject (3) should be taken seriously. Should some non-neutralist treatment fail to entail that SC is to be rejected then we can dismiss its credentials from the outset. For this reason, neutralist treatments serve as a kind of minimal adequacy condition on any more specific, substantive remedy.[44]

47. Neutralist Conceptual Engineering with Concepts

Component One: *Isolate the non-specific culprit from a neutral perspective*. Give sufficient reason, from a theory-neutral perspective, to think that some (non-basic) premise or rule of inference or presupposition invoked in a proof is false or invalid; or, give sufficient reason to "bite the bullet" and endorse the conclusion. Here the culprit will be non-specific in the sense that it is entailed by the conjunction of two or more conceptual principles (each weaker than the non-specific culprit) and where the treatment is neutral as to which of these principles is false. Thus, the derivation is blocked (or the conclusion turns out to be acceptable after all).

Component Two: *Explain contagiousness*. As above, our very understanding of the concepts used in the proof pulls us to accept a set of incompatible propositions.

Component Three: *Explain intractability*. The explanation is two-fold: as above, our conceptual competence makes any proposed solution difficult to swallow; but also we have hitherto overlooked the possibility of neutralist solutions—and so we have, in the first instance at least, been looking for a treatment in the wrong place.

Component Four: *Revise or replace*. To prevent the paradox from returning we need to either suitably revise our concepts so that the conceptual principles deployed in the proof no longer, when taken together, entail a contradiction. Or, if engaging in conceptual replacement, we need to ensure that the surrogate concepts do not themselves give rise to a related paradox.

48. Meaning Engineering

There is little agreement as to what concepts are, where they live, how they survive, and what role they play in a theory of meaning and understanding. Arguably, the concept of a concept is not in great shape.[45] Ironically, it is in desperate need of some Conceptual Engineering.[46] That gives some initial reason to think that Conceptual Engineering without concepts is the more promising view.[47] Such a view aims to solve philosophical problems by revising the meanings—the intensions—of our words. Call that view *Meaning Engineering*.[48]

[44] Neutral treatments are akin to Tarski's minimal adequacy condition on any substantial theory of truth. For discussion of Neutralism in relation to philosophical progress, see Greenough (MSb).

[45] Worse shape than, for example, the notion of intension.

[46] See, for example, Machery (2009) for a version of Concept Eliminativism.

[47] One further prominent reason to reject concepts stems from Williamson (2006) who argues that there are no conceptual truth/principles, and so no concepts.

[48] See Cappelen (2018) for this kind of view.

49. Inconsistent Meanings (Intensions)

Without concepts, there are no inconsistent concepts. How does the Meaning Engineer make sense of the source of paradox? One surrogate for inconsistent concepts are inconsistent meanings (or intensions). These are like inconsistent concepts in that they are composed of principles—intensional principles—which cannot all be true. These principles play an extension-determining role: they fix (or partially fix) the extension of the relevant word (relative to some world). Such inconsistent meanings are rather exotic entities. How do they (partially) fix an extension? Is an idealized user disposed to accept them? I don't propose to answer these questions here. That would take us too far afield. Instead, let's look at an alternative view.

50. Inconsistent Words

One alternative surrogate for inconsistent concepts are *inconsistent words*. Roughly, inconsistent words have uses which are in conflict. Indeed, these uses compete to become the privileged use which manages to confer a consistent meaning onto the word. (In order to make sense of paradoxes, this conflict had better be intra-personal in order to account for why a single subject can be drawn into a paradox.) Take Sameness Skepticism. A competent subject uses the words "evidence" and "knows" such that they are *initially* disposed to accept both (2) and (3). It's easy to develop an account of contagiousness and intractability from there.

51. Neutralist Meaning Engineering

To make Neutralism available to the Meaning Engineer we need to replace talk of inconsistent concepts with inconsistent words in the account of Neutralist Conceptual Engineering. Component Four then becomes: To prevent the paradox from returning we need to suitably revise or replace the inconsistent words deployed in the paradox. There are (at least) two ways in which a revisionary form of Neutralist Meaning Engineering can proceed.

52. Type I Neutralist Meaning Engineering

Suppose we have some paradox, involving some term T, which consists of two initially plausible premises A, B, and an initially implausible conclusion Z which follows from A & B via initially plausible reasoning. To simplify matters, suppose that rejecting logic is not on the table for this paradox. The broad options for an effective treatment are: reject Z (and so reject the conjunction A & B), or accept A & B—and thus bite the bullet and accept Z. Suppose further that our current use of T is unable to break the (relevant) symmetry between these two options—we are not able to currently work out which of the two options is the correct one. The Semantic Engineer now suggests that, by suitably revising the use and meaning of the term T, we can break this symmetry such that it is obvious, after this revision, to a competent subject that the conjunction A & B is to be rejected.[49] Thus the paradox is blocked and we can happily comply with our initial disposition (before the revision) to reject Z. Moreover, if this revision is successful then it becomes common ground amongst all competent users of T that the conjunction A & B is to be rejected. In other words, this conjunction is rejected from a suitably neutral theoretical standpoint. It may well be, however, that the semantic revision of T is not sufficiently fine-grained to tell us which of the two premises A, B, is to be rejected. So, after the revision, we must remain neutral on this issue.[50] Furthermore, we also remain neutral on the issue as to whether or not a more specific revision is available which enables us to readily work out which of these two premises is false. All three axes of Neutralism are thus satisfied. This kind of Neutralist Semantic Engineering uses semantic revision to take us from a paradox to a fully neutralist treatment.[51]

53. Type II Neutralist Meaning Engineering

Suppose we have a paradox with the same structure as that just given—except that prior to any revision of the term T we *are* able to reject the conjunction A & B from a theoretically neutral standpoint—thus blocking the paradox. Suppose also that our (descriptive) treatment is unable to tell us which of the two premises A, B, is the false premise. Furthermore, suppose, we remain neutral as to just *why* this is so. Consequently, the (descriptive) treatment is neutralist on all three axes of Neutralism. The revisionary Semantic Engineer then proposes that our use and meaning of the term T is revised such that either we are disposed to accept (2) but not (3), or we are disposed to accept (3) but not (2). So, after such a revision, one of these premises will no longer yield any kind of initial pull on the competent subject. Before such a revision, there is no obvious basic culprit to blame.[52] After the revision, we can easily identify a more basic culprit. This kind of Neutralist Meaning Engineering uses semantic revision to go from a (descriptive) neutralist treatment to a treatment which, on one axis of Neutralism at least, is less neutral, more specific, and more happy-face in character.

[49] To be properly engaged in Meaning Engineering we must simultaneously engage in revising the use *and* meaning of a word. In particular, we must revise those uses which are meaning-determining (and not just revise any old use of the word). See Greenough (MSa) for relevant discussion.

[50] The relevant symmetry of the two options may arise because of indeterminacy: before the semantic revision it is indeterminate, and so unknown, which of the two options is correct; but after the revision it is known (and so it is not indeterminate) that A & B is to be rejected. Or, the relevant symmetry may arise because of indiscriminability: before the semantic revision, the conjunction A & B is true, say, but it is not feasible to find this out; after the revision, A & B turns out to be not only false, say, but obviously so.

[51] If the revision of the use and meaning of T is specific enough to yield knowledge as to whether or not A/B is false then all competent subjects can readily reject one of these premises from a theory-neutral perspective. Would this yield a kind of hybrid between a Happy-Face and neutralist treatment? Not necessarily. As we saw above Happy-Face treatments may well be an unachievable limit case.

[52] Either because of indeterminacy or because of our limited powers of discrimination.

54. Concluding Remarks

The broad goal of this chapter was to show that Conceptual Engineering approaches to Paradox are not limited to (weak) Unhappy-Face treatments (via the Indeterminate Concept View). Conceptual Engineers can also pursue: happy-face treatments; unhappy-face treatments (via The Indiscriminable Concept View); treatments which are intermediate between the limit cases of Happy-Face and Unhappy-Face Conceptual Engineering; and treatments which involve Neutralism—the broad view that philosophical progress can take place when (and sometimes only when) a thoroughly neutral, non-specific theory, treatment, or method is adopted. In this last case, we found that Neutralism can be combined with Conceptual Engineering with or without concepts. Finally, it is noteworthy that Neutralism is a natural consequence of Semantic Engineering—after all, when it comes to treating paradox, the revision of the use and meaning of some term deployed in the paradox surely aims to produce consensus as to what premises are to be accepted and what are to be rejected. That consensus represents a kind of neutral standpoint. Conceptual Engineering naturally leads to Neutralism—though not vice versa.

Acknowledgements

Particular thanks to Alexis Burgess, Herman Cappelen, and David Plunkett for very helpful editorial and philosophical feedback. For equally useful feedback, thanks also to Derek Ball, Dragana Bozin, Josh Dever, Zoe Drayson, Matti Eklund, Paul Horwich, Mirela Fus, Chris Kelp, Anna-Sara Malmgren, Andrew Peet, Gurpreet Rattan, Mark Richard, Sven Rosenkranz, Kevin Scharp, Camilla Serck-Hanssen, Mona Simion, Tim Sundell, Amie Thomasson, Natalia Waights-Hickman, and Elia Zardini. Thanks also to audiences at: Aberdeen, ANU, Bled, Bristol, Glasgow, Graz, Leuven, Oslo, St Andrews, and UWA. Special thanks to Mona Simion and Tim Sundell for their replies at the Zermatt ConceptLab Workshop in February 2018. The work on this chapter has received funding from the project FFI2013-45968-P, financed by the Spanish Ministry of Economy and Competition (MINECO). Many thanks also to ConceptLab, UiO, for hosting me as a research fellow for two months in 2017 when this chapter was written.

References

Brueckner, A. 1994. The Structure of the Skeptical Argument. *Philosophy and Phenomenological Research* 54:827–35.

Burge, T. 1993. Content Preservation. *The Philosophical Review* 102 (4):457–88.

Burgess, A. 2014. Keeping ‘True’: A Case Study in Conceptual Ethics. *Inquiry* 57 (5–6):580–606.

Burgess, A., and Plunkett, D. 2013a. Conceptual Ethics I. *Philosophy Compass* 8:1091–101.

Burgess, A., and Plunkett, D. 2013b. Conceptual Ethics II. *Philosophy Compass* 8:1102–10.

Cappelen, H. 2018. *Fixing Language: Conceptual Engineering and the Limits of Revision.* Oxford: Oxford University Press.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Chihara, C. 1979. The Semantic Paradoxes: A Diagnostic Investigation. *Philosophical Review* 88 (4):590–618.

Cook, R. 2013. *Paradoxes*. Cambridge: Polity Press.

Cuonzo, M. 2014. *Paradox*. Cambridge, MA: MIT Press.

Deleuze, G. and Guattari, F. 1991/1994. *What is Philosophy?* trans. H. Tomlinson and G. Burchell. New York: Columbia University Press.

Díaz-León, E. 2017. Epistemic Contextualism and Conceptual Ethics. In J. Jenkins-Ichikawa (ed.), *Routledge Handbook for Epistemic Contextualism* (pp. 71–80). London: Routledge.

Eklund, M. 2002a. Inconsistent Languages. *Philosophy and Phenomenological Research* 64:251–75.

Eklund, M. 2002b. Personal Identity and Conceptual Incoherence. Nous 36 (3):465–85.

Eklund, M. 2015. Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In C. Daly (ed.), *Palgrave Handbook of Philosophical Methods* (pp. 363–85). London: Palgrave Macmillan.

Eklund, M. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.

Fassio, D., and McKenna, R. 2015. Revisionary Epistemology. *Inquiry* 58 (7–8):755–79.

Greenough, P. 2002. Knowledge, Lies, and Vagueness: A Minimalist Treatment. PhD Thesis, St Andrews.

Greenough, P. 2003. Vagueness: A Minimal Theory. *Mind* 112:235–81.

Greenough, P. 2008. Indeterminate Truth. In P. French (ed.), *Truth and Its Deformities, Midwest Studies in Philosophy*, vol. XXXII, pp. 213–41. New Jersey: Wiley-Blackwell.

Greenough, P., and Kindermann, D. 2017. The Semantic Error Problem for Epistemic Contextualism. In J. Jenkins-Ichikawa (ed.), *Routledge Handbook for Epistemic Contextualism*. London: Routledge.

Greenough, P. 2017. Conceptual Marxism and Truth. *Inquiry*. DOI: 10.1080/0020174X.2017.1287919

Greenough, P. 2019. Neutralism and the Observational Sorites Paradox. *Synthese*.

Greenough, P. MSa. Against Conceptual Engineering. Book ms.

Greenough, P. MSb. Knowledge: In Sickness and in Health. Book ms.

Haslanger, S. 2000. Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs* 34 (1):31–55.

Haslanger, S. 2005. What Are We Talking About? The Semantics and Politics of Social Kinds. *Hypatia* 20 (4):10–26.

Haslanger, S. 2006. Philosophical Analysis and Social Kinds: What Good are Our Intuitions? *Proceedings of the Aristotelian Society*, Supplementary Vol. 80:89–118.

Haslanger, S. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.

Horwich, P. 1990. *Truth*. Oxford: Blackwells.

Leslie, S. J. 2017. The Original Sin of Cognition: Fear, Prejudice and Generalization. *The Journal of Philosophy* 114:1–29.

Machery, E. 2009. *Doing Without Concepts*. Oxford: Oxford University Press.

Mackie, John L. 1972. *Truth, Probability, and Paradox: Studies in Philosophical Logic*. Oxford: Clarendon Press.

McGinn, C. 1993. *Problems in Philosophy: The Limits of Inquiry*. Oxford: Blackwells.

Plunkett, D., and Sundell, T. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Plunkett, D. 2015. Which Concepts Should we use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 59 (6):793–818.

Plunkett, D. 2016. Conceptual History, Conceptual Ethics, and The Aims of Inquiry: A Framework for Thinking About The Relevance of the History/Genealogy of Concepts to Normative Inquiry. *Ergo* 3 (2):27–64.

Pritchard, D. 2005. The Structure of Sceptical Arguments. *The Philosophical Quarterly* 55:37–52.

Richard, M. 2018. *Meanings as Species* Oxford: Oxford University Press.

Russell, B. 1905. On Denoting. *Mind* 14 (56):479–93.

Sainsbury, R. M. 2009. *Paradoxes*. Cambridge: Cambridge University Press.

Scharp, K. 2007. Replacing Truth. *Inquiry* 50 (6):606–21.

Scharp, K. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Schiffer, S. 1996. Contextualist Solutions to Scepticism. *Proceedings of the Aristotelian Society New Series* 96:317–33.

Schiffer, S. 2003. *The Things We Mean*. Oxford: Clarendon Press.

Schiffer, S. 2004. Skepticism and the Vagaries of Justified Belief. *Philosophical Studies* 119 (1–2):161–84.

Sider, T. 2014. Hirsch's Attack on Ontologese. *Nous* 48:565–72.

Simion, M. 2017. The 'Should' in Conceptual Engineering. *Inquiry* 61 (23):1–15.

Sorensen, R. 2001. *Vagueness and Contradiction*. Oxford: Oxford University Press.

Sorensen, R. 2003. *A Brief History of the Paradox*. New York: Oxford University Press.

Spicer, F. 2008. Are There Any Conceptual Truths about Knowledge? *Proceedings of the Aristotelian Society* 108:43–60.

Tarski, A. 1944. The Semantic Conception of Truth. *Philosophy and Phenomenological Research* 4:341–76.

Thomassen, A. L. 2016. What Can We Do, When We do Metaphysics? In G. D'Oro and S. Overgaard (eds.), *Cambridge Companion to Philosophical Methodology* (pp. 101–21). Cambridge: Cambridge University Press.

Vogel, J. 2004. Skeptical Arguments. *Philosophical Issues* 14:426—55.

Weiner, M. 2009. The Mostly Harmless Inconsistency of Knowledge Attributions. *Philosophers' Imprint* 9:1–25.

Williamson, T. 2000. *Knowledge and Its Limits*. New York: Oxford University Press.

Williamson, T. 2006. Conceptual Truth. *Aristotelian Society Supplementary* 80:1–41.

12

Going On, Not in the Same Way

Sally Haslanger

1. Introduction

I don't know what concepts are, or even if there are any (Machery 2009). So it feels awkward to set out to write a chapter specifically on conceptual engineering. The fact is, I've always been much less interested in what our terms mean (or the content of our concepts) than in what in the world is worth talking about (Haslanger 2012: chapter 16), though of course, these two issues are related. I suppose I have myself to blame, however, for I have suggested more than once that valuable projects within philosophy can be *ameliorative*, more specifically, that we should seek not only to elucidate the concepts we have, but aim to improve them in light of our legitimate purposes (Haslanger 2000/2012: chapter 6; see also Haslanger 2017). More specifically, I have argued for ameliorative accounts of *gender*, *man*, *woman*, and *race*; or, in the language of concepts, for ameliorated concepts of *gender*, etc.

Because my claims about amelioration have been sketchy and confusing, and have shifted over time, I begin my discussion with a recap of and reflection on my earlier projects with the aim of situating this chapter in the context of what came before. I will then consider how and why, within an externalist semantics, we might understand the project of improving our concepts in ways that will promote greater justice. I will argue that at least certain concepts should be understood as playing a functional role in enabling us to coordinate and to organize our lives together. As background conditions and assumptions evolve, so do the contents of the concepts. This evolution is mostly not conscious or planned, and is rarely under our control. But in some cases we can demonstrate ways in which our conceptual resources are inadequate and undertake to improve them. This happens in law all the time (consider the evolving legal definitions of 'rape' and consequent changes in sex education and contestation over practices of consent); it also occurs in social movements—both in counter-publics and subaltern communities—and in fascist propaganda.

The idea is not that all we need to do in order to change the world is change our minds. Of course not. But the tools that culture provides us—such as language, concepts, and inferential patterns—provide the frame for coordination and shape our interaction. Contestation over language and meaning is not always "mere semantics" for it shapes our agency and our lives together. Sometimes we should (at least try to) take control over meanings, for if we don't, others will.

Sally Haslanger, *Going On, Not in the Same Way* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Sally Haslanger.
DOI: 10.1093/oso/9780198801856.003.0012

2. Historical and Political Context

When I wrote my paper, “Gender and Race: (What) Are They? (What) Do We Want Them To Be?” (published in 2000, but written in the mid-1990s), I was not thinking much about background philosophical work on concepts, and didn’t have the language to express clearly what I now think is plausible. Although I was aiming to challenge traditional conceptual analysis by suggesting that our concepts might be improved by empirical or pragmatic considerations, I left the arguments schematic and unclear. Let me begin by trying to situate my project within the literature I was addressing.

The 1990s was a moment when feminists were questioning whether and, if so, how we could continue to talk about women; some even suggested that we were entering a post-feminist era (Riley 1988; cf. Butler 1990). At that time, one dominant conception of gender assumed gender to be primarily a matter of *identity.* The idea of ‘gender identity,’ however, took many forms. For example, one approach took gender to be (roughly) a set of psychological predispositions constructed in a process of socializing males and females, with the result that gendered girls and women were those who developed a relational (vs. atomistic) or emotional (vs. rational) mode of being in relation to others and to the world (e.g., Chodorow 1978; Gilligan 1982). Those of us who did not fit this stereotype, or who recognized the diversity of women’s psychological orientations, found this unacceptable.

Those attending to the diversity of gender formations began to think that the effort to define gender was fruitless:

> If one “is” a woman, that is surely not all one is; the term fails to be exhaustive, not because a pregendered “person” transcends the specific paraphernalia of its gender, but because gender is not always constituted coherently or consistently in different historical contexts, and because gender intersects with racial, class, ethnic, sexual, and regional modalities of discursively constituted identities. As a result, it becomes impossible to separate out “gender” from the political and cultural intersections in which it is invariably produced and maintained.
>
> (Butler 1990: 3)[1]

A deeper problem is that a commitment to the social construction of gender “troubles” the political task of feminism:

> Juridical power inevitably “produces” what it claims merely to represent; hence, politics must be concerned with this dual function of power: the juridical and the productive. In effect, the law produces and then conceals the notion of “a subject before the law” in order to invoke that discursive formation as a naturalized foundational premise that subsequently legitimates that law’s own regulatory hegemony. It is not enough to inquire into how women might become more fully represented in language and politics. Feminist critique ought also to understand how the category of “women,” the subject of feminism, is produced and restrained by the very structures of power through which emancipation is sought. (Butler 1990: 2)

[1] There is a huge literature on this issue. See, for example, Mohanty (1984); Harris (1990); Crenshaw (1991).

In effect, gender is constructed and then assumed to be natural or given. Women, as constructed, are the political subjects in whose name feminism speaks.[2] But if gendered constructions are part of the problem, then feminism's effort to support *women* functions as a way to buttress and sustain the construction of gender. It might seem that an alternative would be to support prediscursive agents against the forces of gender. However, many feminists rejected this alternative: following the reasoning of the quote above, *any* notion of the subject is a juridical construction, so there are no *prediscursive* agents. Even if there were, they would not be women, and so not feminism's proper concern. I reject both of these arguments. Of course, I don't think there are humans who exist outside of culture, but the idea of a social/political agent—used as an analytical tool—is as much a resource for oppositional movements as it is for the dominant juridical structure and can be reappropriated to promote justice; and feminists can legitimately speak on behalf of all agents (not just women) against the binary construction of men and women.

At the time, I was interested in the idea that a critical theory should be *emancipatory.* But, as Butler notes, how could we emancipate ourselves from the structures that construct us as political subjects? To respond, it is helpful to draw on the tradition of critical theory (Geuss 1981). Start with the idea that we all participate in social structures and enact social practices that are unjust, but most of the time this is not obvious to us, even when we are the ones disadvantaged. Many of us get up in the morning and do our best to get our kids to school and ourselves to work on time, not thinking about the racialized school to prison pipeline, the exploitation of janitors who have cleaned our offices overnight, or our own enactment of the gendered division of labor. In much of ordinary life, our practices and our identities present themselves to us in ways that mask the broader system.

To explain our unthinking participation in unjust and oppressive structures, it is useful to have the concept of *ideology.* The concept of ideology is employed in different ways within different traditions. On my view, a *cultural technē* is a set of social meanings—including concepts, scripts, background assumptions ("analytic" truths), inferential patterns, salient metaphors, metonyms, conceptual oppositions, and (broadly speaking) grammar—that provides tools for interpreting and responding to each other and the world around us, and does so in ways that facilitate (better or worse) forms of coordination.[3] (See also Balkin 1998: 102.) The cultural technē provides the frames of our (fragmented, dynamic) practical orientation—or "practical consciousness" (Marx 1845; Giddens 1984: xxiii)—that enables us to engage in social life. An ideology is a cultural technē gone wrong: it may fail to provide us the tools to appreciate relevant parts of the world, or what's valuable and how things are valuable; it may organize us in unjust ways. When we are in the grip of an ideology, however, our practical orientation positions us to enact—usually unknowingly and routinely—practices and structures that sustain injustice (Althusser 1970; Hall 1996/2006: 24–5).

[2] The claim is not that embodied individuals are created by culture, but that *political subjects* are. Compare Locke on the distinction between men (sic) and persons. See also Althusser (1970).

[3] I've been increasingly tempted to include in the cultural technē not only abstract "meanings" and such, but also material signs and symbols, for example a stop sign or a traffic light.

One step in a process of emancipation is to see how the local cultural technē is ideological. Geuss (1981) again:

> The very heart of the critical theory of society is its criticism of ideology. Their ideology is what prevents the agents in the society from correctly perceiving their true situation and real interests; if they are to free themselves from social repression, the agents must rid themselves of ideological illusion. (2–3)

A critical theory challenges the understandings that motivate and appear to justify our unjust practices; it can also offer resources to think and act differently, with better epistemic and normative tools.

> The critical theory induces self-reflection in the agents; by reflecting they come to realize that their form of consciousness is ideologically false and that the coercion from which they suffer is self-imposed. (Geuss 1981: 61)

Ideological "falsehood," as I understand it, is not always a failure of a sentence to express a true proposition. Ideological falsehoods can be true in that sense, but still be problematic, for they may employ concepts that are inapt for the purposes at hand (think of *grue*), or they may be part of a broader framework of ideas and assumptions that distorts our thinking and the social reality that our thinking (partly) constructs (Anderson 1995). The distortion may concern what is left out, or what inferences are made easy or salient (and what ones are obscured). Working with an externalist framework, commonly shared false semantic beliefs are also a potential site of ideology, for they can mask or distort how and whether our terms track kinds, sometimes contribute to the construction of them, and enable us to avoid taking responsibility for their effects (Haslanger 2012: chapters 2, 13).

My own reading of Frye (1983), MacKinnon (1987, 1989), Beauvoir (1989/1949), Wittig (1993), and others, had been liberating in the sense of unmasking the illusion of gender and of disrupting the ideology that gave shape to my practical orientation as a woman. Work by Collins (1990), Frye (1992), and Omi and Winant (1994) also prompted self-reflection on my investment race and related categories.[4] I vividly recall the moment when I read Wittig, "To refuse to be a woman, however, does not mean that one has to become a man" (Wittig 1981: 49). The obvious truth of this blew my mind, but I had never before considered the possibility it describes. Suddenly, new possibilities for agency became available; philosophy can prompt such openings (Bauer 2015). To notice how the existing practices and structures depend on distorted understandings can itself be liberating, in a sense, for we can begin to frame new intentions, explore different forms of agency, and take on new identities. This enhances our autonomy. But it doesn't make us free. Full emancipation requires also that we (collectively) change the unjust practices that structure our lives, and this requires more than thinking differently.

I saw my work as continuous with feminist and antiracist ideology critique. But the problem remained how to understand gender in such a way that it might be a basis

[4] There were personal reasons for such reflections as well. I have always had a complicated and often unhappy relationship with gender; and in 1994 and 1996 I became a parent of Black children in open adoptions and, as a result, my extended family became mostly Black.

for a feminist movement. Is there a way to speak of women, to struggle on behalf of women, to promote the interests of women, without presupposing a shared "identity," and without reinforcing a system that imposes gender while aiming to liberate us? The same questions could be raised, it seemed to me, to race.

In my "Gender and Race..." (2000) paper, I positioned myself as a feminist, antiracist, critical theorist, and offered definitions of gender, and of race, employing what I then called an *analytic* approach. I later (2012: chapter 6) changed the terminology to *ameliorative* approach, because the term 'analytic' was confusing to some whose background was shaped by the analytic/synthetic distinction and the history and critique of logical positivism. I'm not sure, however, that the new terminology was any better, and was, perhaps, just confusing in other ways. But it has stuck.

I had chosen the term 'analytic' having in mind Joan Scott's now classic paper, "Gender: A Useful Category of Historical Analysis" (1986). Scott's point was to demonstrate how the analytical category of *gender* (vs. sex), allows historians to trace the diverse forms gender takes: the cultural symbols, norms, and practices that shape what it is to be a woman (or man, or other) across time and place. Rather than a tool of homogenization, in her work the category of gender is a tool for theorizing diversity. Most importantly, it wasn't a term for an *identity* but a theoretical term for a process of social formation.

The idea that gender is a social formation has Marxist-feminist roots (e.g., Jaggar 1983; MacKinnon 1989; Young 1990). It would take me too far from the topic of this chapter to provide a history of Marxist and socialist feminism. However, it seemed clear that a way to avoid some of the problems that had plagued earlier efforts to define gender as *identity* was to characterize, in a very abstract way and compatible with many different instantiations, the sorts of social formations that produced women and men. Moreover, the analysis could be made especially apt for feminist purposes, if it focused subordinating social formations. Drawing on Catherine MacKinnon's analysis of gender in parallel to class (MacKinnon 1982),[5] I suggested this (rough versions):

A group G *is a gender* (in context C) iff$_{\text{df}}$ Gs members are similarly positioned as along some social dimension (economic, political, legal, social, etc.) (in C), and the members are "marked" by the dominant ideology (in C) as appropriately in this position by observed or imagined bodily features presumed to be evidence of reproductive capacities or function.

There are two dominant forms of gender at least in the contemporary world; but in some contexts there are others and could be even more:

[5] For what it is worth, I chose reproductive markers, rather than "eroticized dominance and submission" (which is MacKinnon's focus) because I had heard global feminists argue that the focus on sex manifested an American obsession (to the best of my knowledge, this was at MacKinnon's Gauss Lectures at Princeton in 1992). My aim was to find a (bodily) "marker" that ideology latched onto, and (real or imagined) markers of reproductive role—*presumed* to be linked by ideology to social role—seemed to be a good alternative. I assumed that the marker would vary depending on the social context, so in cultures that had different or mistaken ideas about reproduction, it would latch onto what *they* took to be the relevant bodily marker of reproductive role. See Hardt (1993); also Bettcher (2009: esp. pp. 105–7). This allows that there is *a sense* in which "sex" is also socially constructed, that is, what we count as sex depends on one's social context.

> S *is a woman* (in C) iff$_{\mathrm{df}}$ S is systematically *subordinated* along some dimension (economic, political, legal, social, etc.), and S is "marked" by the dominant ideology (in C) as a target for this treatment by observed or imagined bodily features presumed to be evidence of *a female's biological role in reproduction.*
>
> S *is a man* (in C) iff$_{\mathrm{df}}$ S is systematically *privileged* along some dimension (economic, political, legal, social, etc.), and S is "marked" by the dominant ideology (in C) as a target for this treatment by observed or imagined bodily features presumed to be evidence of *a male's biological role in reproduction.*

I also argued for an account of geo-ancestral groupings with races as instances:

> A group G *is racialized* (in context C) iff$_{\mathrm{df}}$ Gs members are socially positioned *as subordinate or privileged* along some dimension (economic, political, legal, social, etc.) (in C), and the group is "marked" by the dominant ideology (in C) as a target for this treatment by observed or imagined bodily features presumed to be evidence of *ancestral links to a certain geographical region.*

For example:

> *Whites are a racialized group in the US* by virtue of the fact that Whites are socially positioned as privileged along virtually all of the relevant social dimensions, and Whites are "marked" by the dominant ideology in the US as a target for this treatment by observed or imagined bodily features presumed to be evidence of ancestral links to Europe.

Theorizing *both* race and gender to be social processes rather than identities, and characterizing them at an abstract level that could allow for ideological specification in time and place, seemed to address some of the serious concerns others had raised.

Importantly, the goal of the project was not to capture *what we have in mind* when we use the terms 'woman,' or 'man,' 'Latinx,' or 'White' to describe ourselves or others. Instead, the idea was to offer a theoretical analysis of the social formations that produce raced and gendered groups of people, that is, it was an effort to resist the idea of gender or race as, primarily, identities. Of course, people develop race and gender identities, but I took that to be derivative from the social *process* that produces people who enact unjust gender and race practices.

However, the question of critical theory returns: If I was simply providing theories of gender and race, how was this supposed to be *emancipatory?* I suggested my accounts were grounded in and justified by the political goals of feminist and antiracist theory, which is an important commitment of critical theory. However, Geuss's (1991) challenge was to illuminate what makes a critical theory any different from an ordinary theory: "To be more exact: a critical theory has as its inherent aim to be the self-consciousness of a successful process of enlightenment and emancipation" (1981: 58). How does a theory do this?

My strategy (which, obviously, is a common strategy amongst social constructionists) was to appropriate the terms ordinarily used for identities, for the social processes and relations that make those identities available. The goal was to unmask the ideological assumption that gender and race are "natural," "given," or grounded simply in features of one's body, by shifting attention to the sources and consequences of those identities. The thought is that agents who come to understand the historical and political context of their gender and race identification and the role of their identification in perpetuating their own oppression and the oppression of others, will be taking a first step in a process of emancipation. (I recognized that simply introducing new theoretical terms for the categories was a possibility, but it would not have had the kind of personal and political effects I was hoping for.) Disruption of this sort does not assume that people who come to question their identity can or should immediately act on it. They may not have the power, security, or resources to do so. Social emancipation must be a collective effort and change more than minds. But under conditions of ideological oppression, ideology critique matters. It invites and sometimes produces a shift in one's practical orientation.

I was very explicit in my discussion that appropriation of the terms ordinarily used for race and gender identities should be handled carefully, for there were contexts in which such appropriation was either unwarranted or potentially harmful. I also suggested that there were significant questions of (my) authority in making such a move. The appropriation of existing terminology is risky, for the potential for a theory to function as emancipatory is very context sensitive. Although at this point I have no qualms about affirming the significance of the categories I defined (and still believe an understanding of these social formations are important to feminist efforts), the appropriation of the terminology was even more problematic than I then realized.

For example, by appropriating the terms 'woman' and 'man,' I problematically excluded some women from being counted as women and some men from being counted as men. Although my view does *not* require that one have male genitalia to be a man or female genitalia to be a woman, it does require being subject to subordination/privilege that is linked by ideology to the local bodily markers of reproductive role. This is a mistake: some women are prevented from presenting as women, and some men are prevented from presenting as men, and so do not meet the conditions I proposed (Bettcher 2009, 2012, 2014, 2016 (and comments); see also Jenkins 2016; Kapusta 2016).[6] There are also reasons to think that emancipatory identities that do not build in hierarchy should be available to those who are gendered and raced (e.g., Alcoff 2015). However, my accounts do capture something about the social formation of the public categories of men and women through dominant ascriptions; it is consistent with this that some men and women are excluded from these ascribed positions, and this is a problem that feminism should address.

On this reading of my earlier work, the accounts I offered did not simply appropriate the language of gender and race, but instead revealed features of our meanings that we were mostly unaware of. Drawing on a kind of semantic externalism, I went on to claim that the disruptive accounts I proposed in "Gender and Race..." might provide a better account of what we *actually* mean in dominant

[6] We normally grant someone the authority to avow who they "really are," for example, I'm really an artist (but cannot live as an artist due to economic/political/physical conditions), and this affirmation of existential identity should be extended to those who avow a gender other than one they were assigned at birth (see especially Bettcher 2009).

contexts than what we take ourselves to mean (2012: chapters 13, 14; see also Saul 2006). In the (then) contemporary United States, the dominant use of gender terms is exclusionary, and both gender and race terms actually track hierarchical social formations. It is important to have a way to capture this. At this point in the dialectic, my aim was to combine an account of the kind proposed by post-Quinean scientific essentialists with the sort of critical, ameliorative project I was committed to. Although early scientific essentialists (Putnam 1975; Kripke 1980; and others) focused on *natural kind terms* and thought that only *natural* scientists could be relied on to find the essences of things, I aimed to broaden the reach of their projects to include not only social science, but critical social theory more broadly. Working with an externalist framework, commonly shared false semantic beliefs (beliefs about what we mean by our terms) are a potential site of ideology, for they can mask or distort how and whether our terms track kinds, sometimes contribute to the construction of them, and enable us to avoid taking responsibility for their effects (2012: chapters 2, 13).

One lesson I draw from this is that linguistic choices that might be emancipatory at one moment, or for some individuals, or in response to a certain threat, may be inadequate in a broader context and even deepen other forms of oppression. This is not a new lesson and is to be expected. Although I gestured at these risks in early work, I didn't do enough to guard against a number of linguistic and political harms that could have been foreseen. Interestingly, this shift in my understanding of my own accounts altered the political import of my project: emancipation, it seems, involves at least two steps or moments. One moment is negative: we need to understand the failures of our current practices; another moment is positive: we need to suggest better alternatives.[7] My early accounts of *gender*, *race*, etc., might be employed in a negative moment to illuminate the exclusionary assumptions embedded in our use of certain social kind terms; however, they fail to offer adequate replacements (Jenkins 2016). This suggests that, more generally, efforts to provide tools for emancipation are not only context-sensitive, but may require both a disruptive moment that targets a set of existing practices, and also what we might call a visionary moment that gives us resources to create something better.

But where does this leave me with respect to the project of conceptual engineering? In the late 1990s, I was not really trying to do conceptual or semantic analysis. I was doing critical social theory and appropriating everyday terminology for the purposes of disrupting our identification with unjust social practices, that is, aiming for an emancipatory exposure of ideology. Later (in the mid-2000s), I became more invested in the idea of improving what we think and mean—of ameliorating—our conceptual and linguistic tools. Amelioration comes in many forms because are multiple ways to make things better. To clarify some of the different ways, it is helpful to work with a particular account of concepts.

[7] For example, as I see it, MacKinnon on sex (1987, 1989), Mills on race (1997), and Manne on misogyny (2017) are engaged in a negative moment; Barnes (2016), Jenkins (2016), and Alcoff (2015) are engaged in a positive moment.

3. Concepts?

As said before, I do not have an account of concepts. In this section I will lay out the basics of an externalist account of (coarse-grained) content that will provide a backdrop for my arguments.[8] I choose to proceed with these background assumptions, first, because I find the approach plausible; and, second, because I think it is a useful exercise to consider how amelioration might work within an externalist account of this sort. I am not the first to do this (Cappelen 2018), but I hope that I can provide a different approach that allows us to be at least hopeful about the power of ideology critique.[9]

I am anti-Fregean about meaning, content, and semantics: our utterances and our mental states do not have senses or concepts as their content (Stalnaker 1998). We express, believe, suppose (etc.) propositions, and propositions should be understood in terms of informational content, that is, "as truth conditions, propositions as functions from possible circumstances to truth values, or equivalently, as sets of possible situations" (Stalnaker 1998: 343). There are no "core commitments" associated with words that cannot be overturned or negotiated. Although in some sense we represent the world—propositions are abstract entities that carry information and are, to that extent, representational—the "mode of representation" is not part of the informational content of what we say and think. This allows you and me to think the same thing, the same proposition, even if we access what we are thinking differently.

How do utterances and mental states get the content they have? This is a project for *metasemantics.* Metasemantics, among other things, investigates how linguistic practices and conventions link utterances (occurring in response to others, in different parts of the world, and in other possible worlds) with propositions (cf. Plunkett 2011). It also concerns the ways in which mental states, whether linguistic or not, process and carry propositional and sub-propositional information. Some philosophers and psychologists are apt to suggest that concepts do an important part of this work. When I believe, for example, that the cat is on the mat, I have a concept of *cat* and *on* and *mat*, and various logical connectives, and I combine them in a thought. Another way to think about what's happening is that I have a set of capacities for processing information: capacities for attention, categorization,

[8] I develop the material in this section and those following in (Haslanger forthcoming a).

[9] In the conclusion of his recent book on conceptual engineering, Capellen (2018) maintains that "the changes that happen [in linguistic and conceptual change] are the result of inscrutable external factors that we lack control over" (p. 199) and that "Anyone who spends time thinking and talking about large-scale normative matters should do so without holding out too much hope that their talking and thinking will have significant or predictable effects on the relevant aspect of the world" (p. 200). Although I sometimes fall into such pessimism, I think Capellen doesn't fully appreciate the ways in which culture makes a difference to social life and the sociology of social movements. Feminist interventions have had a huge difference on the construction of gender, and although law (Title IX), technology (especially oral birth control), and economics (such as the abandonment of the "family wage") have made huge differences, the stability of social hierarchy depends on an interdependence between multiple factors, including culture. Such social change did not happen by accident and relied on tools in the cultural technē to challenge existing gender norms. Work by Beauvoir, MacKinnon, Butler, and many others has demonstrated that a book that reframes our concepts can have profound ramifications. Cappelen's discussion suggests that philosophers who hope to contribute to social movements are naïve. However, I haven't found much history, sociology, political theory, or even social/political philosophy, in his book to back up his claims.

interpretation, memory, language, inference, affect, and the like, and these capacities organize inputs in ways that represent information discursively. Propositions—that, as mentioned before, can be understood just as sets of possible worlds—are encoded in a variety of different ways. I have easy and direct access to some propositions, whereas my access to others is more mediated or inferential.

Some of the mental capacities we have are hard-wired. However, humans and some other non-human animals have tremendous perceptual, cognitive, and affective flexibility that enable them to adapt to a variety of settings. For humans and other social animals, adaptation is deeply connected to coordination. Learned mechanisms of coordination require selective attention to public entities that serve as signals in response to which we do our part. A red light might take up only a tiny space in our visual field, but drivers are highly responsive to it, for failure to see it or respond appropriately may be a matter of life or death. Language is one form of public information exchange that gives us a basis for coordination, but language itself depends on more basic capacities we have for picking up information from others, and from our environment, and sending it to others (Zawidzki 2013). In responding to and transmitting information, we develop predictable patterns of behavior that others come to expect; these patterns, when upheld by the coordination group, constitute practices. Non-human animals have such capacities well.

For example, dogs and humans coordinate. One crucial task for this trans-species coordination is the timing of the opening and closing of outside doors. Dogs need to go out to "do their business" and usually cannot open the door for themselves; they need humans to do it for them. In our house, we have a bell hung by the door. When Sparky wants to go outside, he rings the bell. We come to the door and let him out. Sparky had to be taught to use the bell. The bell does not have a "natural meaning" (Grice 1957), but it has a meaning in the ecosystem of our home. In response to his need, Sparky rings the bell, expecting that we will come open the door; we hear the door and expect Sparky to be waiting near the door to go out. The bell does not have linguistic meaning, but it has, what I call, social meaning. The ringing bell provides information that we—Sparky and the other family members—are able to access due to a process of learning from each other in an effort to coordinate. Our capacities for attention, interpretation, categorization, etc., have adjusted to take in this information and act on it in expected ways. Not all social meanings are about coordination, but it is plausible that the capacities that make social meaning possible originated in the need for coordination. Humans, though, are able to take delight in social meanings for their own sake and use them to develop cultures that have lives of their own (Balkin 1998; Zawidzki 2013).

We might say that if one develops a sophisticated set of capacities that enables them to process certain kinds of information, say, about *X*s, then they have the concept of *X*. Consider Yalcin:

> To possess a concept is to have an ability to cut logical space in a certain way, to distinguish possibilities in terms of the sorts of things that answer to the concept.... A concept determines a matrix of distinctions. For example, the concept/subject matter BACHELOR corresponds to the partition of logical space distinguishing possibilities depending on what's happening with the bachelors at each world—so that two worlds will belong to the same cell just in case they

> don't differ in their bachelor respects. To possess a concept, on this idea, is to be capable of entering states of mind sensitive to the associated distinctions.
>
> (Yalcin 2016: 14; also Pérez Carballo 2016: 466ff)

The content of the concept is a partition of logical space.[10] From a psychological point of view, however, *possession* of the concept may occur by virtue of different cognitive mechanisms and give rise to very different dispositions in different individuals. What it means to have a certain concept of *X* is not just what you can articulate, but how you respond to and coordinate with others in your environment, that is, how your capacities for attention, categorization, interpretation, memory, language, inference, affect, and the like, are marshalled for the purpose to coordinating (and refusing to coordinate) with others in response to particular kinds of information.

For example, we may have the same concept of *cat* —the informational content of the concept *cat* is the same for each of us—but our possession of it occurs in somewhat different ways so that certain inferences are more direct for me than they are for you, or that I am more ready to apply the concept than you. Or it may be that because you know more about *cat*s, you have a sensitivity to different kinds of *cat*, so your partition of logical space is more fine-grained.

Consider the concept of war. "William III of England believed that England could avoid war with France. Did he believe that England could avoid nuclear war with France?" (Yalcin 2016: 12) The content of William III's concept of war might be represented as a division between war (land or sea), and non-war (see Figure 1). Logical space is divided into regions that contain worlds with sea wars, worlds with land wars, and worlds with no wars (considered timelessly); worlds (like ours) can occur in more than one partition of space because we have had both land wars and sea wars. Utopian worlds, perhaps, have no wars. However, once we learn about the possibility of nuclear war, we can carve the space of possibilities in a more fine-grained way (see Figure 2). And as we learn more about war, and as military technology and tactics evolve, we might want to not only add complexity, but also redraw distinctions that seemed exhaustive before (see Figure 3).

William III could not draw the distinctions between kinds of wars that we draw. In fact, William III couldn't even imagine nuclear war or cyber war. Our concept of war is more fine-grained than his and we have knowledge about more kinds of war. That is to say, we can draw distinctions between kinds of wars that he was unable to draw. But that doesn't mean that we don't, in an important sense, employ the same broad concept when he, and we, think and talk of war. We may be able to form more true beliefs about water, but amelioration, as we shall see, isn't all about increasing truth. In the social domain, in particular, some facts depend, in part, on what cultural, linguistic, and conceptual resources are available, because our cultural technē shapes

[10] Many people have developed this view in different ways. I think there is enough of a shared background so that one need not be a thoroughgoing externalist to accept much of what I say here. I adopt a Stalnakerian framework, but there are other ways of making the same points. See, for example, Jackson (2000). Thanks to David Plunkett for pointing this out. Note also that Yalcin and Pérez Carballo are expressivists about certain kinds of content. I don't, here, mean to embrace their full views but am simply drawing on the passages I cite.

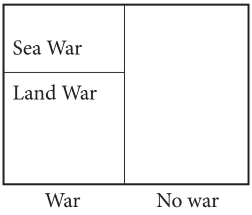

Figure 1

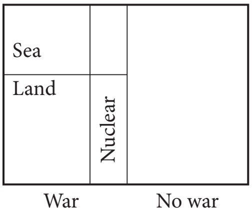

Figure 2

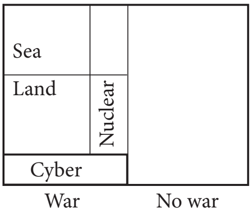

Figure 3

the world. Simple examples are moves in games: the cultural technē of soccer made possible the fact that there was a total of four red cards at the 2018 World Cup.

Suppose we start with this background picture: mental and linguistic representations have informational content; the *informational content* of a concept is a partition of logical space that divides possibilities.[11] To *possess* a concept is to have some cluster of capacities and mechanisms for using that grid of possibilities at some level of resolution, that is, for making the distinction(s) in question, and processing and storing the relevant information.[12] We are now in a better position to distinguish some different forms of amelioration.

[11] There are different ways to spell this out. For example, it might be a distinction between different sets of possibilia (an intension or modal profile (Schroeter and Schroeter 2015: 441)), or a distinction between sets of propositions with respect to their subject matter (e.g., Yablo 2014). For the purposes at hand, I find it best to work with the intuitive idea that the partitions serving as the content of concepts are of possible individuals rather than worlds. So the content of a concept (DOG) is determined by a function from a world (w_1, w_2, w_3)—or a world-time pair ($<w_1, t_1>$, $<w_2, t_2>$. . .)—to a set of individuals in that world (dogs$_1$, dogs$_2$, dogs$_3$).

[12] In order to avoid confusion, I use the term 'distinction' and 'distinguish' or 'classification' and 'classify' for the linguistic/conceptual acts of noting or marking differences, and the terms 'difference,'

- *Epistemic amelioration*: we improve our understanding of the informational content of the concept.
  - *Refinement:* We use concepts without having a very solid grasp of them. We may not be able to apply the concept to some possibilities, and there may be gaps in our judgments about cases. So we refine our concept based on a broader or deeper knowledge of the phenomenon, for example, by gaining empirical knowledge, undertaking conceptual genealogy, and/or insight into logical space at a more fine-grained resolution.
  - *Experiential access:* we improve our access the informational content, gaining more reliable or illuminating access by different modes of presentation, for example, those who have experienced war with "boots on the ground" have a different appreciation of what war is.
- *Informational/semantic amelioration:* we change what partition of logical space the term or concept represents, that is, we undertake to change our thought and talk to do better in tracking reality. Better how?
  - A*lethic*: we are improving the resources available to track truths, for example, a biological account of race prevents us from tracking important truths about race. Making these truths articulable using a social constructionist account can unmask ideology; it can also shine a light on new (emancipatory) possibilities. Ordinary scientific research can also shift content, for example, biological theory can prompt changes in the distinction between animals and other kingdoms, with the result that the extension of 'animal' shifts.[13]
  - *Pragmatic*: what we track with our language and our concepts can make life easier by shifting terms of coordination, for example, 'lunch' once picked out a light meal at any time of day or night. Now when we invite a friend for lunch, we convey, with our term, information about the time of the day when we might meet.
  - *Moral:* because what mean can affect what we do and what there is, semantic amelioration can also be (broadly) *moral*, for example, if the informational content of (legal) 'marriage' excludes same-sex couples, this is a moral wrong.

Connecting these distinctions with my earlier projects, epistemic amelioration is especially apt for debunking projects in the negative moment of critical theory. We expose what kind we are simultaneously tracking and creating. Semantic amelioration is especially apt for the positive moment when we undertake to revise what information that is expressed when we use certain terms or conceptualize our options. We highlight and recommend new partitions of logical space—an expansion or contraction of informational content—as a basis for our future coordination. It is fairly clear how the epistemic amelioration might be undertaken. But is semantic amelioration really possible? If a word or concept's content is essential to it (and what

'differentiate,' or "division" for the ontological basis for distinctions when they divide the world, that is, we distinguish objects that are different; our distinctions aim to capture what differentiates the objects we're interested in, we draw distinctions along divisions.

[13] Putnam (1962/1975) provides an important early discussion of the importance of amelioration in science and the dangers of maintaining the analytic/synthetic distinctions.

else could be essential to it?), then semantic amelioration would seem to be incoherent.[14] At best we are instead recommending the adoption of new terms and concepts.[15]

Note that both epistemic and semantic amelioration assume that the informational content of our concept is *given*. In the first case, we are looking to improve our grasp of it; in the second case, we are considering whether to adjust it. There comes a point, however, when there isn't enough overlap to assume we are thinking or talking about the same thing. Suppose I think that a bachelor must be "on the make" or available for marriage, and you don't. How do we adjudicate our disagreements? We also find terminological expansions that seem to go too far, for example, we now talk of the "war on drugs" and the "war on terror." Are these really wars, or just metaphors? It would seem that we should adjudicate this by reference to the contents of 'war,' 'bachelor,' or other term in question.

But what is the content? How is the content determined, and how do we know what it is? I will consider this question in the next section; however, it might be useful to have some examples before us. Water, gold, and jade have been standard examples in the literature about natural kinds. Marriage and family provide examples of controversial social kinds. Gender and race are examples where there is disagreement over whether they are natural or social.

- Is water essentially H_2O? Is H_2O the informational content of our thought and talk about water?
- According to the dominant understanding of *marriage* prior to 2004 in the United States, marriage is a legal and/or religious status restricted to one man and one woman. A marriage between two men or two women was, for many, unintelligible. Same-sex marriage is now legal in the United States, and in some religions, it is fully accepted. Both the institution of marriage and our understanding of it have changed. A consequence of this is that not only how we think of families, but how families are constituted has changed. How should we understand these changes? What is content of the concept of marriage? Have we improved our concept of *marriage* (or of *family*) or adopted a new concept?
- According to Kant, races are groups of humans who have evolved to have certain distinguishing physical and psychological traits, that "resist further transformation." So, for example, the native American, having had to endure the extreme cold, suffers from a "half-extinguished life power" (Kant 1775/2000, 17). And the Negro, because he has benefited from the rich land of Africa, is "strong,

[14] Consider two partitions of logical space D and D*. Our concept of *bachelor*, say, has D as its content: a particular set of all and only possible bachelors. Suppose an ameliorator comes along and suggests that, instead, the concept of *bachelor* has D* as its content, that is, a different set that is a proper subset of D. It might seem tempting to say that D* can serve as the content for *a* concept of *bachelor*, but that would be a different concept from ours because *our concept* has D as its content and concepts have their informational content essentially.

[15] Dan López de Sa has helpfully pressed me to articulate more clearly why it matters whether we change our concepts or adopt new ones, especially if we can use the same word for new concepts. I think that in some cases it doesn't matter and we can see it as a political choice. But in other cases it is helpful because we can position ourselves as correcting the errors of our past. For example, the informational content of 'rape' has changed, but we were wrong *about rape* (using the same concept) when we neglected to count forced sex between a husband and wife as rape.

> fleshy, and agile. However, because he is so amply supplied by his motherland, he is also lazy, indolent, and dawdling" (Kant 1775/2000, 17). Whites, however, have diverged least from the original form and the "noble blond form" characterized by its "tender white skin, reddish hair, and pale blue eyes" that inhabited the northern regions of Germany, is the strongest. This form itself does not constitute a race, but only a lineage within the white race. However, "This stock would have gotten on well enough to persist as a race if the further development of this deviation had not been so frequently interrupted by interbreeding with alien stocks" (Kant 1775/2000: 20). Social constructionists about race reject Kant's claims about race and propose that race is a social category (Mills 1997; Hardimon 2003, 2017; Taylor 2004; Mallon 2004, 2006; Haslanger 2012: chapter 6; Jeffers 2013). Are social constructionists improving our concept of race or introducing a new concept? What is the relationship between Kant's concept of race and theirs?

4. "Conceptual Analysis"[16]

For an externalist, to answer the question "What is the content/meaning of *X*?" one should focus on the question "What is X?" And when we ask "What is X?"—not about a particular (silverfish, for example) but a type or kind—usually a better way to put the question is: *What it is to be* (an) *X*? When natural phenomena such as water, silverfish, or oak trees, it would be, at the very least, odd to answer the questions by consulting our linguistic intuitions. Our judgments about when to use the term 'silverfish' don't tell us what a silverfish is. However, there are a variety of "What is X?" questions that many philosophers seem to think can be answered by discovering the meaning of the term(s) substituted for *X*, as determined by our disposition to apply the term(s) in question, for example, 'knowledge,' 'moral worth,' 'justice,' 'a person,' 'causation.' In some of these cases, one might think that this *a priori* methodology is warranted because the boundaries of these kinds depend in some way on us and our practices. Perhaps moral worth, justice, personhood, and the like, don't exist independently our judgments of what counts as moral worth, justice, and personhood. So, of course, we should at least begin by investigating our judgments and putting them in order. (This is more plausible in some cases than in others: the answer would have to be more complicated in cases such as 'causation' or 'intrinsic property.')

But the idea that (some) philosophical kinds "depend on us," is not entirely clear; nor is it clear why our *a priori* (linguistic) reflections should be sufficient to provide an adequate theory of them. For example, "What is a sheriff?" What counts as a sheriff depends entirely on human stipulation; but even if you are a competent user of the term 'sheriff,' you may not be able to tell me what a sheriff is. A full answer would presumably require information about the jurisdiction of sheriffs, what their responsibilities are, how they are chosen, etc., as determined by law. We might need to

[16] Some paragraphs in this section also appear in my chapter, "What is Race? Tracing its Socio-Political Reality," and my "Replies," both in Glasgow et al. (2019).

consult experts in civics to get answers (and the answers will depend on what country we are in). We can't just depend on common sense or linguistic intuitions; there are no sheriffs outside of a humanly constructed system of government.

In the case of 'sheriff,' there will be a well-defined role specified by statute, and someone who knows the relevant statutes will know the answers to our questions. But there are also social phenomena that in some sense "depend on us" but are not stipulated or planned by us. Such social phenomena range from macro-scale economic depressions, globalization, urbanization, and gentrification, to more local social practices and relations, for example, within a town, religious congregation, or family. These phenomena call for explanation, and the social sciences (broadly construed) endeavor to provide theories that enable us to understand them, usually identifying kinds of institutions, economic relations, cultural traditions, social meanings, and psychological predispositions to do so. The kinds in question are social kinds, in the sense that they are kinds of things that exist in the social world (and so, in some sense, depend on us). But we discover these kinds through empirical enquiry just as we discover chemical kinds through empirical inquiry.

For example, accounts of gentrification often make reference to the "urban pioneer," sometimes characterized as artists and "bohemians" who take advantage of low rents in poor neighborhoods. Once single people who share rent enter a neighborhood, businesses (such as cafés and pubs) take interest, and landlords see opportunities to raise rents, which drives out the locals. *Urban pioneers* is a functional kind that identifies a particular role in an evolving real estate market. The term 'pioneer' is chosen due to the perceived parallel with pioneers who "settled" the western United States, displacing the local population. If someone were to object to the term 'pioneer'—perhaps thinking that it carried an overly-positive connotation—this would not undermine the explanatory claims.[17] The adequacy of explaining gentrification by reference to singles moving into an urban neighborhood does not depend on our linguistic intuitions about applying the term 'pioneer' to them. The choice of terminology was intended to illuminate a parallel; if the terminological choice doesn't work, then another term could (and sometimes should) be used as a substitute.

However, insofar as philosophical kinds such as *justice* and *personhood* "depend on us," it is not in the sense that we stipulate what they are (like *sheriff*), or in the sense that they serve in explanations of social phenomena (like *urban pioneers*). In the case of *sheriff*, you might think that there aren't any independent facts we're trying to accommodate. (Oversimplifying), we simply create sheriffs and then talk about them. In the case of *urban pioneer*, the prior understandings of 'pioneer' are not crucial to the explanation provided by the theory. In the case of *person*, there is something we are aiming to understand that is not simply constituted by what we say, but at the same time, our conclusions cannot float completely free of our practices.

[17] Metaphors and analogies can play an important and even ineliminable role in theorizing and can aid in explanation. My claim here is only that the choice of terminology for the functional kinds in the proposed mechanisms of gentrification (specifically the influx of singles) is not essential to the success of the model for some purposes (though it may be for others).

How might we explain this? Note that in the philosophical cases, we are not situated as anthropologists trying to understand the social life of the "natives." Nor are we legislators specifying new practices. We seek an understanding of practices in which we are currently engaged as participants. The practices are not fully understood, however. And they are open-ended, revisable, possibly self-defeating. In making sense of them, we are making judgments about how to better understand what we are doing, and how then to go on.[18]

This is not primarily a linguistic exercise: we aren't just deciding how to use existing terminology to pick out things in the world, but how to collectively orient ourselves towards the world and each other. As I've discussed above, language, concepts, symbols, and what, more generally, I've called *cultural technē*, are not only devices that carry information and allow us to communicate, but are also, and as importantly, tools for coordination. Coordination is a fundamental human task. Language helps us coordinate, not only by providing a means to interpret and predict others, but to shape each other so that complex forms of coordination are possible. As Tadeusz Zawidzki puts it:

> Our social accomplishments are not by-products of individualized cognitive feats... Rather, through a form of "group selection," simultaneously interpretive and regulative frameworks that support our social accomplishments, including pervasive, institutionalized cooperation and coordination, language, and so on, have evolved. In the mindshaping metaphor, distinctively human social cognition is conceptualized as a group accomplishment, involving simultaneously interpretive and regulative frameworks that function to shape minds, which these frameworks can then be used to easily and usefully interpret. (Zawidzki 2013: xiii)

Language provides a means to communicate information, but of course, the world presents an information overload. Coordination requires us to select *what* information is important, ways of linking information, and drawing consequences for action. So when we consider "how to go on," language—a practice within practices—is itself is a proper target of philosophical inquiry. We are situated within a tradition of linguistic practices that have already shaped our world; so ignoring those practices would be a mistake. We are situated inquirers, and the question is how we should go on, given where we have been, where we are now, and where we are trying to go (Wittgenstein 1958: esp. sections 185–243; Kripke 1982: esp. fn. 13, pp. 18–19; Lear 1986).

5. Representational Traditions: 'Water' as an Example

Laura and François Schroeter (2015) offer an account of meaning that situates our linguistic activities within our broader social practices. They focus on the example of 'water,' and suggest that to determine what 'water' means, we should undertake an inquiry into what water *is*. "It's important to notice that from the first-person perspective, the object-level question 'what is x?' is equivalent to the metalevel

[18] David Plunkett has pointed out to me that this sounds a lot like Ronald Dworkin's (2011) interpretivism. I agree; however, as I hope will become clearer as we proceed, there are important differences between my view and Dworkin's.

question 'what is the reference (or, more generally, the semantic value) of my term "x"?'" (2015: 419).

But how do we determine what water is? We cannot assume from the start that this is a task for the chemist, for when the chemist says that water is H_2O, she may be using the term in a technical sense, in which case it would not provide an account of what the ordinary person means by 'water.' But neither can we just undertake reflection on linguistic usage or common sense.

> Before you explicitly reflect on the question of what water is, your own assumptions about the topic are bound to be heterogeneous, incomplete, and partially contradictory—and this heterogeneity is only exacerbated when you take your whole community's views into account. Thus justifying an answer to a 'what is x?' question is nothing like slotting some missing values into an implicitly grasped formula. Your goal in rational deliberation is to find some principled way of prioritizing and systematizing your own and your community's commitments about water, so as to identify the appropriate normative standards for evaluating the truth and acceptability of beliefs about the topic. (2015: 430)

The broad idea is this: when we deliberate about what X [water, race, freewill, moral worth, gender . . .] is, we have to start with *something*. In the sorts of cases we are considering, we can take ourselves to be situated within a broad representational tradition concerned with X (we are not starting from scratch and stipulating the meaning of theoretical terms). And we may assume that the tradition has a certain epistemic ambition, so we may "take our words and thoughts to represent genuinely interesting and important features of the world—not just whatever happens to satisfy our current criteria" (2015: 436). So scientific inquiry is also relevant, since it discloses some parts of the world that are important for many of our purposes. But where do we begin?

The Schroeters (2015: 426) give a sample of inputs to deliberation in the case of water:

- *Particular instances*: there's water in this bottle, in Port Phillip Bay, Lake Michigan, etc.
- *Perceptual gestalts*: the characteristic look, taste, odor, tactile resistance and heaviness of water.
- *Physical roles*: water's rough boiling point, its transformation into steam, its role as a solvent, the fact that it expands when it freezes, etc.
- *Biological roles*: water's necessity for the survival of plants and animals; how it's ingested; the effects of water deprivation; etc.
- *Practical roles*: the roles water plays in agriculture, transport, washing, cooking, surfing, etc.
- *Symbolic roles*: water is strongly associated with cleanliness and purity, it plays an important role in many religious rituals, etc.
- *Explanatory roles*: water has a non-obvious explanatory structure, which explains many of its characteristic roles; water is composed of H_20.
- *Epistemology*: water is easy to spot but hard to define; our beliefs about water may be mistaken or incomplete; observation of instances of water grounds induction to unobserved cases.

The aim is to answer to the "What is X?" question. Their project is not semantic but *meta-semantic*: what would make it the case that the informational content of 'water' is all and only possible instances of H_2O? The inputs just considered help us narrow down the target kind so we can investigate it further. As we proceed, we may find that some of our background beliefs are false and our theoretical efforts misguided. But what do we do with these inputs? How do we balance various considerations? Schroeter and Schroeter (2015) propose that

> ...ideal epistemic methods for answering 'what is x?' questions hinge on rationalizing interpretation of one's representational traditions. You need to diagnose the most important representational interests at stake in a representational tradition with 'x', and you should identify the correct verdict about the nature of x as the one that makes best sense of those interests. (2015: 430)

A rationalizing interpretation, on their view, is not determined by reports of beliefs and intentions of participants in the tradition, nor is it a causal explanation of the tradition:

> From the deliberative perspective of a rational epistemic agent, the interests that are relevant to adjudicating 'what is x?' questions are those that help justify or rationalize that tradition. Ideal methods for adjudicating 'what is x?' questions don't simply construe representational practices as meeting psychologically or causally fixed representational interests. Our interpretive methods construe them as meeting representational interests that help make sense of our practices—that help construe them as having a point or rationale. (2015: 435)

On the Schroeters' (2015) view, there is a best interpretation of the representational tradition, where the scope of that tradition is determined by commitment to *de jure* sameness of reference and shared linguistic and epistemic practices (p. 428). What I mean is not just a function of what I think water is, or any old interpretation of our representational tradition: I can get the meaning wrong if I don't do justice to the interpretive task. For example, if I decide that, given our interests and collective uses of the term, water is the alcoholic beverage also known as 'beer,' I would be wrong. I would have failed to capture a reasonable interpretation of our representational tradition. But I could also be wrong if I miss what is worth talking about:

> As rational epistemic agents, we normally take our words and thoughts to represent genuinely interesting and important features of the world—not just whatever happens to satisfy our current criteria. When asking about the nature of water (or free will, color, etc.), we don't assume that we (or our community as a whole) already implicitly know the right answer. (2015: 436)

We postulate ambiguity or opt for an error theory only as a last resort. In the case of 'water,' plausibly H_2O wins because it organizes and explains the relevant inputs to deliberation. So the informational content of 'water' (and similar words in other languages), is a partition in logical space on which chemical composition is the differentiating feature.

The Schroeters' metasemantics enables us to determine when we are gaining knowledge about the meaning of a term, and when we are talking about something altogether different: If Aristotle thinks that water is essentially a fundamental element and I think it is essentially H_2O (and not an element), are we talking about the same thing? Yes, because the term/concept functions in importantly similar ways in our communities.[19] He was wrong about the chemical composition, but that isn't surprising, given the technological and scientific limitations of his day.

6. Conceptual Amelioration

This gives us a helpful model for understanding at least some conceptual amelioration. We are part of a complicated representational tradition with various threads and various purposes for employing the concept *X*. Although we are adept at using the concept in many contexts, its exact informational content is obscure. Perhaps we find ourselves in controversy, or in a situation where more precision or understanding is needed. We engage in reflection on the representational tradition and find that certain ways of going on—ways of interpreting our past practices projecting them into the future—requires adjustment in our judgments about what counts as an *X*, or what is true of *X*s. This isn't just an epistemic amelioration: we adjust the informational content—the partition of logical space—in order to do justice to the complex role of the concept in our practices.

Although sympathetic with the Schroeters' view, I am doubtful that there is a single "best" interpretation of a representational tradition. And although I agree that language is situated within our practices, I am not convinced that the right position from which to evaluate the tradition is that of the "rational epistemic agent."[20] Of course, whether this is adequate depends on how we characterize the rational epistemic agent and what bits of language we are considering; but I think we need to bring to the forefront the role of language in shaping us for coordination. As Zawidzki says in the quote I included above, "Our social accomplishments are not by-products of individualized cognitive feats." Especially in the case of social kinds, the evaluation of a linguistic/conceptual tradition should consider how it shapes us to be responsive to each other (and the world), what forms of social reality are created (institutions, practices, artifacts, identities), and whether or how we might do better.

But a variation on the Schroeters' metasemantics can also give us a guide that will set limits on amelioration. We use many of our terms/concepts for multiple purposes,

[19] It is important to note the difference between a *functional concept* and the *function of a concept.* A functional concept is a concept whose informational content divides logical space between those things that fulfill a certain function and those things that don't. The concept of *heart* is a functional concept because hearts are hearts by virtue of pumping blood, that is, of having that function in a body. If we grant that water is (essentially) H_2O, then *water* is not a functional concept, because water isn't what it is because of how it functions. It is what it is by virtue of its composition. However, it is compatible with this that *the concept of water* has a function, that is, that the division of logical space along the water/non-water axis and the ability to track that distinction has an important function in society. For example, our ability to distinguish water from other similar but toxic chemicals assists our survival; the concept may function to help us be aware of the distinction, teach it to our offspring, etc.

[20] There is a sense in which the Schroeters' externalism depends on an internalist metasemantics—this is something they acknowledge and a way of capturing what they take to be the advantages of both internalist and externalist insights. I am aiming to be less internalist in my metasemantics for I see metasemantics as not just about how a rational epistemic agent would judge cases, but about how language is taken up and informs our practices. Agents can play a role in this process, hence the possibility of amelioration. More in this below. Thanks to David Plunkett for urging me to clarify this.

not all of which point to the same thing. It is reasonable for us to settle disputes by arguing for a particular way of extending a thread that has been part of our representational tradition, but it is not necessary for us to adhere to the tradition strictly, especially if the threads are tangled or if the world has changed so that background assumptions are no longer valid.[21]

Water is perhaps not the best example when we are thinking about amelioration. Consider instead the concept of *family*. On the approach I have been proposing, we should consider the function of a *family* concept in organizing our lives together and in shaping our self-understandings to engage fluently in the practices that enable coordination. Plausibly, to have a *family*-concept is (roughly) to have a cluster of mechanisms for processing information about the coordination of domestic life, for example, intimacy, sex, raising of children, economic partnership, intergenerational transfers of traditions and property.[22] These are tasks that any culture has to manage somehow and many do it through the construction of families. Since its inception, anthropology has explored kinship systems, and the formation of social identities that "fit" within such systems. A generic conception of family is not enough, however; more specific ideas about "what it is to be a family" are required in order to coordinate effectively. So language evolves to pick out certain social formations and roles; symbols, metaphors, and narrative tropes take hold so individuals can form expectations about how to manage these tasks. The bionormative nuclear family is not a "natural" phenomenon, and societies that rely on it for coordination provide a script. Many have organized their lives around the script: first comes love, then comes marriage, then comes a baby in the baby carriage. There are video games for girls in which they plan fantasy weddings, and even adult women envision themselves as Cinderella. (In 2012, Weddingbee had 14,974 members who were not engaged (Baker 2013; Patterson 2014).) Such observations are relevant to evaluating "our" concept of family, just as observations about water were in the Schroeters' example.

[21] Herman Cappelen (2018) argues against ameliorative projects such as mine because "there simply isn't a good way to identify 'the phenomenon' except disquotationally and the disquotational identification is unresponsive to the challenge of articulating the limits of revision" (p. 184). But why is disquotation the only option? The Schroeters have identified multiple routes for identifying 'the phenomenon' that is the subject matter for our talk of water that goes well beyond disquotation. Below, I will discuss "the phenomenon" of family which, I assume, we can identify by the multiple roles that talk of family plays in ours and (assuming translation) other cultures; anthropologists have been doing this at least since the nineteenth century. Of course we cannot determine 'the phenomenon' *a priori*, but as externalists, why should we expect otherwise?

[22] Cappelen (2018: chapter 16) argues that a functional approach to concepts, such as the one I suggest here, is implausible because concepts don't have functions: "Of course, people have goals and aims and purposes when they use words on particular occasions. But I don't think concepts have purposes and certainly not words (or extensions or intensions)" (p. 181). I am puzzled. There is, for example, decades of work on functional and structural explanation. Functions are attributed to parts of a system when they contribute to the ongoing working of the system (hearts, carburetors). Although it is misleading (my mistake!) to speak of *the* function or purpose in such cases, things can a function in the social domain without human intention, for example, the gendered division of labor functions to keep women's wages low, which functions to make exit from marriage more difficult than men's, which functions to sustain women's subordination to men. Language can also be part of a homeostatic social system because of the way it enables us to call attention to certain phenomena and facilitates communication.

For many generations, in the United States at least, the concept *family* conveyed a specific informational content: it included a husband, his (same-race) wife, and her biological offspring.[23] This social formation was legally and culturally entrenched. Other ways of arranging domestic life, although acknowledged, were either unimaginable due to lack of technology (ART), or cognitive bias (queer families), or were tolerated only insofar as they mimicked the dominant formation (adoption, step-families, unmarried and mixed-race couples with children). "Childless" couples didn't really count as families (note: "when are you going to start a family?"). At this point, heterosexual bionormative nuclear families (HBNFs) constitute one kind of family, but it is broadly recognized that families include domestic arrangements made by adoption, donated gametes, families with single parents, same-sex, trans and genderqueer parents, unmarried parents, and extended families of various kinds. I believe that conceptual amelioration has occurred, both as a result of pressure by social movements and by the development of reproductive technology: that is, the informational content of the term 'family' has changed, due to a change in the social conditions. More ways of organizing domestic life have become normalized. But our concept of family has evolved, improved, at least in part, through the political work done by LGBTQ and adoption activists in conceptual engineering with slogans such as "we are families too."

What sort of amelioration is this? Is it epistemic or semantic? If we are thinking of families as ways of organizing domestic life, generally speaking, then it would seem to be epistemic: there are more ways of organizing domestic life than in HBNFs, even in the actual world, and those who confined their understanding of family to HBNFs were mistaken. Like William III with respect to *war*, they just didn't have a sense of the fine-grained partitions in the logical space of family. However, if we are thinking of the informational content expressed when we speak of families, then it seems to be a semantic amelioration: the partition of logical space communicated changes when we include more kinds of domestic arrangements as families. But this shift doesn't require us to form a new concept. I suggested above that possession of a concept concerns how you respond to and coordinate with others in your environment, that is, how your capacities for attention, categorization, interpretation, memory, language, inference, affect, and the like are marshalled for the purpose to coordinating (and refusing to coordinate) with others in response to particular kinds of information. Our capacity to coordinate with others in organizing domestic life—even considering the specifics of how we do it around here—does not necessarily break down as we expand who counts as family. The concept—our capacities for processing relevant information—evolve as we recognize the possibility of new social formations and new norms.

Suppose, however, that Albert thinks of families as heterosexual married couples and the biological offspring of the wife (HBNFs), and only such families. Albert has *a* concept of family. Albert may resist calling same-sex couples raising children a "family" on moral grounds; he may be opposed to divorce and take adoption to create family-like groups, but not real families. We may undertake to epistemically

[23] Currently the conditions for being the legal father of a child are complicated and vary from state to state (FindLaw n.d.); in some cases it is still assumed to be the husband of the woman who gives birth. Historically this was generally the case in the United States.

ameliorate Albert's concept of family by pointing out to him that within our representational tradition, *family* is a functional notion concerned with the management of domestic life. This might be to offer him a better grasp of what he means. However, he may resist this intervention. He may have a different interpretation of our representational tradition, or he may not care what the rest of us think and instead choose to consider a different representational tradition as his own, for example, a particular religious tradition. Because Albert's capacities for processing information about domestic life is at odds with the broader community, he may have trouble communicating and coordinating with others. And it would be plausible to say that he doesn't share *our* concept of *family*. "Our" concept of family is embedded in our practices and our laws, our forms of intimacy and love. At this point in time, families include more than HBNFs.

But simply pointing to the informational content that most of us (around here) express and rely on to coordinate doesn't seem to be sufficient to capture what some want in the idea of conceptual amelioration. Yes, some people are out of sync with the broader representational tradition of which they are a part, and if they want to conform to that tradition, they should adjust their understandings of what they are talking about. But the idea of conceptual amelioration is especially valuable when we are cutting against the grain—there are times when the community seems to be (or has been) committed to a particular understanding that should be resisted. Conceptual amelioration should not just be a matter of demanding that the outliers conform to the dominant understanding of what we are talking about. Critical (feminist, antiracist, queer, disability) theorists, in particular, seek to ameliorate by shifting our concepts away from dominant understandings. This calls for a more robustly normative sense of amelioration.

Return to the concept of *family*. Families are one kind of social formation. They typically consist of individuals who engage in sets of practices distributing things of (+/−) value: sex, childcare, homecare, eldercare, emotional labor, work in income generation and transfer, social networking, etc. There are, of course, many ways of accomplishing these social tasks. What makes a group of people performing such tasks a family? Let's imagine that according to Albert, a group of individuals engaged in the activities mentioned above constitutes a family$_A$ just in case it is a group consisting of a heterosexual couple that has been blessed by a priest living together (most of the time), and the children born of those parents (by the union of the parental gametes). *Family$_A$* is not a functional concept, that is, in order to count as a family$_A$, a group does not need to fulfill a certain social function. (One might say that families$_A$ are not grounded in functional role.[24]) Still, *the concept of family$_A$* might play a functional role in Albert's society. If this concept of family is dominant, that is, if families$_A$ are the primary site for love, intimacy, primary rituals around life and death, non-wage-based economic cooperation, and such, then this distinction will

[24] There is, I think, and important connection between Epstein's (2015) distinction between grounding and anchoring and my claims here about the distinction between the function of a concept and functional concepts. Concepts can have a function by virtue of anchoring facts without that function being part of the grounds for the category. But if we take the anchoring purposes to be what is essential to the concept (rather than the grounds), then grounds can adjust. I'm not sure I have this right, but it is worth pursuing, I think.

play a crucial role sustaining social coordination. Moreover, in such a society, non-married or same-sex couples, children born out of wedlock, adoptive families, and such will not be recognized as the proper site for recognized family activities.

When we consider whether Albert's concept of family is adequate, the task seems to be not simply whether it captures how things are, in fact, organized locally, but whether tracking, communicating, and coordinating around the distinction between families$_A$ and not-families$_A$ is a good idea. There is no question about whether Albert's distinction exists or whether we can track truths using the distinction. The question is whether we should collectively develop the capacities to notice that distinction, reinforce the distinction in law and policy, and structure society around it. Does this way of coordinating people domestically bring out the best in us? Does it oppress us? Does it make for a well-functioning society? Does it result in unnecessary suffering?

So, on one hand, there is a descriptive project of characterizing the possible ways of organizing domestic life, or the ways we do it (or have done it) around here. As anthropologists, we talk about different kinds of families, allowing that our structures of domestic life are one form among many. But on the other hand, there is a normative project. If the concept of family makes a difference to how we live—if we coordinate by developing capacities to register and respond to a particular distinction between domestic arrangements—then we should not just engage in description. We should evaluate how we have been thinking of families and decide how to go on. This is what I was earlier characterizing as a philosophical, as opposed to an anthropological, project. It involves, as the Schroeters suggest, an interpretation of our past practice (and the practices of others), but deciding how to interpret the past and how to project ourselves into the future involves normative considerations. (See also Schapiro 2003.)

We know that some ways of organizing family life lead to various kinds of social dysfunction, and others are immoral. So one normative axis of evaluation is functional: Because we process certain kinds of information mainly for the purpose of coordination, we can judge the adequacy of our ways of doing so by reference to how well the coordination works. Better and worse concepts of family might then be evaluated in terms of how well they achieve coordination in relation to the broader social context. Another axis of evaluation may be broadly moral/political. Perhaps Albert believes that there is something morally problematic about non-HBNF families, so we should organize domestic life by creating and supporting HBNFs; all other forms are somehow morally defective and so should be discouraged. On his view, only HBNFs are *real* families. He is not denying that we call many domestic arrangements 'families.' He would agree that the issue is how we should go on, but thinks that we are currently going in the wrong direction. He prefers a more narrow way of understanding what a family "really" is, for he believes that to be a family is to participate in those sets of practices that are (morally) legitimate, perhaps *even if* they do not promote local coordination. A group of individuals counts as a *family* only if they exemplify good ways of organizing domestic life.[25]

[25] One way to develop this idea would be to treat 'family' (and other social kinds) as *thick concepts*. However, whether we need to take the normative requirements to be part of the semantic value of the concept or whether it is pragmatic is a very controversial issue (see Vayrynen 2013). I will proceed here as if it is pragmatic, but I won't defend that position.

How is this related to semantic amelioration? Suppose what is essential to the concept of family is its role in social coordination, rather than its informational content, that is, the partition in logical space it tracks. The concept of family might be (roughly) the concept that we use to organize us in domestic arrangements. There are many kinds of family because there are many ways to go about such organization. But there are also better and worse concepts of family because there are better and worse forms of domestic organization. If I live in a situation in which Albert's concept of family is dominant, it is "our" concept of family there, then I would urge semantic amelioration, that is, I would demand that my family count as a family too. This would not be to demand that the concept of *family* change its role in our lives, but that it change its informational content. Given that, on the model I'm proposing, its role is essential and its content is not, this would be semantic amelioration—a change in the informational content—and not conceptual replacement.

To understand the tension between the descriptive and the normative aspects of amelioration, it may be useful to consider what Joshua Knobe and Sandeep Prasada call "dual character concepts" (2011; also Knobe et al. 2013, Leslie 2015). Note that some kind terms allow for a distinction between being a good exemplar of the kind and being a true exemplar. Usually these to evaluations go together. A *good* scientist is a true scientist; a true musician is a good musician. However, these evaluations seem to be based on connected but distinct criteria. They argue that there is a set of concepts—the dual character concepts—that

> ... are represented via both (a) a set of concrete features and (b) a set of abstract values that the concrete features are seen as realizing. These two representations are intrinsically related, but they are nonetheless distinct, and they can sometimes yield opposing verdicts about whether a particular object counts as a category member or not. (2011: 2965)

One of their paradigm examples concerns the concept of *scientist.* On the one hand, we might characterize a scientist in terms of the sorts of things they do, their qualifications, their job description, etc.; on the other hand, we might characterize a scientist (roughly) in terms of their intellectual virtues. They note that these two characterizations of a scientist might come apart (2011: 2965). One might say of a person who works in a lab but is dogmatic and does the minimum required each day:

(1) There is a sense in which she is clearly a scientist, but ultimately, if you think about what it really means to be a scientist, you would have to say that she is not a scientist at all.

Or one might say of a person not employed in a lab or academic institution and who "has never been trained in formal experimental methods but who approaches everything in life by systematically revising her beliefs in light of empirical evidence" (2011: 2965):

(2) There is a sense in which she is clearly not a scientist, but ultimately, if you think about what it really means to be a scientist, you would have to say that she truly is a scientist.

Other examples that scored high as dual character concepts include: "Friend, Criminal, Love, Mentor, Comedian, Minister, Theory, Boyfriend, Artist, Argument, Teacher, Poem, Soldier, Sculpture, Art Museum, Musician, Mother, Rock Music, Scientist, Novel" (Knobe et al. 2013: 256).

What is going on in these cases? Knobe et al. argue that there are two kinds of normativity playing a role in the case of dual character concepts. On one hand, we can evaluate the extent to which one meets certain conditions for being a member of the kind in question, for example, a *good* musician plays fluently, has advanced skills (even though they may have little creative spark). On the other hand, we can evaluate the extent to which one exemplifies certain abstract values, for example, a *true* musician lives to make music, and does so creatively and with passion (even though they may not have fabulous skills). The goodness vs. the trueness of a musician are different dimensions of evaluation. In the case of dual character concepts:

> ... people appeared to employ two distinct sets of criteria. When a given object met one set of criteria but not the other, participants tended to say that it was a category member in one sense but was not a category member in another sense. As such, the experiment provided evidence that dual character concepts provide two bases for categorization. (2013: 248)

This is what would be expected in the case of practices, generally. We must rely on practices to coordinate us, and in doing so they give us reasons to act and a basis for evaluation. However, practices themselves can be evaluated, both functionally and on other normative grounds.

I am not suggesting here that all concepts have their social function (rather than their informational content) essentially. That would be to take a stronger stand than I am prepared to endorse at this point. Knobe and Prasada suggest:

> We have seen that some concepts are unified through hidden causes (natural kind concepts) and others through abstract values (dual character concepts), but perhaps these are just two of the many possibilities, and there are also yet other kinds of concepts that are unified in quite different ways. For example, there might be concepts in which all of the concrete features are unified in that they all tend to make an object suitable for the same basic function (e.g., the concept COMPUTER). People might then associate these concepts with both (a) a list of concrete features and (b) the more abstract notion of the relevant function. (If so, such concepts would be like the dual character concepts studied here in that they would provide two bases for categorization, but they would be unlike dual character concepts in that they would not provide two bases for normative judgment.) (2013: 255)

A plausible hypothesis is that our concepts have different roles in mindedness.[26] Sometimes we categorize for the purpose of explanation, sometimes to capture what's of value, sometimes to identify a functional role (and as Knobe et al. suggest, possibly for other purposes as well). There can be controversy, then, over not only what conditions must be met in order to be included in a category, but also the point or purpose of the category, and what gives the category its unity. My suspicion is that inquiry into concepts sometimes rightly privileges the (paradigm) instances and

[26] I don't mean to suggest that we have and develop concepts for explicit and intended purposes; we don't. We are socialized into the local conceptual/linguistic scheme and no one designed it. I do think, however, that concepts play a functional role in systems of communication and coordination, and I have this sort of purposiveness in mind. See also Haslanger (forthcoming b).

allows our understanding of the purpose to adjust; and sometimes we are rightly invested in the purpose and reconsider the instances. Note, however, that the partition of logical space that actually serves the purposes of our distinction (once we figure out the purpose), may be very different from what it was before, based on our judgments of membership, or even based on our conception of what's at stake. (Baseball doesn't cease to be baseball when we change the rules, or "improve" the ball to allow more home runs so that it attracts more fans.[27]) This provides a basis for allowing ameliorative accounts to be, in some sense, a way of improving and not just replacing our concepts.

Let's return to Albert's concept of family as consisting of only HBNFs. It would seem likely that he would grant that, under current circumstances, same-sex families count as families "in some sense," for example, legally, but nevertheless aren't *true* families. That is because a broader practice of family formation that includes same-sex families (etc.) doesn't realize certain abstract values. This would allow him to say that families are understood as structures for coordinating domestic life, but only a subset count as "true" families, that is, those that do so in keeping with certain values that he supports. To capture the disagreement between Albert and the others in his milieu, then, we could see it as concerned with the values around which we coordinate. The concept of family is a focal point for such coordination. To determine what "true" families are we cannot simply consider our past practices, or past judgments of memberships, for our past practices may all be terrible. Normative discussion is needed in order to decide how to go on. The adequacy of an ameliorative proposal should be considered with respect to its prospects of either (a) disrupting our current unjust or dysfunctional practices, or (b) improving our practices should the revision become part of the cultural technē. These two options correspond to the two moments (negative and positive) of critical theory's work towards emancipation I mentioned at the end of section 2.

7. Conclusion

There is much more that needs to be discussed at this point. However, I've argued that there are two ways to think about conceptual amelioration on a model according to which the content of a concept is to be understood as informational content, that is, a partition of logical space. On the one hand, we can ameliorate our *understanding* of the relevant space of worlds—both what worlds it includes and how they should be sub-divided. Such improvement in our understanding may be based on advances in empirical knowledge and technological advances (think of war, water, and offspring created by artificial reproductive technologies), on a reinterpretation of the history of the concept, or on information gained by new perspectives that have access to the content from different modes of presentation.

[27] I am not assuming, in this example, that baseball is a historical particular (an institution), but as a type of game with instances. The concept of *baseball* partitions logical space into worlds in which there are baseball games and those in which there aren't (and along more fine-grained baseball facts). But which worlds have baseball games? If we change the rules, do we have a new concept that partitions things differently, or do we allow that the baseball partition changes? I prefer the latter option. Thanks to Ari Koslow for raising this issue.

On the other hand, we can ameliorate not simply by reinterpreting our past practice but by *correcting* it. It is not plausible to me that the "best" interpretation of our representational tradition concerning race is an interpretation according to which race is a social category. But, according to the Schroeters, this is what a social constructionist about race would be committed to. Nor am I convinced that *marriage* has always included in its informational content worlds in which there were same-sex marriages. I'm not even sure how those judgments would be adjudicated. But on my view, this doesn't matter. The terms 'marriage,' 'family,' and such, function in our culture to focus us on certain relationships as the "proper" basis of domestic life. To have the concept of marriage is to have the capacity to track certain kinds of information that are relevant in our current milieu to coordinate around domestic life and related social tasks. But the terms of coordination evolve, and we must be able to track that evolution with the term. More importantly, we can contribute to that evolution by challenging local ideas about what should be the "proper" basis for organizing domestic life. This is what happens in social movements, both radical and conservative, and through the work of critical theory. Disrupting the local cultural technē is difficult, because we are all deeply invested in maintaining the terms of coordination in our milieu, and going against the grain is costly. However, to ask, what is marriage, *really?* is to ask what forms of domestic partnerships (if any) promote a well-functioning and just society. When activists have claimed that same sex couples can be married, or that LGBTQ domestic arrangements are families, it wasn't based on what we have meant all along, but on what we should have meant. And what we should mean going forward, at least for now. A new answer, if it is incorporated into our practical consciousness, can be emancipatory and can change our social world, for we shape that world through what we do and who we think we are.

Some of our concepts are organized around values. Others are organized around functions. Some are organized around both. This is because we have an interest in carving logical space in order to coordinate with each other, to draw distinctions that serve our purposes as social beings and to realize our values. The best way to do this changes as we develop new technologies and as we come to appreciate new and different values. When social change happens, there is likely to be controversy and disagreement about how to extend the concepts we've been using to do the work we now need them to do. Such changes should be acknowledged as such, and should not be held hostage to what we have thought we were doing all along, and how to continue that. Our conceptual frameworks should be forward-looking and give us the tools to envision and create better lives together.

Acknowledgements

Thanks to Ekaterina Botchkina, Thomas Brouwer, Alexis Burgess, Alex Byrne, Herman Capellen, David Chalmers, Robin Dembroff, Esa Díaz-Leóna, Daniel Elstein, E. J. Green, Samia Hesni, Jerome Hodges, Mark Jago, Katharine Jenkins, Karen Jones, Ari Koslow, Dan López de Sa, Mari Mikkola, Deborah Mühlebach, Greg Restall, Mark Richard, Miriam Schoenfield, Laura Schroeter, Francois Schroeter, Daniel Stoljar, Tim Sundell, Pekka Väyrynen, Aness Webster, Robert Williams, Stephen Yablo, and Seth Yalcin for conversation and feedback on my ideas. Special thanks to David Plunkett, whose extensive and insightful comments on an earlier draft helped me improve the chapter considerably.

References

Alcoff, Linda Martín. 2015. *The Future of Whiteness*. New York: Polity Books.

Anderson, Elizabeth. 1995. Knowledge, Human Interests, and Objectivity in Feminist Epistemology. *Philosophical Topics* 23 (2):27–58.

Althusser, Louis. 1970. Ideology and Ideological State Apparatuses. https://www.marxists.org/reference/archive/althusser/1970/ideology.htm

Balkin, J. M. 1998. *Cultural Software: A Theory of Ideology*. New Haven: Yale University Press.

Baker, Katie J. M. 2013. We Need to Stop Obsessing about Walking Down the Aisle. *Jezebel*. https://jezebel.com/5979476/we-need-to-stop-obsessing-over-walking-down-the-aisle

Barnes, Elizabeth. 2016. *The Minority Body*. Oxford: Oxford University Press.

Bauer, Nancy. 2015. *How to Do Things with Pornography*. Cambridge, MA: Harvard University Press.

de Beauvoir, Simone. 1989/1949) *The Second Sex*. Trans. H.M. Parshley. New York: Vintage.

Bettcher, Talia. 2009. Trans Identities and First-person Authority. In Laurie Shrage (ed.), *You've Changed: Sex Reassignment and Personal Identity* (pp. 98–120). Oxford: Oxford University Press.

Bettcher, Talia. 2012. Trans Women and the Meaning of 'Woman'. In Nicholas Power, Raja Halwani, and Alan Soble (eds.), *Philosophy of Sex: Contemporary Readings* (6th edn) (pp. 233–50). New York: Rowan & Littlefield.

Bettcher, Talia. 2014. Feminist Perspectives on Trans Issues. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (spring edn). https://plato.stanford.edu/archives/spr2014/entries/feminism-trans/

Bettcher, Talia. 2016. *Precis of* and Comments on "Katharine Jenkins" Amelioration and Inclusion: Gender Identity and the Concept of Woman. *Pea Soup* January 27. http://peasoup.typepad.com/peasoup/2016/01/ethics-discussions-at-pea-soup-katharine-jenkins-amelioration-and-inclusion-gender-identity-and-the-.html

Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Cappelen, Herman. 2018. *Fixing Language*. Oxford: Oxford University Press.

Chodorow, Nancy. 1978. *The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender*. Berkeley: University of California Press.

Collins, Patricia Hill. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. New York: Routledge.

Crenshaw, Kimberlé. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43 (6):1241–99.

Dworkin, Ronald. 2011. *Justice for Hedgehogs*. Cambridge, MA: Harvard University Press.

Epstein, Brian. 2015. *The Ant Trap*. Oxford: Oxford University Press.

FindLaw. (n.d.). Legal Definition of 'Father' by State. https://family.findlaw.com/paternity/legal-definition-of-father-by-state.html (accessed July 1, 2018).

Frye, Marilyn. 1983. *The Politics of Reality*. Freedom, CA: The Crossing Press.

Frye, Marilyn. 1992. *Willful Virgin*. Freedom, CA: The Crossing Press.

Geuss, Raymond. 1981. *The Idea of a Critical Theory*. Cambridge: Cambridge University Press.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press.

Gilligan, Carol. 1982. *In a Different Voice: Women's Conception of Self and Morality*. Cambridge, MA: Harvard University Press.

Glasgow, Joshua, Haslanger, Sally, Jeffers, Chike, and Spencer, Quayshawn. 2019. *Four Views on Race*. Oxford: Oxford University Press.

Grice, H. Paul. 1957. Meaning. *The Philosophical Review* 66 (3):377–88.

Hall, Stuart. 1996/2006. The Problem of Ideology. In Kuan-Hsing Chen and David Morley (eds.), *Stuart Hall: Critical Dialogues in Cultural Studies* (pp. 24–45). New York: Routledge.

Hardimon, Michael. 2003. The Ordinary Concept of Race. *Journal of Philosophy* 100 (9): 437–55.

Hardimon, Michael. 2017. *Rethinking Race: The Case for Deflationary Realism*. Cambridge, MA: Harvard University Press.

Harris, Angela P. 1990. Race and Essentialism in Feminist Legal Theory. *Stanford Law Review* 42 (3):581–616.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs* 34 (1):31–55.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique.* Oxford: Oxford University Press.

Haslanger, Sally. forthcoming a. How Not to Change the Subject. In Teresa Marques and Åsa Wikforss (eds.) *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation.* Oxford: Oxford University Press.

Haslanger, Sally. forthcoming b. Cognition as a Social Skill. *Australian Philosophical Review.*

Herdt, Gilbert. (ed.) 1993. *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History.* New York: Zone Books.

Jackson, Frank. 2000. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.

Jaggar, Alison M. 1983. *Feminist Politics and Human Nature*. Lanham: Rowman and Littlefield.

Jeffers, Chike. 2013. The Cultural Theory of Race: Yet Another Look at Du Bois's "The Conservation of Races". *Ethics* 123 (April):403–26.

Jenkins, Katharine. 2016. Amelioration and Inclusion: Gender Identity and the Concept of *Woman*. *Ethics* 126 (2):394–421.

Kant, Immanuel. 1775/2000. Of the Different Races of Human Beings. In Robert Bernasconi and Tommy L. Lott (eds.), *The Idea of Race*. Indianapolis: Hackett.

Kapusta, Stephanie. 2016. Misgendering and Its Moral Contestability. *Hypatia* 31 (3): 502–19.

Knobe, Joshua, and Prasada, Sandeep. 2011. Dual Character Concepts. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Boston, MA: Cognitive Science Society.

Knobe, Joshua, Prasada, Sandeep, and Newman, G. E. 2013. Dual Character Concepts and the Normative Dimension of Conceptual Representation. *Cognition* 127 (2):242–57.

Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language.* Cambridge MA: Harvard University Press.

Lear, Jonathan. 1986. Transcendental Anthropology. In Philip Pettit and John McDowell (eds.), *Subject, Thought and Context* (pp. 267–98). Oxford: Oxford University Press.

Leslie, Sarah-Jane. 2015. 'Hillary Clinton is the only man in the Obama administration': Dual Character Concepts, Generics, and Gender. *Analytic Philosophy* 56 (2):111–41.

Machery, Edouard. 2009. *Doing Without Concepts*. Oxford: Oxford University Press.

MacKinnon, Catharine. 1982. Marxism, Method, and the State: An Agenda for Theory. *Signs* 7 (3):515–44.

MacKinnon, Catharine. 1987. *Feminism Unmodified.* Cambridge, MA: Harvard University Press.

MacKinnon, Catharine. 1989. *Towards a Feminist Theory of the State.* Cambridge, MA: Harvard University Press.

Mallon, Ron. 2004. Passing, Traveling and Reality: Social Constructionism and the Metaphysics of Race. *Noûs* 38 (4):644–73.

Mallon, Ron. 2006. Race: Normative, Not Metaphysical or Semantic. *Ethics* 116 (3):525–51.

Manne, Kate. 2017. *Down Girl: The Logic of Misogyny*. Oxford: Oxford University Press.

Marx, Karl. 1845. The German Ideology. https://www.marxists.org/archive/marx/works/1845/german-ideology/index.htm

Mills, Charles. 1997. *The Racial Contract*. Ithaca, NY: Cornell University Press.

Mohanty, Chandra Talpade. 1984. Under Western Eyes: Feminist Scholarship and Colonial Discourses. *boundary* 2 (3):333–58.

Omi, Michael, and Winant, Howard. 1994. *Racial Formation in the United States*. New York: Routledge.

Patterson, Te-Erica. 2014. 8 Reasons Why a Woman's Wedding Day Is Her Greatest Fantasy. *Elite Daily*. https://www.elitedaily.com/dating/8-reasons-why-a-womans-wedding-day-is-her-greatest-fantasy

Pérez Carballo, Alejandro. 2016. Structuring Logical Space. *Philosophy and Phenomenological Research* 92 (2):460–91.

Plunkett, David. 2011. Expressivism, Representation, and the Nature of Conceptual Analysis. *Philosophical Studies* 156:15–31.

Putnam, Hilary. 1975. The Meaning of "Meaning". In H. Putnam (ed.), *Mind, Language and Reality. Vol 2 of Philosophical Papers* (pp. 215–71). Cambridge: Cambridge University Press.

Putnam, Hilary. 1962/1975. The Analytic and the Synthetic. In H. Putnam (ed.), *Mind, Language and Reality. Vol 2 of Philosophical Papers* (pp. 33–69). Cambridge: Cambridge University Press.

Riley, Denise. 1988. *'Am I that Name?' Feminism and the Category of 'Women' in History*. Basingstoke: Palgrave-McMillan.

Saul, Jennifer. 2006. Philosophical Analysis and Social Kinds II. *Proceedings of the Aristotelian Society Supplementary Volumes* 80:119–43.

Schapiro, Tamar. 2003. Compliance, Complicity and the Nature of Nonideal Conditions. *Journal of Philosophy* 100 (7):329–55.

Schroeter, Laura, and Schroeter, Francois. 2015. Rationalizing Self-Interpretation. In Chris Daly (ed.), *The Palgrave Handbook of Philosophical Methods* (pp. 419–47). Basingstoke: Palgrave-Macmillan.

Stalnaker, Robert. 1998. What Might Nonconceptual Content Be? *Philosophical Issues* 9:339–52.

Taylor, Paul. 2004. *Race: A Philosophical Introduction*. Cambridge: Polity Press.

Vayrynen, Pekka. 2013. *The Lewd, the Rude, and the Nasty: A Study of Thick Concepts in Ethics*. Oxford: Oxford University Press.

Wittgenstein, Ludwig. 1958. *Philosophical Investigations*. Oxford: Basil Blackwell.

Wittig, Monique. 1981. One is Not Born a Woman. *Feminist Issues* 1 (2):47–54.

Wittig, Monique. 1993. *The Straight Mind*. New York: Basic Books.

Yablo, Stephen. 2014. *Aboutness*. Princeton: Princeton University Press.

Yalcin, Seth. 2016. Belief as Question Sensitive. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12330

Young, Iris Marion. 1990. *Throwing Like a Girl: And Other Essays in Feminist Philosophy and Social Theory*. Bloomington: Indiana University Press.

Zawidzki, Tadeusz. 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press.

13

The Theory–Theory Approach to Ethics

Frank Jackson

1. Methodological Preamble

As Wittgenstein noted, "naming something is like attaching a label to a thing".[1] But what are labels good for? They help us interact successfully with the things we have labelled. It is not for nothing that we give the front door key a different label from the back door key. What is more, it is part and parcel of these successful interactions that our labels ground information-preserving causal chains. That's how proper names help us arrive at the right cities when we travel. It explains, for example, why typing them into search engines or uttering them to travel agents produces the information we need. Reflections like these tell us that there has to be something essentially right about the causal theory of reference for proper names, though some (myself included) insist that we should understand it in its causal descriptivist guise.

Suppose we approached ethical terms in the same spirit, asking questions like: What do we do with them? What are they good for? How could we do something similar but do it better? This way of thinking changes the terms of the debate in ethics. Someone tells us, as it might be, that what's morally right is that which an idealized version of ourselves would resolve to do given full information about the available options. The questions to ask become ones like: What purpose might be served by giving people this information? and, Why would it be good to establish that some course of action is the one an idealized version of ourselves would resolve to do?

This essay will explore the implications of this way of thinking about ethical terms, concepts, and properties, and explain why a theory–theory story about them is so attractive when we approach matters from this perspective.[2]

2. Background Presumptions

One has to start somewhere and it can help to make explicit one's starting presumptions. Here are mine. (i) When someone affirms, "I believe that abortion is not always

[1] (Wittgenstein 1963: section 15).

[2] By the theory—theory I mean the view of, for example, Jackson (1992), Jackson and Pettit (1995); details are given below.

Frank Jackson, *The Theory–Theory Approach to Ethics* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Frank Jackson.
DOI: 10.1093/oso/9780198801856.003.0013

wrong", this should be taken at face value, as stating what they believe. And their belief is true just if it is the case that abortion is not always wrong, and that's the case just when the sentence "Abortion is not always wrong" is true. (ii) We marshal evidence for and against different views about what ought to be done. (iii) We change our minds about what we ought to do, sometimes as a result of the impact of new factual information (think, e.g., of those who have changed their minds about whether the water supply should be fluoridated) and sometimes as a result of a change in what's valued at some fundamental level (think, e.g., of those have who moved from being average utilitarians to being total utilitarians). (iv) We debate issues like, Does an act's being the morally best of those that are available entail that it is obligatory?

Taken together, considerations like these strongly support cognitivism in ethics, where by cognitivism I mean the view that ethical beliefs are beliefs properly speaking. For example, to believe that some course of action is morally wrong is to represent it as being a certain way, as having the property of being morally wrong, and the belief will be true just if it has the property in question. Moreover, a sentence like "Abortion is not always wrong" represents things as being a certain way, and will be true just when things are that way. Of course, strong support can be defeated, and some will offer one or another consideration to blunt the case for cognitivism. But I think that the overall case for cognitivism is more than strong enough to make it a reasonable starting point.

3. The Metaphysics of Ethical Properties

Cognitivism gives us license to talk about ethical properties in a metaphysically serious sense. Right acts differ from wrong acts in how they are in the same sort of way that tall people differ from short people in how they are, and the countryside in France differs from the countryside in England. Here we are not talking about properties in the sense of fundamental joints in nature or anything like that, and we are not talking about properties in the sense of universals.[3] We are talking about properties in the sense that is in play when people wonder *how much* the housing market has dropped, or which antibiotic is the *most effective*, and, more generally, are wondering about what things are *like*, their nature.

Cognitivism also means that we know how to talk about ethical properties: use ethical terms. If we are English speakers, we will use "morally right", "immoral", "evil" and so on. But do we have to use ethical terms? Is it compulsory?

Many insist that we do not have to use ethical or moral terms. We have other terms we can use instead. I will call this view naturalism. Sometimes the case for it is made by reminding us of the plausibility of the view that the kinds of properties that figure in the natural sciences, broadly construed, give us a complete picture of the properties to be found instantiated in our world. But then each and every instantiated property can be picked out in the terms of the natural sciences, and those terms do not contain ethical vocabulary. This means that—provided that ethical properties are

[3] To put matters in the terms of Armstrong (1978).

instantiated—each and every ethical property can be picked out without using ethical terms. Sometimes causal considerations are brought into the picture. We want ethical properties to figure in causal explanations, including explanations of our beliefs that one or another action possesses one or another moral property. How else could we know about them? And how else, given the plausibility of broadly causal approaches to content, could we have beliefs with the content that such and such an action is immoral, and that so and so a person is morally admirable, for examples? But we know enough about how our world works to know that the properties that figure in causal explanations are one and all properties we can pick out without recourse to ethical expressions. Sometimes supervenience enters the picture. The account of how things are given in ethical terms supervenes on the account given in non-ethical terms. How could this be the case, runs one argument in support of naturalism, unless the properties we pick out using ethical terms were the same properties (under different names, of course) that we pick out using non-ethical terms?

I have just stated naturalism (in ethics) in terms of language, as the view that we do not have to use ethical terms to pick out ethical properties, and reminded you of some familiar reasons for favouring naturalism, so understood. Why not put matters directly in terms of properties? For when you worry about what you ought to do, you aren't worrying about the words that apply to a proposed course of action. You are worrying about what properties it has (or so we cognitivists insist). True, if cognitivism is correct, the action will be what you ought to do just if "is what you ought to do" applies to it, or as we say it above, just if the property picked out by "is what you ought to do" is possessed by it. All the same, it is the property, and whether or not it is possessed, that concerns you, not the words *per se*. The wrongness or otherwise of abortion is not a question about words. The trouble with stating naturalism directly in terms of properties, is that one ends up saying something like: ethical properties are identical to non-ethical properties. This invites the thought that the view can be dismissed out of hand; its falsity comes from the meaning of the prefix "non-". It is, I think, safest to think of naturalism as holding that ethical properties can be picked out without using ethical expressions.

A view sometimes called soft naturalism[4] holds that ethical properties are one and all natural properties, but that the ethical cannot be analysed in non-ethical terms. Have we somehow failed to make a place for soft naturalism? No. Soft naturalism isn't the view that right acts, for example, have a property we cannot pick out using non-ethical terms. When soft naturalists explain their view as holding that that something like

Being right = maximizing expected utility

is true, but that being right, or that "is right", or that the concept RIGHTNESS, etc., cannot be analysed in non-ethical terms and, in particular, cannot be analysed in terms of maximizing expected utility, they are not saying that being right requires ethical terms to pick it out. They cannot be saying that. If being right is identical with maximizing expected utility, Leibnitz's Law tells us that there is a way to pick it out without using ethical expressions—use "maximizing expected utility". For "maximizing

[4] To use the terminology of Parfit (2011: chapter 25).

expected utility" picks out maximizing expected utility (obviously), and if being right is identical with maximizing expected utility, then "maximizing expected utility" picks out being right. Of course, there is an important distinction between versions of naturalism that hold that ethical terms, or concepts or . . . can be analysed in terms of non-ethical ones, and versions of naturalism that deny that any such analyses are possible. But all versions agree—they have to agree—that we do not need ethical terms to pick out ethical properties.

Should we go along with naturalism? Mooreans—those who insist that it is compulsory to use ethical terms if one wishes to pick out ethical properties, on the grounds that ethical properties are quite distinct from the kinds of properties talked of in the natural sciences—have replies to each of the considerations outlined earlier. But I think, as do many, that, taken together, the considerations make a very strong case against Mooreanism and for naturalism. Moreover, there is much dissension within the Moorean camp. I don't mean in the replies to the considerations sketched above (though there is dissension there also), I mean concerning the distribution of the ethical properties they insist cannot be picked out without using ethical terms. One should, I urge, be suspicious of claims to detect properties that are *sui generis* in the sense that they can only be picked out and explained in terms special to them, and whose causal explanatory roles with respect what happens in our world are, to say the least, unclear. But if those who claim to detect these special properties display an impressive degree of non-collusive agreement about the distribution of these special properties, sceptics have to sit up and take notice. It is, however, an only too familiar fact that there is great disagreement over the distribution of these allegedly special properties, especially over exactly which properties picked out in non-ethical terms ground one or another ethical property.[5] For example, although Mooreans agree that, by and large, killing humans is wrong, as do nearly all of us, they differ greatly over exactly why it is wrong—over which property of killing humans picked out in non-ethical terms makes this the case—as the debate over eating meat makes vivid. The same goes for keeping promises, causing pain, allowing people to make free choices and all the rest. The general agreement that they are, by and large: right, wrong, and right, respectively, masks huge differences over why they earn these verdicts. Of course, dissension is not limited to Mooreans, but it is a special problem for them. When a group of theorists claim: one, to detect properties that are opaque to science; two, are unable to say much directly about their nature (that's the cash value of saying that the properties are *sui generis*); three, grant that these special properties are grounded in ones that are transparent to science; while, four, being unable to agree among themselves about how the grounding works, surely scepticism about these properties is warranted.

4. Hunting for the Ethical after Moore

Turning one's back on Moorean thoughts about ethical properties—that is, embracing naturalism as specified above—has a profound effect on how we should think

[5] As Mackie (1977: 37) highlights.

about debates in ethics. This point has been noted but its full impact sometimes escapes explicit recognition, or so it seems to me. It can be very hard to think outside the Moorean box. The impact is that once one has a full inventory of all the properties we can pick out in non-ethical terms of some proposed course of action, or maybe all the properties anyone sensible could possibly think might be relevant, the question of whether or not the action is morally right is settled; there is nothing more to find out about the action. The perennially tempting thought that even when that full inventory is to hand, it is, in some good sense, a further question whether or not the action is morally right is a mistake.

Well it is not quite like this. The inclusive sense we are giving the notion of a property means that there is a sense in which one cannot have full inventory. As well as the properties picked out by, for example, "is in Australia" and "is causing pain", there are different properties picked out by "is in Australia and causing pain", "is in Australia or causing pain", "is causing pain and contains an odd number of electrons", and so on. There is, accordingly, no such thing as a full inventory in the sense of a very long but finite list that covers every property we can pick out in non-ethical terms. What we can have is a full inventory in the sense of a list of all the properties we can pick out in non-ethical terms that might plausibly be relevant (containing an odd number of electrons is an example of a property that is not plausibly relevant), along with a grasp of the various ways of combining those ingredients to form expressions that pick out further properties that might possibly be relevant, as happens when someone insists that what's crucial for being a value is not being desired but being what's desired to be desired, or someone else says that to be a right action is to be one which it is rational to desire everyone in similar circumstances perform, or some such.

When I say above that once one has a full inventory of all the properties we can pick out in non-ethical terms of some proposed course of action, the question of whether or not the action is morally right is settled, I do not mean by this that we should stop using words like "is morally right", "is evil", and so on. But the rationale for using them cannot be that they serve to pick out properties that have been left out of the inventory—there are no such properties to be picked out. The rationale for using them can only be that they pick out, from the properties already inventoried, properties that are best for some purpose or other, where "best", in this context, is—*of course*—itself to be understood in non-moral terms. To fail to see this last point would be precisely to slip back into the Moorean camp. But now a key question becomes how to understand this sense of "best". What does it take for some selection of properties we pick out using non-ethical terms to be the best ones to pick out using ethical terms?

5. Making a Start on What It Takes to be the Best Properties for Ethical Terms to Pick Out

Here is a simple example to set the scene. Physicists use the term "the centre of mass" to pick out an important property of a system of particles. There is a sense in which the term is not essential. They could have picked out the very same property using the terms for mass and position. Although having such and such a centre of mass is distinct both from having so and so a mass, and having this or that position, the terms for mass and position, suitably deployed, pick out the very same property as having such and such a centre of mass. That's what gets done when the centre of mass of a system of particles is specified in a physics text. But there is, of course, no mystery about the rationale for having the term "centre of mass" in physics, the value of the concept, and the importance of the property picked out: centres of mass play important predictive and explanatory roles.

What makes this example an easy one to think about (which is, of course, why I chose it) is that we know the relevant criterion for being a property that it would be good to pick out in physics and, thereby, a property it would be good to have a term for—it is being good for doing what physicists' do. Information about a system of particles' centre of mass is useful for predicting and explaining the system's behaviour over time, and predicting and explaining the behaviour of systems of particles over time is among what physicists do. What is far less clear is what it takes to be a good property to pick out in the ethics case. We know what physics is for, and that guides us in evaluating, say, $1/2Mv^2$ versus $1/2M^2v$, and explains why we give a tick to the first (when v is not too near c) and a cross to the second, and why we give a tick to centres of mass. But what is ethics for? That's the question we need an answer to if we are to find out which properties ethical terms pick out.

6. The Problem of Finding Purpose in Traditional Debates in Ethics

What ethics is for is a question about its purpose. But much of the debate in ethics does not look like a search for its purpose. It looks instead like a complex exercise in intuition swapping, combined with attempts to shift one's opponents' intuitions by noting one or another tension in their intuitive responses to different cases. Indeed, it does not just look like this; that's what much of it is. And some have, understandably, expressed concerns about this methodology. Do we do serious science by intuition swapping? Now, in fact and of course, intuitions do play a role in science and in theory building more generally. We perforce have to start our theorizing somewhere, and, naturally and properly, start from that which we find most intuitively appealing. The worry is the extent to which debate in ethics is driven by intuitions. A huge amount of it consists in painting pictures of possible situations using written and spoken words, and urging that one or another intuitively appealing response is indeed the correct response to one or another described situation. Sometimes the protagonists describe themselves as doing conceptual analysis; sometimes the exercise is given a more empirical flavour, as happens in the work of those who describe themselves as experimental philosophers. But it is intuition swapping all the same. The difference is that those doing conceptual analysis are more likely to insist that certain intuitions—their own, after due reflection—are the only defensible ones, whereas those doing experimental philosophy are more likely to want to collect intuitions from a range of subjects. But both are harvesting intuitions from word pictures, and building their theories on what gets harvested.

Is there a way of understanding the process I have just given a rough description of that makes it defensible and, moreover, fits with our leading idea that we should look for the properties it would be good to use ethical terms for—the idea we introduced in the first section by noting how a causal theory of reference for proper names drops out of asking what proper names are good for? Perhaps surprisingly, the answer to this question is yes, and the way to see this is to reflect on what we learn from a famous argument for analytical functionalism in the philosophy of mind, or so I will argue.

7. Analytical Functionalism

Analytical functionalism is a view about the meanings of certain words, words like "belief", "desire", and "feeing hungry". The view is that they are susceptible to an analysis in functional terms. But combined with this view about mental language is a position on the metaphysics of mental states. Sometimes the claim concerning the metaphysics of mental states is that they are functional states; sometimes the claim is that they are states (brain states, in fact) that get to be the mental states they are by virtue of the functional roles they play. The difference does not matter for our concerns. What matters for us is how the two positions, one about words and the other about the nature of mental states, are related. And here's a plausible answer to that question, or so it seems to me and many. When we describe human beings (and other animals) using mental state terms—"belief", "desire", and all the rest—something remarkable happens. Our ability to make successful predictions and explanations concerning their interactions with the environment and each other goes up dramatically. This strongly suggests that in using mental terms to describe humans, we are picking out properties humans in fact have. That's the obvious explanation of the utility of using these words. What properties might they be? Well, what's remarkable is, as we noted, the increase in predictive and explanatory power concerning certain interactions, and we know that functional properties are especially valuable for predicting and explaining interactions. This leads us to the hypothesis that mental words pick out functional properties. This leaves us with a choice between saying that mental states are the functional properties themselves, or are states having the functional properties in question. Either way, we have an explanation of why mental language is so useful, and one that connects facts about words with the metaphysics of mind.

The purpose of this reminder is to nail down the point that what happens when we use certain words can tell us a lot about the nature of that which we are using those words for, and that's something you can sign up to even if you dissent from one or another detail in the little example given above.

I hope the next question on my agenda will now be obvious. What do we learn from our use of ethical words, and how does this relate to the intuition swapping methodology of so much of the debate in ethics?

8. The Effects of Using Ethical Words

What happens when we describe matters in ethical terms? When dealing with psychopaths, nothing much. But when dealing with most of us, something very striking happens. Certain kinds of disputes get settled, or more nearly settled, when we use ethical terms. Suppose that Crusoe was alone on that island; in that case, he can do what he wants to do. He does not have to balance what he wants against what others want. But for most of us, a recurring issue is how to balance what I want against what you want. I want to see the doctor first, so do you. We cannot both get what we want. I want the roads near my house to be of good quality, you want the roads near your house to be of good quality. But there's only so much funding to go around; we cannot both get what we want. And so it goes. The fact that we live in communities and that resources are limited means that we cannot get exactly what we want, and sometimes cannot even get close to what we want. Fighting is one solution but its drawbacks are notorious. The solution that is often adopted is to describe matters in ethical terms. Sentences like "Your illness is more serious than mine (or you were here first), so the doctor ought to see you first" and "The road past my house services many more people than does the road past your house, so the bulk of the funding should go to it" start to get uttered. And when they do, the dust settles. I take it to be a commonplace that using ethical terms can help when dealing with clashes in wants. The exact story about the how and why of this might take many forms, but what matters for our purposes is that it happens. Framing issues in ethical terms helps settles hard decisions about the distributions of goods in the sense of the things we desire. I emphasize that many have said things like this. Here is a recent statement (Sterelny and Fraser 2017: abstract) "moral facts are facts about co-operation, and the conditions and practices that support or undermine it". Our concern is with the fact that using moral terms helps resolve conflicts; their concern is with moral facts and the right evolutionary story concerning them. But the song is the much same nevertheless. The words could hardly be so useful unless they picked out properties in the world, and the instantiation of these properties are the facts Sterelny and Fraser are talking about.

Although I am prescinding from the details of how using ethical terms helps resolve disputes and aids co-operation, we can say this much in the broad. Part of the process involves motivation. When matters get described in moral terms, it affects people's motivations. If that were not true, there would be no reason to use moral terms. When someone uses "is what ought to be done", or the like, of some proposed course of action, it is no accident that they tend to bring that course of action about. I want to see the doctor first. You want to see the doctor first. I learn the details of your situation, and come to a conclusion that I express in the words, "You ought to see the doctor first". I make way for you, and you feel good about my doing so. And so it goes.

We can now see how to make sense of what's going on when we swap all those intuitions, and do so in a way that makes the process a reasonable one. We are swapping intuitions about word pictures, as we say above. Why should we pay attention to these intuitions? Well, the words are our words; the intuitions are our first up opinions about the application of those words to those situations. (We don't use an "intuition meter"; we simply say what we believe about those situations using the words in question.) But our words would not be much use for describing the world if our opinions about when they do or do not apply had no weight. So, to turn to the ethics case, our intuitions are a guide to what properties our ethical terms pick out. What's more, as we have lately been outlining, we have good reason to hold that the properties our ethical terms pick out are good for resolving the kinds of conflicts that are part and parcel of being creatures that need to resolve conflicts in interests and, more generally, co-operate as members of communities. A ditty runs "Things go better with coke"®. Communities go better when issues get framed in ethical terms and the results of those framings are implementations of the actions that got described as ethically right.

9. How Ethical Terms Can Get to Make Things Go Better: A Bit More Detail

Using ethical terms makes things go better, especially in situations where a group of people cannot each get what they separately want, and in cases where entering co-operative arrangements have clear benefits. That's what we have just been saying. Let's now spell things out a bit more.

The way we use ethical terms can be divided into three categories.[6] One consists of connections between matters described in non-ethical terms and matters described in ethical terms. Killing is typically wrong. One ought as a rule to keep one's promises. If some action is impossible, it isn't something that ought to be done. Pain is bad. And so on. I will call these clauses input clauses. They get to be called input clauses because they can be framed as conditionals with non-ethical antecedents and ethical consequents. For example, the last illustration can be rephrased as: if x is pain, then x is bad; and of course the penultimate example is already in the form in question. If using ethical language is to make things go better, input clauses are vital. We need them to provide ethical discourse with a good part of its subject matter—the actions, motives, policies, etc., that are up for evaluation in ethical terms. We talk above of the intuition swapping that is so prominent a feature of debates in ethics. Much of the swapping concerns input clauses. The debate engendered by the famous trolley car problem is one illustration. That debate is essentially a parade of variations on the theme of killing one to save many, accompanied by invitations to assign one or another verdict framed in terms of what ought to be done to the variations, all in the hope of finding a consistent pattern in the verdicts, or, better, a consistent pattern in the considered verdicts.

A second category consists of inter-connections between matters described in ethical terms. Here are some rough examples. Rights imply duties to protect those rights. What ought to be done is what's best out of the available options. You ought to do whatever is a necessary condition for what you ought to do. I call these clauses internal clauses. They are a vital part of the way we reason about connections between matters described in ethical terms. A simple example is what happens when we move from matters described in terms of what is good or bad to what we ought to do, maybe balancing the extent to which we have special moral obligations towards those close to us. Internal clauses provoke their share of the intuition swapping that dominates debates in ethics. Think of the intuition swapping that

[6] What follows draws on, for example, Jackson (1992 and 1998).

goes on in discussions over how to balance obligations to those we are close to against a general duty of beneficence. Or, again, think of the intuition swapping that goes on in discussions of whether or not an obligation to do A and B implies an obligation to do A, and in discussions of how being morally permissible relates to being morally obligatory.

Finally, there are output clauses. They connect people's beliefs framed in ethical terms with what they are motivated to do. They are every bit as much a subject of intuition swapping as are the input and internal clauses. Some find it intuitively compelling that someone who believes that they ought to do A will be motivated to some extent to do A. Dissenters present counter-examples—subjects who agree that A ought to be done but show no inclination whatever to do A. Dissenters to the dissenters respond by urging that these cases are one's of subjects who are only agreeing in "words" that A ought to be done; they haven't taken on board what believing that A ought to be done really amounts to. But, however this debate pans out, it had better be the case that, very often, describing matters in ethical terms as makes a difference to behaviour, as we noted in an earlier section.

I hope that what I have just been saying sounds pretty commonsensical. How could using ethical terms have good effects if there were no agreement about which actions, policies, characters, etc., are apt for description in ethical terms, no agreement about how to reason using ethical terms, and no agreement about what often happens when someone uses one or another ethical term to describe an action, policy, etc.? We need those input, internal, and output clauses. But insisting that there must be some agreement is not the same as insisting that there must be great agreement, and there isn't. It is an only too familiar fact that the intuition swapping we have mentioned a number of times has not led to widespread agreement on a core set of precisely specified input, internal, and output clauses. There is a reasonable amount of agreement in the rough, but not once we seek precise statements of the clauses.

10. How to Respond to the Disagreement

Someone may say, "So what", when asked about the disagreement. Perhaps they say that they have just published a book in ethics and in that book they nail down exactly why it is typically wrong to kill, and exactly what the exception clauses look like, and likewise for the other input clauses. They also insist that they capture the logic of ethics in the sense of the logical relations that hold between being right and being good, between being morally permissible and being right or wrong, between an obligation to do A and B and an obligation to do A, and so on. There is also a chapter on the connection between moral judgement and motivation, where, our author claims, that perennial issue is sorted once and for all. Perhaps they finish up by saying that they are sorry that there is so much dissension among theorists in ethics but hope that their book will reduce this and, anyway, when was philosophy an exercise in counting heads?

But how did our author make their case in their book? With experiments? Not in the sense of experiments in science. Finding the one true set of input clauses, for example, is not like finding that gold resists corrosion, or that light is a first signal. And the same goes for the internal and output clauses. With experiments in the sense experimental philosophers sometimes have in mind—that is, by garnering responses to vignettes and reflecting on those responses? But that's a version of intuition swapping, and the failure of this method to deliver consensus is exactly what has led to the discussion we are now having. By finding a new way to probe the true nature of the special properties that are picked out by ethical terms? No; to say that is to forget the implication of turning one's back on Mooreanism. There are no such special properties. By appeal to the causal theory of reference, arguing that careful investigation of the causal origins of our use of ethical terms will reveal their referents and allow us to read off from these discoveries the one true set of (precisely specified) input, internal, and output clauses? But the right theory of reference for ethical terms has to be one or another kind of description theory, where by a description theory for a word "W", I mean a theory that says that "W" applies to x just if x has a certain nature, that is, instantiates a certain property.[7] Why do I say this? Because of the point we have adverted to already. People who worry about the morality of abortion are not worrying about words. They are worrying about the kind of action abortion is; they are worrying, that is, about abortion's nature, whether or not it has a certain property. We perforce use words to discuss ethical questions, but these words need to pick out properties if our words are to engage with what matters ethically. I know non-cognitivists will object, but we cognitivists insist that the job of ethical terms is to describe, and that our ability to address what's really at issue in debates in ethics depends on this fact. When we produce sentences like "Abortion is sometimes morally right" and invite discussion, a necessary condition for what follows to be relevant is that "is sometimes morally right" applies to abortion if and only if it has the very property people care about when they debate the morality of abortion.

How then might our imagined author make their case? And this is, of course, a pressing question for real authors of works in ethics. The intuition-driven nature of so much of the debate in ethics, combined with the diversity in intuitions, raises a serious question as to how any author might sensibly hold that they'd got it right. I think that there are only two ways our author can respond.

The first is to remind us of the hope many of us had when we first heard about the trolley problem. It is the hope that drives players of what has come to be called "Trolleyology". It is the hope of finding a set of consistent and compelling responses to the original problem and its many variants. By consistent, I do not mean logically consistent. I mean that the set no-where contains different responses to, let's say, version 3 and version 5 of the problem, where the only difference between the two versions is a property that no-one thinks is morally relevant. By compelling, I mean that the set is such that anyone who understands the various versions, and does not themselves fall into inconsistency in the sense just explained, will come to share the responses. So, the first way for our author to respond is to urge that when people reflect on what their book has to say, something remarkable will happen. There will be a massive reduction in the extent to which intuitions about ethics' famous problem cases diverge, and that, although there will be holdouts (that's how it is in

[7] In *this* sense, a description theory of reference is true for the word "circle" in English. It applies to x just if x has a certain geometric property.

philosophy), it will be plausible that the holdouts are in some way confused or in some sense mean something different from what most of us mean when we use ethical terms.

Some will say, "Good luck with that". I am not that pessimistic. As someone sympathetic to consequentialism, I have always taken the following line of argument seriously. Make the hypothesis that being right = maximizing expected utility, and see what ensues. What ensures, as we all know, are a range of verdicts about what's right that clash with commonsense morality. Or at least they seem to, but, consequentialists urge, on examination matters are not so clear-cut. That's the burden of, for example, Smart (1961) and Kagan (1982). So one way to go in response to the divergence in intuitions is to urge that, after due reflection, the detection of confused thinking and so on, the divergence will disappear.[8] But, nowadays, my money is more on the second way to go, and here I am influenced by Alexander (1891–2), the paper by Sterelny and Fraser mentioned earlier, and, most importantly, by the examples of proper names and mental state terms.

We noted that it is matter of record that describing matters in ethical terms makes things go better, especially when dealing with problems that arise from the fact that we live in communities and need to adjudicate between competing claims. It is very hard to believe that this is an accident, as we noted. But if it isn't an accident, there must be a story to tell about the properties our ethical terms are picking out, which explains this happy result. The situation is akin to that with names and mental state terms. It is a matter of record that the way we use names helps us find our way to conferences, etc. It is a matter of record that the way we use mental state terms helps us explain and predict what people will do. Both facts call for explanation. In first case, the explanation will advert to some version or other of the causal theory of reference for names.[9] In the second, the explanation will advert to some version or other of functionalism, or so I suggested.

What's the right story in the case of ethical terms? We can say this much, based on our earlier remarks. The story will have input clauses, internal clauses, and output clauses. We saw above that they are essential if moral terms are to be of use to us in negotiating our interactions with others in our communities. The task then is to find the best candidates to be the properties picked out by our moral terms to make true the input clauses, internal clauses, and output clauses in a way that best explains how they—our use of moral terms by virtue of the properties they pick out—make things go better in our interactions with our fellows. But, I hasten to add, properties first identified via their role in facilitating relatively amicable negotiations within a community can and will be possessed outside that context. Nothing I say here implies, for example, that we lack moral obligations towards those outside our own community.

Finally, this way of looking at things does not preclude the consequentialist's answer. Maybe taking "is right" to pick out maximizing expected utility, along with

[8] That is, the identity of being right with maximizing expected utility will drop out of *mature* folk morality, in the terms of Jackson (1998: 133).

[9] In my view (as I said earlier), the version known as causal descriptivism. See Kroon (1987) and for the particular version I favour Jackson (2010: lecture five).

the corresponding choices for what “is morally good”, etc., pick out, deliver the best explanation. But maybe what does the best job is taking “is right” to pick out that which an idealized, fully rational version of ourselves would resolve to do given full information about the available options, along with the corresponding choices for what “is morally good”, etc., pick out.[10] Or maybe what does the best job is taking “is wrong” to pick out actions which violate principles for the regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement, along with the corresponding choices for “is permissible” etc.[11] Or maybe. . . . I deliberately leave this key question open.

Acknowledgments

Thanks to the many who have discussed ethics with me over the years, with special thanks to David Braddon-Mitchell, Philip Pettit, David Plunkett, Peter Railton, and Michael Smith.

References

Alexander, S. 1891–2. Is the Distinction between ‘Is’ and ‘Ought’ Ultimate and Irreducible? *Proceedings of the Aristotelian Society* 2 (1):100–7.

Armstrong, D. M. 1978. *A Theory of Universals*. Cambridge: Cambridge University Press.

Jackson, Frank. 1992. Critical Notice of Susan Hurley: *Natural Reasons*. *Australasian Journal of Philosophy* 70 (4):475–87.

Jackson, Frank. 1998. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jackson, Frank. 2010. *Language, Names, and Information*. Oxford: Wiley-Blackwell.

Jackson, Frank, and Pettit, Philip. 1995. Moral Functionalism and Moral Motivation. *Philosophical Quarterly* 45:20–40.

Kagan, Shelly. 1982. *The Limits of Morality*. Oxford: Oxford University Press.

Kroon, Fred. 1987. Causal Descriptivism. *Australasian Journal of Philosophy* 65 (1):1–17.

Makie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Harmonsworth: Penguin.

Parfit, Derek. 2011. *On What Matters*, vols 1 and 2. Oxford: Oxford University Press.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Smart, J. J. C. 1961. *An Outline of a System of Utilitarian Ethics*. Melbourne: Melbourne University Press.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

Sterelny, K. and Fraser, B. 2017. Evolution and Moral Realism. *British Journal for the Philosophy of Science* 68 (4): 981–1006.

Wittgenstein, L. 1963. *Philosophical Investigations* (2nd edn). Oxford: Blackwell.

[10] To draw on Smith (1994).

[11] To draw on Scanlon (1998).

14

Conceptual Ethics and the Methodology of Normative Inquiry

Tristram McPherson and David Plunkett

Introduction

One of the striking features of normative theorizing in philosophy (as well as in related fields, such as political theory) is the diversity of concepts that feature centrally in it. In particular, it is commonplace for different theorists to offer different glosses on the overarching normative questions they are interested in, using what appear to be distinct concepts.

Consider, for example, broadly *ethical* inquiry. Suppose we focus our attention on a specific agent in a completely determinate set of circumstances. Normative inquirers ask:

- what the agent *ought to do, all things considered*
- what the agent *has most normative reason* to do
- what it would be *immoral* for her to do
- what it would be *rational* for her to do
- whether one of her options would constitute *injustice, or exploitation, or betrayal*
- etc.

Something similar holds for broadly *epistemic* inquiry, where inquirers ask, of a given agent, in a given set of circumstances:

- what she *ought all things considered to believe*
- what she *knows*
- what she has *adequate epistemic justification* to believe
- what it is *rational* for her to believe
- whether her beliefs have been formed in an *epistemically responsible* way
- etc.

A similar variety of questions are posed by inquirers in aesthetics and political philosophy. In all of these cases, normative inquirers frame their investigations in terms of a range of (what appear to be) different normative concepts.

Even if we bracket the idea that normative inquirers *do* frame their investigations using different normative concepts, it is very plausible that they *could*. And given this possibility, normative inquirers face an interesting set of questions concerning what

Tristram McPherson and David Plunkett, *Conceptual Ethics and the Methodology of Normative Inquiry* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020).

DOI: 10.1093/oso/9780198801856.003.0014

normative concepts they *should* use. This is a central part of what we call the *conceptual ethics* of normative inquiry. We take *conceptual ethics* to encompass a range of issues about the normative and evaluative assessment of concepts and words. If we focus on a specific agent in a specific context, two central questions in conceptual ethics are: *which concepts* that agent should use, and *which words* she should use to express those concepts.[1]

The aim of this chapter is to explore two main questions in the conceptual ethics of normative inquiry. The first question concerns whether to orient one's normative inquiry around *folk* normative concepts or around *theoretical* normative concepts. For example, KNOWLEDGE and IMMORAL are arguably folk normative concepts: non-philosophers engage in a rich range of thought and talk about what various people do and do not know, and when and why they have been immoral.[2] By contrast, ADEQUATE EPISTEMIC JUSTIFICATION and PRO TANTO PRACTICAL REASON are arguably technical concepts that have their home in a range of theoretical activities, most markedly systematic epistemic and ethical investigation.

The second question that we explore is whether to orient one's normative inquiry around concepts whose *normative authority* is especially accessible to us, or around concepts whose *extension* is especially accessible to us. For example, the normative authority of OUGHT TO DO ALL THINGS CONSIDERED might seem especially clear, while its extension will seem intensely controversial in many important cases. By contrast, it is relatively clearer what falls in the extension of BETRAYAL, but more controversial how normatively significant betrayal is in many important cases.

In this chapter we do not aim to adjudicate these two questions in conceptual ethics. Instead, we have two central goals. First, we aim to make vivid a range of possible positions that one might occupy with respect to these questions. Attention to this range of options can be valuable for properly grasping the relationships between many existing normative theories, and methodological approaches to normative inquiry. It can also help to reveal approaches to normative inquiry that might otherwise remain obscure. We often survey these options in a relatively neutral manner. However, we do not mean to imply that all of these positions are ultimately equally good. With this in mind, our second goal is to highlight a range of schematic arguments favoring certain options over alternatives. In so doing, we hope to contribute to the long-term goal of adjudicating among those options.[3]

[1] We here draw from Burgess and Plunkett (2013a) and (2013b). As emphasized in Burgess and Plunkett (2013a), in calling this area of inquiry conceptual "ethics", we do not mean to privilege the idea that broadly moral/political norms matter here more than others. It could be, for example, that we should (at least in certain contexts) use those concepts that best carve reality at its objective joints, regardless of any broadly moral/political norms (e.g., regardless of whether it makes our lives go better or worse).

[2] In this chapter, terms in small caps (e.g. CAT) pick out concepts. Single quotation marks (e.g. 'cat') are used strictly to mention linguistic items. Double quotation marks (e.g. "cat") are used for a variety of tasks including quoting others' words, scare quotes, and mixes of use and mention.

[3] For simplicity, we tend to frame our questions *individualistically*. But normative inquiry is usually a collective endeavor. This raises two important possibilities that we do not discuss in the chapter: (i) that it might be best overall if some groups of individuals oriented their inquiry around one normative concept, while others oriented their inquiry around another; and (ii) that what concept it makes sense for an individual to orient her inquiry around might depend in important ways on what other inquirers are doing.

The question of whether to orient inquiry around folk or theoretical concepts arises in many different kinds of inquiry. By contrast, the question of whether to orient inquiry around *authority-revealing* or *extension-revealing* concepts lacks obvious parallels in other areas of inquiry. As we aim to show, these two issues can interact in interesting ways, making it fruitful to explore them together.

We organize our discussion into four sections. Section 1 focuses on the choice between folk and theoretical normative concepts. Section 2 introduces questions about which words to use to express the folk or theoretical concepts we deploy. Section 3 concerns the choice between authority-revealing and extension-revealing normative concepts. These discussions explore arguments *within* the conceptual ethics of normative inquiry. Section 4 steps back from this focus, to a more theoretical question about the very practice of engaging in reflection about conceptual ethics. In particular, it explores the question of how we should understand—as well as how we should *choose*—the normative standards and concepts that we deploy when engaging in conceptual ethics.

1. Folk vs. Theoretical Concepts

This section concerns the question of whether normative inquiry should be oriented around folk concepts or theoretical concepts. For example, in theorizing about norms for action, should we focus on trying to understand which actions satisfy the folk concept MORAL RIGHTNESS or instead Allan Gibbard's theoretical concept THE THING TO DO?[4] We canvass several reasons to favor using folk normative concepts, and theoretical normative concepts, respectively. Before proceeding, however, we first clarify several key assumptions that guide our discussion.

First, the issue we are focusing on here is not whether to *use* folk or theoretical concepts in normative inquiry. We take it that we will likely want to use both such concepts in any reasonably developed inquiry. The question is rather: which such concepts should we use to *state* our central theoretical questions, and the answers to those questions that we seek? When a concept plays both of these roles in an inquiry, we will say that the inquiry is *oriented* around that concept.

Second, we are investigating a question in the conceptual ethics of *normative inquiry*. As we shall understand it, this encompasses inquiry into what agents should do, think, or feel, and into what should or ought to be the case. We also take it to encompass inquiry into *evaluative* questions, concerning what is good or bad, valuable or disvaluable; *deontic* questions, concerning what is permissible, required, or forbidden; and *aretaic* questions, concerning what is virtuous or vicious. Further, for our purposes here, we also treat it as encompassing inquiry oriented around concepts such as COURAGEOUS or POLITE, whose status as normative is more controversial than the sets of questions just canvassed.[5] So understood, normative inquiry occurs in many parts of philosophy, including in ethics, political philosophy, epistemology, aesthetics, legal philosophy, and the philosophy of science, as well as

[4] For Gibbard's articulation of the concept THE THING TO DO, see Gibbard (2003).

[5] See Väyrynen (2013) for an important argument for the idea that evaluation might be only pragmatically implicated by use of thick concepts like COURAGEOUS.

outside of philosophy (e.g., in debates about public policy). Given that conceptual ethics is itself a normative topic, this chapter itself involves normative inquiry. (This is something that raises complications that we will return to in section 4.)

Questions of how to individuate concepts and words are highly relevant to the foundations of conceptual ethics.[6] We cannot, however, responsibly argue for answers to these questions here. Instead we aim to clarify our own usage.

We will think of *concepts* as components of thought, and *words* as linguistic vehicles. Somewhat more specifically, we will take concepts to be individuated by something like their inferential role at the level of thought (for suitable concepts, we will assume that inferential roles determine intensions). And we will take concept-users' inferential dispositions to provide significant (although defeasible) evidence of the inferential roles of the concepts tokened. Turning to words: we will not take words to be individuated by their meanings. For example, we will allow that a single word 'bank' can have one meaning that is about the sides of a river, and another that is about a kind of financial institution. And we will understand the meanings of words (perhaps only at an instance of use) in terms of the concepts they are used to express.

These are regimenting assumptions intended to simplify discussion; much of what we say could be straightforwardly reframed given different assumptions on these matters. For example: if you think that there are two different English words written as 'bank', everything we say could be re-phrased in terms of groups of homophonous but distinct words. Another, perhaps more important example, concerns our use of the term 'concept'. For much of our discussion, what will really matter to us are patterns of inference that are closely associated with the use of a given word (e.g., the word 'moral'). This is because it is these patterns of inference that will be most significant to the effects of the use of a given word in the context of normative inquiry. When we treat these divergent patterns of inference as evidence for the presence of distinct concepts, much of what we want to say could instead be represented in terms of multiple systematic patterns of inference associated with the same concept.

We assume the following rough distinction between *folk* and *theoretical* concepts. Theoretical concepts have their home within a community dedicated to a certain relevant inquiry, and are used for the purposes of that inquiry. For example, consider the concepts one acquires when learning advanced physics, or advanced linguistics. Folk concepts have their home in thought and talk beyond the specialists in a given field of systematic inquiry. These characterizations surely admit of hard and borderline cases. However, they are often clear enough. For example, inquirers sometimes find it useful in their work to deploy the folk concept STRING (as in: "we tied the samples together with some string"), and the folk sometimes use the theoretical concept STRING THEORY. In these cases, we have no difficulty determining which side of the folk/theoretical divide is the natural home of the relevant concept.

[6] For further discussion, see Burgess and Plunkett (2013b); Cappelen (2018); Braddon-Mitchell (Chapter 4, this volume); Greenough (Chapter 11, this volume); and Haslanger (Chapter 12, this volume).

It is also worth emphasizing that theoretical concepts are diverse. One standard function of distinctively theoretical concepts is to provide manifest precision: it can be theoretically useful to offer stipulations, explications, operative definitions, etc., in order to allow the assessment and transmission of relatively precise theses. Other theoretical concepts are introduced to allow us to aptly discuss and investigate certain important worldy phenomena. Where a worldy phenomenon is imperfectly understood, we may need to choose between explicit precision and tracking the important worldy phenomenon. In light of this, the virtues and vices that we explore below may not apply to all theoretical concepts.[7]

We now examine two classes of reasons to favor orienting normative inquiry around folk concepts. The first is motivated by the idea that it makes sense to inquire into questions that we care about. Consider the questions *do I know anything?* or *is it morally wrong to eat meat?*, as well as many other questions that lead people to become interested in normative inquiry in the first place. It is natural to think that such questions are framed using folk concepts. Given this, if one instead orients one's normative inquiry using theoretical concepts, there is a danger that one will simply have *changed the subject*, and failed to address the question one cared about in the first place.[8]

Notice that this argument starts with the issue of how the *questions* we address in inquiry are framed, rather than the materials that we use to answer those questions. Thus, it is compatible with using highly theoretical concepts in one's *answer* to the question of whether it is wrong to eat meat. However, many who are sympathetic to this line of thinking hold that the answers to questions must also be framed at least partly in terms of the same folk concepts (e.g. KNOWLEDGE, MORALLY WRONG) deployed in the question itself, in order to be an apt answer to *that* question.[9]

This sort of argument suggests an important concern to be attended to. However, there are at least two reasons to be cautious about its force. First, on at least some natural ways of individuating folk concepts, a folk concept may include elements that you do not care about in a given instance when you are deploying that concept. Consider the following example. Carrie Jenkins motivates her exploration of flirting by pointing to the practical significance of judgments about flirting in our lives. She then goes on to offer an account of flirting that proposes (among other things) that the flirter must proceed in a *playful* manner.[10] Suppose that playfulness is part of the folk concept FLIRTING. In many contexts, this aspect of the concept may be irrelevant to what people care about when they deploy that concept. For example, if two people are behaving in ways disposed to raise the salience of romance between them, the question of whether they are doing so in a playful manner might well not be practically significant to them, or to other interested parties. In these contexts, a

[7] The idea that there is an interesting distinction between folk and theoretical *concepts* might be controversial, especially given certain theories of content. For example, on some views, the content of folk natural kind concepts might be fixed by facts about expert usage. (See, for example, Burge 1979 and Schroeter 2008). But recall how we are using the term 'concept': we think that versions of many of the points here could be reformulated within the sort of theory of content just mentioned, by focusing on contrasting inferential implications of certain uses of words.

[8] See Strawson (1963) and Jackson (1998) for articulations of this kind of worry.

[9] See, again, both Strawson (1963) and Jackson (1998).

[10] Jenkins (2006).

person may frame much of her thinking and discussion about this interaction in terms of FLIRTING despite the fact that this concept includes elements that are irrelevant to her interests in this case. She might do so simply because this is the most useful concept for her to deploy, among those concepts readily available in her current conceptual repertoire.

This example brings out a more general issue. One reason why someone, in a given context, might not care about certain aspects of a concept's precise inferential role is that she was primarily interested in discussing a certain *property*, which one could (at least in principle) think about using different concepts. It might be that one's interest in a certain folk concept is explained by the fact that it enables one to discuss or think about that property in a given context. Suppose that we can make sense of goals of inquiry identified at the object-level (e.g., in terms of properties or facts) rather than at the representational-level (e.g., in terms of concepts). This would allow for us to change which concepts we use without a "change in topic" in the sense relevant for that inquiry.[11]

A further reason to be cautious about this reason to favor folk concepts is that inquiry itself may *change* what you care about. For example, Carlos might start out caring a great deal about MORAL RIGHTNESS. But inquiry might change this attitude: for example, he might become convinced by an account of the history of this folk concept that destroys his interest in deploying it in normative inquiry.[12] One simple example: he might conclude that MORAL RIGHTNESS is an ideological instrument for the ruling class, and that this fact infects either the intension of this concept or its fruitfulness as a locus for inquiry.[13] This might then in turn motivate him to introduce a theoretical concept in the vicinity of MORAL RIGHTNESS that answers to his *theoretically informed* interests.[14] In general terms, inquiry might reveal that our folk concepts are *defective* in a range of different ways. For example, they might lead us into irresolvable paradoxes.[15] At the same time, inquiry might reveal that we might be able to replace those concepts with others, which preserve the core of what we should care about in the vicinity.

Theoretical inquiry may be misleading some or even much of the time. But, at least in principle, it seems fully capable of providing or revealing reasons for investigating properties that are not picked out by our folk concepts. This is particularly hard to deny given that the line of reasoning above was not exclusively focused on normative

[11] For further discussion about preserving "sameness of topic" despite shifts in concepts, see Cappelen (2018) and Thomasson (Chapter 21, this volume).

[12] See Plunkett (2016) for relevant discussion of the significance of conceptual history for the conceptual ethics of normative inquiry.

[13] For a brief discussion of this sort of broadly Marxist-inspired take on morality, see Sinnott-Armstrong (2006: 208). For helpful critical discussion of Marx's own (more subtle) views on morality, see Wood (1999).

[14] Notice that on some sophisticated theories of folk concepts (e.g., Jackson 1998), the content of our folk concepts might turn out not to vindicate all of ordinary speakers' inferential dispositions for using those concepts. On this sort of picture, even if folk *inferential dispositions* for using a concept are objectionable in some way, the intension of the concept might not be objectionable in this way. Jackson also grants that sometimes theoretical reflection can reasonably lead us to care about something other than the folk concept we start with Jackson (1998: 44–5).

[15] For a clear example of this line of thought with respect to the concept TRUTH, see Scharp (2013).

inquiry as such, or even philosophical inquiry as such. As the toy example of conceptual history above underscores, what we care about in normative inquiry could be affected by *non-normative* theoretical inquiry from areas outside of philosophy. For example, discoveries in sociology, psychology, or climate science might create a context where certain normative questions become especially salient and pressing.

A second reason for orienting inquiry around folk concepts arises from the idea that our ability to learn about certain normative facts might be mediated by trained cognitive capacities, which may be (a) largely implicit and (b) largely accessible via deployment of our folk concepts. For example, decently-raised children internalize a torrent of context-relevant guidance about which acts and arrangements are *fair*. Suppose that such guidance is crucial to building the competence of inquirers to investigate certain broad normative topics. Then seeking to replace such concepts with precise theoretical alternatives risks depriving the inquirer of the ability to use her implicit knowledge.

Similarly, suppose that one's aim in developing a theoretical concept is to capture more precisely what we most care about in the relevant topic. In many cases, it can prove difficult to understand precisely *what* we care about in the topic at hand, let alone make it fully explicit. Given this, there is a substantial danger that in shifting our focus by using a new or "reformed" concept, we will lose the ability to focus on what we cared about in deploying the folk concept—some of which, of course, might also be things we *should* care about. This might in turn lead us to miss important aspects of the normative issues at hand. In short, the switch to theoretical concepts might leave us with worse tools for investigating our normative topics at hand (even if such a switch still preserved "sameness of topic" in relevant ways). Indeed, we might end up with worse tools here without even fully realizing that this is so; something which might well make the shift to the new "reformed" concepts particularly dangerous.[16]

Tied to this worry is the question of how good we are at judging the merits and dangers of attempts to depart from conceptual folkways. Perhaps, as a general matter, one takes a dim view of our abilities here. This might then be combined with a strong version of the thesis that our ability to learn about key normative facts is mediated by trained cognitive capacities, which are largely implicit and largely accessible via deployment of our folk concepts. This combination of views might be used to support a kind of intellectual analogue of Burkean conservatism in social/political philosophy, according to which we should largely defer to our current conceptual repertoire, or at least let it change slowly and gradually and be highly skeptical of attempts to radically change it quickly with "conceptual revolutions".[17]

The fact that folk concepts are central sites of social learning, however, can also provide grounds for caution about orienting one's normative inquiry around folk concepts. For the influences that shape our implicit grasp of normative concepts need not tend towards accuracy or reliability. For example, moral and political concepts, from IMMORAL to JUST to CHASTE, are exceptionally natural targets for

[16] This sort of criticism is prominent in Velleman's (1988) criticism of Brandt's reforming definition of 'good' (as in Brandt 1979/1998).

[17] Compare Burke (1790/1982).

ideological forces. Such ideological contamination might render our intuitive grasp of the relevant concept unreliable, even if it does not infect the inferential role of the concept itself.[18]

Next consider two types of reason supporting orienting normative inquiry around *theoretical* concepts. First, consider "explicated" theoretical concepts that can be explicitly and uncontroversially characterized in relatively precise terms. One reason to adopt such concepts is that explicitness and precision can be highly useful in normative investigation. And it may be extremely difficult (if possible at all) to achieve an explicit precise characterization of a relevant folk concept. Further, explications can be tailored to match what the inquirer cares about most in inquiry (to the extent this is clear to her).

There are at least two reasons to be cautious about this sort of consideration, however. First, at least on some theories of meaning or content, explication or stipulation provide no guarantee of meaning. For example, some theorists will argue that even if 'atom' was introduced to pick out the simple indivisible constituents of matter, the term functioned in important ways in physics independent of that definition, allowing it to turn out that atoms are neither simple nor indivisible.[19] Second, precision might be positively *misleading*, if the subject matter being studied itself lacks precise boundaries. Aristotle famously emphasized this point, and took it to apply to certain normative inquiries.[20]

A second reason to orient normative inquiry around theoretical concepts is that they, as a general *kind* of concept, have a track record of relevant usefulness. Inquiry in the natural and social sciences often involves the introduction of new theoretical concepts. For example, think of the concept QUARK in physics, or the concept IMPLICIT BIAS in the social sciences. Because the natural and social sciences include important paradigms of successful inquiry, this might suggest that orienting around theoretical concepts can contribute to the success or significance of an inquiry. Plausibly, theoretical words and concepts are often introduced because inquirers identify a need for new concepts in order to promote their aims as inquirers. One might hope that the theoretical concepts that emerge in normative inquiry can similarly help us to better achieve the aims of such inquiry.

One reason for caution about this reason is that its force plausibly depends greatly on the degree of similarity between normative inquiry (and what it investigates) and scientific inquiry (and what it investigates). And there is great controversy concerning this very question.[21] We return to this issue in section 4.

At the start of this chapter, we claimed that, if we focus on a specific agent in a specific context, two central questions in conceptual ethics are: *which concepts* that agent should use, and *which words* she should use to express those concepts. In this section, we have canvassed some reasons for favoring orienting normative inquiry around folk concepts, or around theoretical concepts. This is an instance of the first

[18] For connected discussion, see Railton (2003); Jones (2005); and Eklund (2017: chapter 7).

[19] Compare Schroeter and Schroeter (2014).

[20] Aristotle (2002: 1094b13–b25).

[21] Compare the important (although now dated) review of metaethics in Darwall, Gibbard, and Railton (1997), which marks a central division in metaethics as being between theories that posit *continuity* vs. *discontinuity* between ethics and science.

kind of issue in conceptual ethics. In the next section, we now turn to the second kind of issue, about concept/word pairing.

2. Concept/Word Pairing

There are a wide range of philosophically rich issues in conceptual ethics that bear on which words to use in order to express a given concept. For example, the use of certain words might be advocated because (i) their use in relevant contexts promotes certain broadly epistemic goals (such as the pursuit of knowledge of a given subject matter—for example, in physics or moral philosophy)[22] and/or (ii) their use in relevant contexts promotes certain broadly practical goals (such as the political goal of helping foster a more just or free society, or the ethical goal of living a better life or avoiding causing unjust harms to others).[23]

It is beyond the scope of this chapter to address the full range of such issues as they apply to the conceptual ethics of normative inquiry. Instead, our goal is to focus on a specific cluster of issues connected to our discussion in the previous section of whether to orient one's inquiry around folk or theoretical concepts. The specific cluster of issues we focus on arises from the fact that certain words that are central to existing normative inquiry (such as 'moral' or 'rational') are associated with several significantly different inferential patterns. As we will show, this entails that normative inquirers face interesting questions about which words to use to express either folk or theoretical concepts.

To begin, we distinguish folk from theoretical *words*. As we shall understand them, *folk words* are words used to express one or more folk concepts. Similarly, *theoretical words* are words used to express one or more theoretical concepts. Given this way of drawing the distinction, some words will be both folk words and theoretical words. This is true, for example, of 'rational' and 'moral'.

To make the idea of associated inferential patterns vivid, we will focus on the word 'moral'. Consider the range of competing philosophical accounts of what distinguishes moral thought and talk from other normative thought and talk. Some associate morality with attitude-independent categorical norms.[24] Others associate it with an impersonal point of view,[25] or a distinctively interpersonal one.[26] Still others associate it with the fittingness conditions for attitudes of certain emotions, such as guilt, resentment, and anger.[27] Others associate it with a list of supposed "platitudes" about morality that include both things about the subject matter of morality (e.g., which kinds of activities it regulates) as well as things about its (at least purported) relative normative import.[28] And many assume or defend a kind of *rationalism* about moral obligation, according to which, if you morally ought to perform an action you therefore ought to perform it, in the most authoritative sense of 'ought'.[29]

[22] See, for example, Carnap (1950/1962); Brandt (1979/1998); and Railton (1986).

[23] See, for example, Brandt (1979/1998); Railton (1986); and Haslanger (2000).

[24] Williams (1985).

[25] Sidgwick (1874).

[26] Scanlon (1998) and Darwall (2006).

[27] Gibbard (1990).

[28] See Smith (1994). Cf. Foot (1972).

[29] Korsgaard (1996).

Our aim here is not to adjudicate among these proposals. Rather, we want to emphasize that—at least in our broad social/historical context—each of these proposals captures an idea that can seem natural to associate with the word 'moral', at least in some circumstances. We can understand these associations as connected to inferential patterns: in at least some circumstances, it will seem natural to infer from the presence of a moral property to the presence of one of these associated properties, or vice-versa.

On some views about how to individuate folk concepts, it might be that *each* of these competing proposals correctly analyzes one among many folk concepts that are sometimes expressed by 'moral'. Suppose that this were so. This would pose a clear danger that many exchanges among normative inquirers could end up being "merely verbal disputes", where speakers "talk past" each other in their discussions.[30] Merely verbal disputes can stall the progress of inquiry, or hinder it in other ways (e.g., by leading to confusion on the part of participants).[31] However, our focus here is *not* on these dangers that verbal disputes may pose for normative inquiry. Rather, it is with related (but distinct) issues: issues that we think are under-appreciated in the practice of contemporary normative inquiry (even if many practitioners of normative inquiry will, on reflection, agree these are issues to be reckoned with).

To see these issues, suppose for the moment that the word 'moral' picks out a single concept MORAL across the range of uses relevant to normative inquiry (i.e. setting aside obviously different uses like 'the moral of the story'.) As we suggested above, there are many distinct inferential patterns associated with the word 'moral' in our social/historical context. Even if one such inferential pattern is no part of the concept MORAL, it could still play an important role in explaining how we in fact reason using that concept. This raises what we call the *unreliable inference* danger when using such a word in normative inquiry. This is the danger that, when attempting to use MORAL, or consulting one's intuitions about it, the inquirer could be guided not by this concept, but by an associated idea that is not constitutive of the concept, and which in fact is inaccurate. This danger can arise even for the inquirer who intends to use the word 'moral' to express the (allegedly unique) concept MORAL.

There are several potential species of such unreliable inferences. One important species that we will return to below involves what we will call *implicit switching*. This occurs when different associated ideas guide inference in a single sequence of reasoning. For example, one might infer that an action was morally required from its being demanded from a distinctively interpersonal point of view. One might then

[30] See Chalmers (2011) and Jenkins (2014) for helpful general discussions of verbal disputes.

[31] We should underscore that we do *not* think that every time there is variation in word meaning speakers are doomed to end up in a merely verbal dispute. In some cases, the speakers might still express genuine disagreements when they utter claims using a term 'X', despite this variation in meaning. For example: the parties might be involved in a *metalinguistic negotiation*. In cases of metalinguistic negotiation, a speaker uses (rather than mentions) a term to advocate for a view about how that very term should be used. Speakers in a metalinguistic negotiation might well express conflicting normative views about how a word should be used—views that will standardly be based on normative considerations about things *other* than words and concepts (e.g., how we should live, how we should organize our social/political institutions, or what objective joints there are in reality)—even if those views are expressed through pragmatic mechanisms (rather than in terms of literal semantic content). See Plunkett and Sundell (2013) for further discussion. See also Thomasson (2016) and Ludlow (2014) for connected discussion.

infer *from* the action's being morally required that one has an authoritative obligation to do it. Unless these associated ideas are both necessary conditions on the concept MORAL, this inference could lead one astray.

Against these dangers, one must weigh a straightforward reason to use familiar, as opposed to obscure or novel words: speech or text composed of such words will generally be *easier to understand*. To see this, imagine a book that begins by introducing a long list of explicitly defined novel terms, and then proceeds to use only the newly defined terms in the text that follows. Such a book will be much harder to understand than a book conveying the same ideas using familiar words. Insofar as folk words are typically more familiar than theoretical words, this issue about comprehensibility suggests a presumptive reason for using folk words.

So far, we have focused on the use of folk words to express folk concepts. However, it is also common for normative inquirers to appropriate existing folk words as vehicles to express their *theoretical* concepts. For example, explications take this form.[32] One reason to use existing folk words to express an unfamiliar theoretical concept is that it can help to orient you (and your audience) to roughly what you want to be talking about. Another reason is that it might be an important element of a campaign to get a group of speakers (e.g., all ordinary speakers, or a specific subset of philosophers, or a group of political activists, etc.) to reform their usage to accord with your preferred usage.[33]

The use of folk words to express theoretical concepts faces a clear form of the unreliable inference danger. For any folk term, a competent speaker will tend to find natural the inferences that they have come to associate with that term. And this may lead them to make these inferences even if they are not licensed by the theoretical concept they are using.

To make this vivid, consider Ronald Dworkin's use of the term 'morality' in *Justice for Hedgehogs*. Dworkin stipulates that there is a "distinction between ethics, which is the study of how to live well, and morality, which is the study of how we must treat other people".[34] This stipulated definition of 'morality' means that we won't have specifically "moral" reasons or obligations (etc.) that stem directly from the welfare of at least many non-human animals, or the status of the natural environment. (And perhaps we will also lack "moral" reasons that arise from the welfare of human infants, depending on how we cash out 'people'.) Because the stipulation is fully compatible with there being weighty 'non-moral' reasons arising from (e.g.) the welfare of infants or non-human animals, it is not clear whether this fact about the consequences of the stipulation is a problem. However, we think this sort of stipulation, and the fact that it excludes considerations about the entities we have highlighted, illustrates important worries in conceptual ethics.

[32] See, for example, Carnap (1950/1962); Carnap (1947/1956); Railton (1986); and Brandt (1979/1998).

[33] This sort of campaign might be explicit. See, for example, the discussion of race and gender terms in Haslanger (2000) and the discussion of 'true' in Scharp (2013). Or it might occur via metalinguistic negotiation. For discussion of this latter option, in the context of the use of philosophical terms, see Plunkett (2015).

[34] Dworkin (2011: 13). It should be noted that others follow Dworkin in this stipulated usage when framing their discussions. For example, see Appiah (2005: xiii) drawing on Dworkin (2000).

We take Dworkin's stipulation to face an especially worrisome instance of the implicit switching danger. Many ordinary people take 'morality' to pick out something that is *particularly* normatively important. It will be hard for many to shake this association. This threatens to make illegitimate inferences more likely: e.g., inferring, without argument, from a claim about something being "moral" (in the stipulated sense) to a claim about its distinctive normative importance, relative to other kinds of considerations. This means that using the term 'morality' to refer to Dworkinian morality threatens to undercut giving certain normative considerations their due in normative reasoning. This is especially worrisome if the kinds of considerations being ruled out involve entities (e.g., many non-human animals, human infants, or humans that lack certain cognitive capacities) that are already objectionably marginalized in our actual social/political practices, normative inquiry, or both.[35]

It is worth emphasizing that this is not intended as a point about *any particular philosopher's* psychology. Insofar as normative inquiry is a social endeavor, these issues will be important when they arise for members of the community of inquirers who attempt to engage with or use this explication.

Notice that a related danger can arise even for philosopher who is careful not to stipulate the meaning of 'moral' in this way. For example, consider T. M. Scanlon's *What We Owe to Each Other* (1998). Scanlon offers a contractualist account not of morality as a whole, but "... of a narrower domain of morality having to having to do with our duties to other people, including such things as requirements to aid them, and prohibitions against harming, killing, coercion, and deception".[36] This is roughly the same part of the normative that Dworkin stipulatively uses the term "morality" to refer to. Scanlon goes on to say the following: "It is not clear that this domain has a name. I have been referring to it as "the morality of right and wrong", and I will continue to use this label".[37] However, even here the narrower usage of 'morality of right and wrong' is so close to an ordinary way of talking about morality as a whole, that (e.g.) implicit switching concerns may still loom large.

Similar dangers show up in many sorts of normative inquiry. For comparison, consider the case of 'rational' in epistemology. Many epistemologists treat 'rational belief' and 'justified belief' as interchangeable. For example, Stewart Cohen claims that "'reasonable' and 'rational' are virtual synonyms for 'justified'", Michael Huemer claims that "another word for what is justified... is 'rational'", and Declan Smithies claims that "to say that one has justification to believe a proposition is to say that it is rational or reasonable for one to believe it".[38] This association with 'rational' may have important consequences.[39] Quite generally, criticism of someone as "irrational" suggests that something has gone wrong in them. Applied to belief, this feature of IRRATIONAL makes for extremely natural inferences from irrationality to failure of epistemic responsibility. This in turn means that the assumed association between 'justified' with 'rational' may lend unearned plausibility to responsibilist

[35] For connected discussion here about the case of 'justice', in particular with respect to the way it interacts with normative concern for non-human animals, see Plunkett (2016b).

[36] Scanlon (1998: 6).

[37] Scanlon (1998: 6).

[38] See Cohen (1984: 283); Huemer (2001: 22); and Smithies (2012: 274), respectively.

[39] For discussion and critique of equating justification and rationality, see Sylvan (MS).

theories of epistemic justification.[40] Even if philosophers stipulate a meaning for 'rational' that does not support such inferences, use of the word may nonetheless influence the inferences or intuitions that inquirers make when using the stipulated notion, rendering them less reliable. Note that these issues closely parallel the ones just discussed above about the Dworkinian notion of morality, even though much of this discussion doesn't involve any specific *stipulation* of terminology.

Thus far, we have considered the use of folk words in normative inquiry. This has led us to say some things along the way about the virtues and vices of using theoretical words. We now turn to that topic more explicitly.

As the preceding discussion makes clear, one reason for using theoretical words is to *disambiguate*. If there are many ideas associated with 'moral', then using a novel theoretical term might help to focus one's attention on the specific concept one has in mind. This can be true even if the concept you wish to deploy is a folk concept. Consider three cases.

First, suppose that one is convinced that the correct understanding of the folk concept MORAL concerns when guilt and resentment are fitting. One might introduce a novel term to talk about this concept, rather than using the word 'moral', because one is worried about triggering associations between the word 'moral' and ideas unconnected to the conditions under which these emotions are warranted.

Second, some terms relevant to normative inquiry are associated with *both* folk and theoretical concepts. (As with 'moral' and 'rational'.) In light of this, one might want to explicitly *flag* that one is using a folk concept by using a theoretical word like 'folk morality' or 'folk rationality'.

Third, if you believe there are multiple folk concepts expressed by a given term, you might want to explicitly flag *which* one you are using, via introducing a technical term. For example, suppose that you believe that the term 'knowledge' is associated with at least two folk concepts: a factive concept, which philosophers have traditionally privileged, and a non-factive concept.[41] One might want to introduce a technical term, like 'non-factive knowledge', to focus attention on which folk concept you intend to pick out. *Mutatis mutandis*, these same considerations can favor the use of theoretical terms to express theoretical concepts.

Next consider potential dangers associated with using theoretical terms. The first thing to note is that many *theoretical* terms (especially those with a significant history of use) have a plurality of ideas associated with them in theoretical contexts. This is true for many prominent theoretical terms in philosophy, like 'grounding',

[40] For one place in contemporary epistemology that relies on this distinction, see Weatherson (2008) and Littlejohn (forthcoming), which both offer replies to the "New Evil Demon Problem" (see Cohen 1984). Note that our calling attention to this conceptual distinction is fully compatible with the possibility that the correct theory of epistemic justification turns out to vindicate a tight connection between epistemic justification and rationality, and hence certain patterns of inference involving them.

[41] This non-factive sense is prominent in the history and sociology of science, where scholars will sometimes talk about the "production of knowledge" concerning claims we now know to be false (but which were thought to be true at the time, or which were relied on in certain ways in scientific practice), as in Shapin (1994), and in psychology, as in Gilovich (2001). See also Michel Foucault's use of "knowledge" in his discussion of his idea of "power/knowledge", as in Foucault (1980), and in earlier work such as Foucault (1966/2000). For an apparently non-factive folk use, see Seuss (1965).

'epistemic', or 'metaethics'.[42] On some semantic views, this divergence might undermine our ability to effectively refer to *anything* with these terms. Consider, for example, certain externalist theories according to which reference is determined in large part by expert usage. If it is indeterminate which (if any) experts one is deferring to in use of one of these terms, one might fail to refer to anything at all.[43]

Even setting this worry about reference failure aside, using theoretical terms with a plurality of associated ideas invites worries about unreliable inference. For example, consider the term 'epistemic'. This is arguably a theoretical word in its central uses by contemporary philosophers engaged in normative inquiry.[44] In some cases, philosophers explicitly claim that properly "epistemic" justification must be the sort of justification that bears an explanatory connection to *truth*.[45] Others use the term 'epistemic' to pick out norms or values that are tied to the *constitutive* standards that govern beliefs, where it is then a further question whether or not those standards are truth-related or not.[46] Finally, some take the epistemic standards to be particularly normatively important or weighty with respect to all-things-considered normative theorizing about beliefs.[47]

For many expert epistemic inquirers, each such association of the word 'epistemic' will have become psychologically entrenched and intuitive. And this means that such experts may become vulnerable to versions of the unreliable inference dangers that we discussed for the use of folk words.[48] Examples like this show that switching to theoretical terminology does not guarantee that one will thereby avoid vulnerability to implicit switching and unreliable inference worries. However, switching to such theoretical terminology may nonetheless help to mitigate those worries. For example, the associations of a theoretical term with a certain range of ideas will often be less psychologically entrenched (even in an expert), and easier to make explicit than associations with folk terms. And use of a theoretical term at least sometimes generates pressure to be explicit about what one intends to communicate via the term.

As we have sought to make clear, the question of which words to use remains complex, even given a decision concerning whether to use folk or theoretical concepts. Many central terms in normative inquiry—both folk and theoretical—are associated with multiple theoretically significant ideas. We have argued that this raises significant dangers for normative inquiry.

These dangers are exemplary of a cluster of general issues in conceptual ethics concerning concept/word pairing. In the next section, we discuss whether to orient normative inquiry around extension-revealing or authority-revealing concepts.

[42] For extensive discussion relevant to 'metaethics', see McPherson and Plunkett (2017).

[43] See Cappelen (2013).

[44] On this point, see Cohen (2016).

[45] We endorse this explanatory connection in McPherson and Plunkett (2015). See Berker (2013: section 3) for references to epistemologists endorsing a range of similar theses. For a view that denies this kind of explanatory connection, see Enoch and Schechter (2008).

[46] See Nolfi (2014). Note that Nolfi herself goes on to deny that belief aims at truth. See also Nolfi (MS).

[47] One unusually explicit example: Schroeder (2018) presupposes that practical and epistemic reasons are of a kind, and explores potential explanatory priority relations between them. Elsewhere, Schroeder suggests that the normative is "all about reasons" (Schroeder 2007: 81), in (at least) the sense that the core normative facts that really matter can all be reductively explained in terms of normative reasons.

[48] For connected discussion, see Cohen (2016).

Many of the complications and dangers highlighted here about concept/word pairing also arise in connection with that issue. Having illustrated the general structure of this cluster issues about concept/word pairing here, we will merely touch on a couple of examples of this in the next section.

3. Authority-Revealing vs. Extension-Revealing Concepts

Consider two things that someone engaged in a normative inquiry might want from a normative concept. On the one hand, we care about which actions, states of affairs, etc., our normative concepts apply to. Given this, it would be attractive if there were a clear way of discovering these sorts of facts about a concept. On the other hand, it would be attractive if we were confident that normative conclusions framed in terms of that concept *really mattered*. Ideally of course, one could find a concept that possessed both of these features to a high degree. But it often seems that one in fact needs to weigh these features against each other.[49]

To illustrate: an inquirer interested in norms that govern our practical lives might orient her inquiry using the concept POLITENESS. Or, she might focus on the concept MORAL WRONGNESS. Or, she might introduce a novel concept that she stipulates to be maximally normatively significant for deliberation or action. Competent users arguably typically have a pretty strong grip on what is polite, but how much politeness normatively matters is highly contested. The situation might be almost reversed for the introduced theoretical concept. If forced to choose between orienting her inquiry around one of these normative concepts, which should she choose? In this section, we consider what can be said about this question.

It will be useful to have names for the two attractive features we contrasted above, as well as to say a bit more about what each of these features involve. We will call a concept *extension-revealing* to the extent that ordinary use of that concept tends to make facts about its extension accessible in a way that is usable to the person employing that concept.[50] Extension-revealingness is gradable along a number of different dimensions. For example, one dimension is how quickly or easily someone can identify what's in the extension. Another dimension is how much of the extension she can identify (e.g., just paradigm cases or much larger swaths of the extension). Moreover, concepts can be extension-revealing for different reasons. In some cases, it might be that users can identify part of the extension of a concept by grasping analytic *a priori* truths about the concept. In other cases, users can identify part of the extension of a concept because of how the relevant community has been

[49] This might be *necessarily* so, given the nature of what we below call 'authoritatively normative' concepts. (For example: if things in the vicinity of Moorean "open-question" qualities of the sort discussed by Gibbard (2003), drawing on Moore (1903/1993), are true of authoritative normative concepts, but not other concepts.) For our purposes here, we leave this issue aside.

[50] We use 'extension' here because it is a familiar and useful label. We intend the point we are making to apply to normative terms even if they lack extensions. For example, suppose that some normative terms function as operators. For such cases the relevant question is how much is revealed about, roughly, the truth-conditions of sentences containing the operator.

trained to use the concept over time, even if the relevant knowledge is neither analytic nor *a priori*.

The relevant idea of *normative authority* is less straightforward, so we will introduce it via an example. Suppose that Priya takes herself to face a conflict between what morality and prudence demand of her. She might ask: given this conflict, *what ought I to do?* Notice that, when Priya asks this question, it does not make sense to interpret her as seeking just *some normative standard or other* that can adjudicate the perceived conflict. For example, if etiquette joins with morality against prudence in this case, this hardly answers her question. Rather, Priya is deploying a normative concept here that aspires to wear a distinctive normative *authority* on its sleeve: a concept such that, when answers are framed using it, those answers purport to settle conflicts such as one between morality and prudence.[51] Or at least she is *trying* to deploy such a concept (e.g., perhaps she fails to token a determinate and coherent concept.) A concept is *authority-revealing* to the extent that competence with that concept tends to make its authority accessible in a way that is usable to an actual competent person employing that concept.[52] For brevity, call the apparently authority-revealing concept that Priya tokens in her deliberation AUTHORITATIVE OUGHT.

It is crucial that the properties of concepts that we are focused on concern what inferences it is *relatively* straightforward for competent users of these concepts to make. For example, we want to leave it open that the concept POLITENESS is in fact maximally authoritative. We insist only that such authority is not transparent to competent users: it would not be particularly puzzling or surprising for a competent user of POLITENESS to claim that this concept simply picks out certain social relations, with little authoritative normative significance. By contrast, it would be puzzling and surprising for a normal competent user of AUTHORITATIVE OUGHT to take this concept to be non-defective, and yet to deny that it has authoritative normative significance.

Conversely, it would be puzzling for the ordinary user of POLITENESS to simply deny that they had any idea what sorts of behaviors were polite in their community. POLITENESS is arguably highly extension-revealing. By contrast, AUTHORITATIVE OUGHT is arguably not extension-revealing. This does not imply that it lacks a determinate extension. Rather, it simply signals that relatively few facts about the extension of this concept are transparent to ordinary competent users of this concept, or at least such users in our social/historical context. For example, it would not be

[51] See McPherson (2018). Note that an authoritative resolution of such a conflict could take many forms. For example, it could involve deferring entirely to the dictates of morality. Or it could involve weighing the verdicts of morality and prudence in any number of more complicated ways. For example: generally favoring the dictates of morality, but not in cases when prudence *very strongly* suggests you should Φ, and morality only weakly recommends you not Φ. Or, for example, it could involve completely ignoring the final dictates of morality and prudence and balancing contributory factors (perhaps including factors that these norms take as morally or prudentially significant).

[52] Notice that if you think there is a hierarchy of degrees of normative authority, then it would be important to distinguish: (i) *what degree of authority* is revealed to the competent user, and (ii) *how manifest* that authority is to the competent user. We will ignore this complication in the text.

puzzling or surprising for a competent user of this concept to deny that she had any idea what is in its extension.

Consider next claims about *genuine* reasons to perform a certain action, or the *genuine value* of certain states of affairs. On at least some ways of understanding these sorts of claims, they purport to deploy authority-revealing *contributory* notions. For example, one might think it is a conceptual truth that what one ought to do is just what one has most reason to do.[53] If this line of thinking is correct, it suggests that there might be whole classes of authoritative concepts, cutting across categories like the evaluative, deontic, aretaic, etc.[54]

The contrast between authority-revealing and extension-revealing concepts might be taken to be equivalent with, or at least deeply tied to, the contrast between thin and thick concepts.[55] In rough terms, thick concepts (e.g. BRAVE, COWARD, KIND, JERK, etc.) involve a mixture of descriptive and normative application-conditions. The kind of mixture they involve is meant to contrast with thin ethical concepts (e.g. OUGHT, BAD, etc.). It is a matter of much debate how to understand the contrast between thin and thick concepts, as well as what exactly each of them *are*. For example, there is much debate about whether the descriptive and normative aspects of thick concepts can be separated from each other, as well as whether the normative aspects of thick concepts are part of their content, or whether they are instead pragmatically associated with their use.[56] We should not treat the distinction between thick and thin concepts as equivalent to the distinction between authority-revealing and extension-revealing concepts, even if it turns out that certain thin normative concepts are most authority-revealing. The basic reason is this: for both thick and thin concepts, there is the question of how authoritative the normativity associated with that concept is. This might vary across thick concepts: for example, contrast TREACHEROUS with BANAL. The same is arguably true for thin normative concepts: famously, the truth-conditions of 'ought' can vary widely across contexts of use, sometimes picking out intuitively non-authoritative standards.[57]

With this orientation in mind, let's return to the question of whether to orient one's normative inquiry around authority-revealing concepts or extension-revealing concepts. This question presupposes that we cannot have a concept that is maximal on both dimensions. This might be denied. Consider, for example, one particularly optimistic constitutivist project, which involves two goals. The first is to identify a theoretical concept of a constitutive norm, perhaps CONSTITUTIVE NORM FOR ACTION, which is maximally authoritative. The second is to argue that we can derive the complete set of facts about what we ought to do in a relatively straightforward way on the basis of reflection on this concept. The rationales we consider below

[53] Douglas Portmore calls this the "least controversial normative principle concerning action" (Portmore 2013: 437). (Whether it is then a *conceptual* truth is a separate matter.) For an argument against understanding 'ought' in terms of 'most reason', see Broome (2015).

[54] See McPherson (2018) for this proposal.

[55] For example, parts of Eklund (2017) seem to presuppose this equivalence. See, for example, Eklund (2017: 18–19).

[56] See Roberts (2017) for an overview of thick concepts.

[57] On this point, see Kratzer (2012) and Finlay (2014).

assume that such a project is not highly promising.[58] It is instructive to contrast this sort of ambitious constitutivist project with an attempt to introduce a concept by stipulating both that it is maximally authoritative, and that it has a certain extension. It is plausible that this fails: we cannot stipulate our way to making anything we like maximally authoritative.[59]

We begin by considering a natural rationale for focusing on authority-revealing concepts. Normative inquiry, like any inquiry, can be motivated by the desire to know (or better understand, etc.) facts about our world. However, it is very natural to think that normative inquiry can also be centrally motivated by the desire to discover answers that can (in some sense) *directly* guide our decisions about (e.g.) how to live, what to believe, or which social/political arrangements to support, promote, or protest. It seems plausible that answers framed in terms of authority-revealing concepts will be most suitable to play this role of direct guidance.

Further, many philosophers who investigate normative topics without using authority-revealing concepts arguably do so because they assume a connection between these topics and authoritative norms. For example, in political philosophy, many philosophers who investigate issues about justice, freedom, and/or equality do so in part because they take considerations about these things to bear in important ways on how we authoritatively ought to arrange our social/political institutions.[60] Similarly, in ethics, many philosophers who investigate morality assume that morality is closely connected to the authoritative norms for action.[61] And within epistemology, it is widely assumed that *knowledge* is intimately tied to the most authoritative norms for the regulation of belief.[62]

If we assume these commitments, one central reason to orient normative investigation around maximally authority-revealing concepts is that orienting it instead around EXPLOITATION or MORALITY or KNOWLEDGE may seem like a puzzling and potentially dangerous detour. At best, one investigates the topic of ultimate interest less directly than one could. But there are more significant dangers, which can be illustrated by considering the example of MORALITY. The fact that many inquirers focus on MORALITY is partly explained by the fact that normative authority is saliently associated with the word ‘moral’. One danger is that the association with

[58] Notice here that we are not setting aside the possibility of a slightly less optimistic constitutivism, according to which there is a concept whose authoritativeness and extension are both derivable via difficult constitutivist philosophical reasoning.

[59] For connected discussion, see Prior (1960).

[60] See, for example, Rawls (1971/1999); Nozick (1974); Anderson (1999); Sen (2009); Satz (2010); Dworkin (2011); and Pettit (2012).

[61] See, for example, Smith (1994) and Korsgaard (1996).

[62] On this front, consider the vast amounts of effort spent on thinking about the nature of knowledge in the wake of Gettier (1963). Some epistemologists might well be content to think of engaging in this enterprise as just investigating the contours of a concept we happen to employ, which picks out something with little connection to authoritatively normative facts about what we should actually believe (or how we should proportion our credences, etc.). However, surely part of why so many epistemologists have been concerned with the nature of knowledge is that they assume that issues about knowledge (e.g., who has it when) are *normatively important*. In a similar vein, consider the rise of so-called “knowledge-first” epistemology in the vein of Williamson (2000), which uses knowledge as the basis for many things we take to have normative import (e.g., evidence).

normative authority might not be vindicated by the correct theory of the inferential role of MORALITY. Even if the association is vindicated, the plurality of ideas associated with 'moral' raises the danger that framing one's inquiry in terms of this concept will make that inquiry vulnerable to unreliable inference dangers.

What can be said against this rationale for orienting one's theorizing around authority-revealing concepts? To begin, consider the idea that MORAL is not itself authority-revealing, but that certain moral facts *ground* the facts about what you authoritatively ought to do. If this were true, it might seem to make sense to investigate moral facts as a way of discovering what one ought to do. One thing to be said for this approach is that—assuming that all of its presuppositions are correct—such investigation might contribute to providing *understanding* of authoritative normativity, by allowing us to discover not just what we ought to do, but *in virtue of what* we ought to do it.[63]

Another rationale for investigating the grounds of normative facts is that one might think this is an especially promising way to discover what those facts are. One difficulty with this rationale is that priority in metaphysical explanation is not reliably correlated with epistemic access. To take a simple example, the fundamental microphysical facts presumably ground facts about the observable features of our environment. But we evidently have more direct access to the latter than we do to the former. So this rationale would require an argument that shows why metaphysical priority is a good guide to epistemic access in the case of the normative. For example, it might be that, at least in certain epistemic contexts, we have no good epistemic access to the relevant facts that doesn't proceed via learning about the facts that determine them.[64] But it is far from clear that this is so in the case of authoritatively normative facts, in the sorts of epistemic contexts that we standardly find ourselves in.

Consider a different epistemic argument for focusing on certain less authority-revealing concepts (e.g., perhaps MORALITY or EXPLOITATION), rather than directly on what one authoritatively ought to do. Recall from section 1 the suggestion that we all have a good deal of (perhaps implicit) trained moral knowledge, as a result of our moral education. One might think that most of us lack a parallel training that is framed in terms of the concept AUTHORITATIVE OUGHT, or associated contributory notions, such as GENUINE REASON FOR ACTION. This might be because AUTHORITATIVE OUGHT (e.g.) is a theoretical, and not a folk concept. But it need not be: the same point would hold even if it were simply a *relatively obscure* folk concept. If this is true, then, when one is considering a possible action, and asking oneself the question "authoritatively ought I to do this?" one might tend to come up blank. Or, one might tend to generate an answer that is based in one's implicit understanding of some *other* concept, such as WHAT SERVES MY PRESENT AIMS, or WHAT IS GOOD FOR ME, or MORAL RIGHTNESS, without interrogating the implicit inference to what one ought to do.

[63] Note that one might view understanding as *one* goal of normative inquiry, or as *the distinctive* goal of normative inquiry. A case for the latter view might draw on parallel arguments that understanding is a constitutive goal of moral inquiry in Hills (2009).

[64] See Greenberg (2006) for discussion of this idea within legal philosophy, in the context of discussing facts about the content of the law (in a given jurisdiction, at a given time).

Suppose next that normative authority was part of the concept MORALITY, such that, at least ordinarily, if one morally ought to perform an action, one also authoritatively ought to perform it. If this were true, it might make sense to try to answer at least some questions about what we authoritatively ought to do by engaging in moral inquiry, and inferring conclusions about what we authoritatively ought to do from the moral conclusions found.

We can imagine parallel rationales for the other cases. For example, one might think that a similar argument could be made for orienting one's inquiry in social and political philosophy around JUSTICE, even if one's ultimate aim was to make claims about how we authoritatively ought, all things considered, to arrange our social and political institutions. Similarly, one might advocate for orienting epistemological inquiry around KNOWLEDGE, even if one's ultimate aim was to make claims about how we authoritatively ought, all things considered, to regulate our beliefs.

The force of this rationale interacts with the dialectic concerning folk vs. theoretical concepts discussed earlier, as well as with the issues we discussed about concept/ word pairing. For example, this rationale may require inquiry that deploys *folk* moral concepts, as opposed to explicitly theoretically refined moral concepts (where it is much less clear that we would have the relevant sort of trained intuitive knowledge). As such, this rationale is also vulnerable to the various worries about using folk words and concepts discussed in sections 1 and 2 (including ideological contamination, implicit switching, etc.).

Note next that the idea that *any* concepts are authority-revealing might be challenged. Consider the way we tried to describe Priya's concept. It might be denied that anything could satisfy such a concept.[65] Or it might be denied that examples like Priya's deliberation even succeed in isolating a coherent concept.[66] Perhaps, as we suggested with 'morality', above, there are several features that philosophers associate with 'authoritative normativity', but no single property or concept could possess all of those features.[67] If this were the case, there might be a range of authoritative-ish concepts for the normative inquirer to choose among in her explication. And such a choice, once clarified, might reduce the force of the natural rationale for focusing on the authoritative concepts in one's normative inquiry.

Other cases raise analogous complications. For example, it is not clear whether there is a single way in which norms are authoritative *for belief*. One way to bring this out is to consider the variety of interesting properties that various philosophers have associated with *epistemic justification*.[68] One possible explanation for this diversity is that—given the importance that many place on the topic of "epistemic justification"—many of these properties are ones that (at least some) philosophers associate with authoritative norms for belief. Another way to bring out the issue is to consider possible competing proposals concerning what it is for a norm to be authoritative for belief. For example, a natural proposal is that being authoritative

[65] See Copp (1997) and Tiffany (2007). See McPherson (2018: section 6) for discussion.

[66] See Baker (2018).

[67] See McPherson and Plunkett (2017: section 2.3) for brief discussion of this possibility. For connected discussion, see Finlay (2019).

[68] See Alston (2005) for a survey of some of the key views here.

for belief is a matter of being related to truth in the right way.[69] Or perhaps it is a matter of being related to authoritative practical norms in the right way.[70] Or perhaps it is a matter of being the constitutive norm for the mental state of belief.[71] One possibility is that each of these proposals captures something that we care about in believing, but there is no single property *being authoritative for belief*.

Let's take stock of where we are. This section has sought to illuminate the interest and complexity of the question of whether to orient normative inquiry around authority-revealing or extension-revealing concepts. As we have indicated in certain places, the plausibility of answers to this question interact with the plausibility of answers to certain questions about folk vs. theoretical concepts. This underscores some of the interest in investigating these two different topics in the conceptual ethics of normative inquiry together.

Importantly, conceptual ethics is itself a kind of normative inquiry. This means that the issues in conceptual ethics we have discussed thus far also interact with how we understand and approach the very normative questions we have been posing in this chapter. We now turn to this topic.

4. Evaluating the Norms and Concepts used in the Methodology of Normative Inquiry

So far in this chapter we have explored the choices between focusing normative inquiry around folk or theoretical concepts (and words), and between focusing on authority-revealing and extension-revealing normative concepts. In doing so, we have introduced a series of normative considerations and arguments that bear on these choices. In seeking to assess these considerations and arguments, we can ask: *what sorts of norms* should we be deploying in our methodological evaluation of normative inquiry? We can further ask: for any such answer, what would *explain* the aptness of that answer? In this section, we briefly consider three broad approaches to answering this latter question: (i) considering the significance of inquirer aims, (ii) appealing to the results of metanormative inquiry, and (iii) engaging in further conceptual ethics, of the kind we have been discussing in this chapter. We will also explore some of the ways that these approaches interact.

We begin by considering a natural proposal: that the normative facts about which normative concepts and words an agent should use in a given context are determined

[69] We defend a related thesis about *epistemic justification* in McPherson and Plunkett (2015). As we just discussed, epistemic justification might well be taken by some to be tightly associated with authoritative normativity. On this front, consider our earlier discussion of those who take justification and rationality to be virtually equivalent. In turn, many such philosophers who accept that (at least near) equivalency also take considerations of rationality to bear in significant ways on what we (authoritatively) ought to believe. (See, for example, Wedgwood 2012, in connection to Wedgwood 2007).

[70] This might be thought of as a kind of "pragmatist" explanation. See McPherson and Plunkett (2015: 111) for discussion of a related thesis about the epistemic. For connected discussion, see Enoch and Schechter (2008).

[71] For example, Shah and Velleman (2005) defend a constitutivist account of epistemic normativity. For connected discussion, see Nolfi (2014).

by facts about what *promotes the aims* that an agent has.[72] One way to motivate this *instrumentalist* idea is to notice that it seems to explain certain obvious data points. For example: if an agent aims to do work in sociology, then the concepts she should deploy are arguably different than those she should deploy if seeking to do work in physics. Or, to take another example: the concepts she should deploy when engaged in public political advocacy are arguably different than those she should deploy if seeking to do advanced theoretical work in political philosophy. In both of these cases, it seems that what matters is what promotes the aims the agent has in each case. One elegant hypothesis is that, if the aims of inquirers explain normative facts about concept choice in *these* cases, perhaps they also do so in *all* cases.[73] Or more modestly, perhaps the aims of inquirers play this role in the relevant range of cases about conceptual ethics that we are considering in this chapter. This would suggest a clear way in which (at least in principle) one could assess the choices we have highlighted thus far in this chapter: get clear about one's aims are, and investigate how best to serve those aims. (Note that, given that mature normative inquiry is often a collective endeavor, it may be more appropriate to speak of *our* or *their* aims here, rather than the aims of a given individual agent.)

It is not difficult to see that this sort of instrumentalist picture requires substantial defense. Instrumentalism is intensely controversial as a foundational theory of most significant types of norms and values (e.g., epistemic norms, reasons for action, well-being, etc.). Nor is it obvious that it has special credibility for the conceptual ethics of normative inquiry in particular. One way of making this vivid is to consider inquirers with substantively bizarre or awful aims; we may resist the idea that achieving those aims constitutes their doing normative inquiry *well*.[74]

It is possible to draw a different lesson from the example just used to motivate instrumentalism: perhaps our aims function to determine *what* we are investigating (e.g., *sociology* or *normative ethics*). Once we have settled on a topic, it is something about the topic itself that determines how we ought to inquire into it.

Consider an example to see how this idea might apply to a *normative* inquiry. Suppose that our aims make it the case that we are engaged in *moral* inquiry, and that the word 'moral' is univocal enough for this aim to fix a topic. We might then ask the following: what is the nature of moral thought and talk? And what, if anything, is that thought and talk distinctively about? These questions form the heart of what we understand to be *metamoral inquiry*. This inquiry, as we understand it, aims to explain how actual moral thought and talk—and what (if anything) such thought and talk is distinctively about (e.g., moral facts, properties, etc.)—fit into reality.[75] On a prominent class of metamoral theories, moral concepts are fundamentally *practical* in nature, such that they must be sharply distinguished from standard, descriptive concepts.[76] Recall

[72] See Haslanger (2000) and Anderson (2001) for two places where this sort of idea is advanced. See Burgess and Plunkett (2013b) for both sympathetic and critical discussion of this idea.

[73] The motivation we have suggested here for instrumentalism about the norms of inquiry is adapted from the motivation that frames Schroeder's defense of the Humean theory of reasons in Schroeder (2007).

[74] For further discussion, see Burgess and Plunkett (2013b: 1105).

[75] See McPherson and Plunkett (2017).

[76] See, for example, Korsgaard (1996), Blackburn (1998), and Gibbard (2003) for versions of this thought.

that one rationale proposed in section 1 for orienting inquiry around *theoretical* concepts appealed to an analogy to scientific inquiry. If moral concepts are of a radically different kind than scientific concepts, this might undermine that rationale.

By contrast, now consider different views about the nature of the reality that moral thought and talk is distinctively about. (Call this part of reality "moral reality".) On certain *naturalistic* views of moral reality, it is metaphysically continuous with—or a part of—the reality studied in the natural and social sciences.[77] This may support the idea that the study of moral reality *should* be modeled on the methods of the sciences, arguably strengthening the case for using theoretical concepts. By contrast, consider a non-naturalistic realist metamoral view, according to which moral reality is metaphysically discontinuous with the reality studied in the natural and social sciences.[78] Perhaps *that* reality can only be accessed using specific kinds of *a priori* reasoning that, at its core, must involve the deployment of folk concepts all of us have prior to any particular theoretical training.

The possibilities sketched here for moral inquiry extend to other kinds of normative inquiry. In general, how we should proceed in normative inquiry may in significant part be determined by our progress in metanormative inquiry.[79] We understand *metanormative inquiry* as aimed at explaining how actual normative thought and talk—and what (if anything) that talk is distinctively about (e.g., normative facts, properties, etc.)—fit into reality.[80] As these brief examples suggest, the connections between overall metanormative views and the topics in conceptual ethics we have been discussing in this chapter are rich and worth exploring further. However, it is also worth flagging some of the limitations of this approach for informing the methodology of normative inquiry.

First, one might reasonably be more confident in certain relevant methodological claims than one is in any metanormative theory determinate enough to have the sorts of methodological implications sketched. This might make it unreasonable to change the former in light of the latter.[81]

Second, the significance of metanormative inquiry for the sorts of normative questions we have been asking in the first three sections of the chapter is arguably limited. For example, if one has decided to focus on *morality*, some of the metamoral theses just canvassed may help to determine whether to deploy folk vs. theoretical concepts. But it is harder to see how these sorts of theses could tell you whether to orient your inquiry around MORALITY as opposed to EXPLOITATION or AUTHORITATIVE OUGHT. These theses might well help us to better understand (a) the sort of thing(s) that we investigate when we investigate morality and (b) how to investigate morality were we to choose to do so. But understanding those things can't by itself settle our decisions

[77] For example, see Railton (1986) and Boyd (1997). Note that moral thought and talk might have some features that made it quite distinctive, despite the metaphysics and epistemology of morality being quite continuous with that of some sciences. See, for example, Copp (2001).

[78] For example, see Shafer-Landau (2003); Fitzpatrick (2008); and Enoch (2011).

[79] See McPherson (2012) for defense of this idea.

[80] For further discussion of this characterization of metanormative inquiry, see McPherson and Plunkett (2017) and Plunkett and Shapiro (2017).

[81] See McPherson (2012) for further discussion of this idea.

about *what* to investigate, even if it might be helpful in understanding what possible things we might investigate.

One way to put the basic point here is as follows. Metamoral inquiry is a hermeneutic enterprise, aimed at explaining our *actual* moral thought and talk (and what, if anything, it is distinctively about). It can't, by itself, tell us that we should be investigating moral thought, talk, and reality, or, more generally, make our decisions for us about what to investigate.[82] Given that we could imagine *that* decision process in instrumental terms, it is not clear how deep the contrast with the instrumental approach really goes here.

To push this line of thought further, consider the choice about which concepts to orient normative inquiry around. We can directly ask a normative question here: *how ought we to make this choice?* One possible answer is: *look to your aims*. But this is only one possible answer. The answer might instead be grounded in facts *independent* of the interests of the inquirer. For example, perhaps some normative truths are intrinsically more valuable to know than others (or some topics intrinsically better to investigate).[83] Or perhaps some kinds of knowledge are more *morally or politically valuable* than others, and thus should be privileged in normative inquiry. For example, perhaps we should orient normative inquiry around certain concepts that help *reveal the basis for an unforced political consensus* among people with widely divergent ethical and political views. Such a consensus might allow such people to live relatively harmoniously together (rather than killing each other, as in the history of religious warfare), and perhaps also allow for a certain kind of freedom.[84] Or perhaps we should aim to uncover theories and practices of reasoning that are "plausible to and usable by moral agents in the case at hand, nonabusive of social power or vulnerability, and capable of delivering feasible conclusions".[85] Or perhaps we should aim to uncover theories that help give "agents a kind of knowledge inherently productive of enlightenment and emancipation".[86]

Of course, we can now apply the dialectic of this chapter to the norms that apply to the question *how ought we to choose the normative concepts to orient normative inquiry around?* For what sort of OUGHT concept should be deployed in this question? If there is a most authoritative concept that applies to normative inquiry, then perhaps it makes sense to structure inquiry around that (modulo the sorts of competing concerns discussed in section 3). On the other hand, suppose there is *not* a most authoritative norm that applies here. Then it appears that, if one is guided by a norm in selecting among one's options, this norm will have no distinctive

[82] We here echo what we say at the end of McPherson and Plunkett (2017). For connected discussion, see Eklund (2017).

[83] For a systematic articulation of this idea for inquiry in general, and not just in the case of normative inquiry in particular, see Sider (2012).

[84] This kind of goal underwrites key parts of the liberal tradition in political theory, as well some of the ideas behind the ideal of "public reason". For relevant discussion, see Rawls (1996); Gaus (2011); and Quong (2013).

[85] Jaggar and Tobin (2013: 413). They articulate this in the context of developing an overall account of moral epistemology, which they take to be rooted in core political ideals of feminism.

[86] Geuss (1981: 2). Geuss takes this to be the central goal of *critical theory* as such, in the vein of the so-called "Frankfurt School" of critical theory. For connected discussion, see Horkheimer (1937/1975) and Habermas (1968/1987).

authority relative to competing norms you might have used for the purposes of selection. On this hypothesis, it would thus seem that arbitrary choice must play some role in the foundations of the conceptual ethics of normative inquiry.[87]

5. Conclusions

Our central aim in this chapter has been to explore two central questions in the conceptual ethics of normative inquiry. First: should normative inquiry be oriented around *folk* or *theoretical* normative concepts? Second: should it be oriented around concepts that are *authority-revealing*, or ones that are *extension-revealing*? Along the way, we have also addressed questions about concept/word pairing. As we have tried to show, there is an important and interesting class of considerations that can be used to support different answers to these questions in conceptual ethics, as well as a complicated further set of questions about what sorts of normative standards we should be using to answer these questions in conceptual ethics. We take this to warrant attention: these questions are philosophically rich, relatively unexplored, and potentially have striking implications for how we should conduct normative inquiry.

Does this mean that normative inquiry needs to halt while we turn our attention to these methodological questions? Not at all. Some methodological inquiry in philosophy appears motivated by the idea that we cannot legitimately proceed in our other inquiries without first rebutting skeptical challenges, or otherwise using methodology to put our inquiry on a more secure footing. Tied to this, the boldest methodologists have Cartesian aspirations of finding indubitable methods and starting points to replace our shabby-seeming ordinary attempts to answer philosophical questions. Others dream of a method that—even if not indubitable—can usefully be followed by all, no matter how benighted their initial opinions. We take such motivations to be misguided. It might well be impossible for us to achieve either of these goals, at least in the near future. It is also far from clear how much we *need* such foundations to have reasonable confidence in the products of our philosophical theorizing. Finally, methodology is itself philosophy, so it is hard to see how it could provide a decisive antidote to worries about the legitimacy of philosophy.

Our approach is rooted in a different understanding of the value of methodology. Being *more philosophy*, methodological inquiry can be interesting and worthwhile for the same reasons other areas of philosophy can be. For example, it might be interesting because we seek understanding of ourselves, the world we live in, and our thought and talk about that world. In this spirit, we find it natural for (at least some) philosophers engaged in normative inquiry to be curious about the activity they are engaged in, and the arguments that can be made for pursuing that activity in one way or another. With this in mind, we hope that this chapter spurs more interest in the conceptual ethics of normative inquiry.

Further, even without Cartesian aspirations, we can hope that methodological reflection will help us to improve our inquiry. We think that this chapter can

[87] Wrestling with this hypothesis is one of the central themes of Eklund (2017). See also Burgess (Chapter 6, this volume) for connected discussion of the charge of 'hypocrisy' in conceptual ethics.

potentially help normative inquirers to do better in at least two ways. First, we take it that normative inquirers not only *could* make a wide variety of choices concerning which concepts to orient their normative inquiry around, but that they are in fact *actually making* different choices here. This diversity of practice is not always clearly signaled (including in some of our own previous work). We suspect that this is partly because the range of relevant options is rarely salient to those inquirers, with the result that there is no felt need to clarify one's target. We think the lack of clear signaling is also partly explained by the lack of a clear vocabulary for communicating the relevant orientation. The discussion in this chapter can help us to understand each other's work better, by providing (what we hope is) a useful framework within which to locate distinct projects. Second, we hope that this chapter paves the way for further methodological reflection that helps to adjudicate some of the central questions that we have asked in this chapter. If so, this may enable some normative inquirers to focus their efforts more successfully on the questions it is most worthwhile to investigate, whatever those turn out to be.

Acknowledgements

Earlier versions of this chapter were presented at University of Oslo (ConceptLab), University of Birmingham, University of Sydney, the NYC Philosophy of Language Workshop, University of Pennsylvania, University of Colorado Boulder, The California Philosophy Workshop, University of Leeds, Hebrew University of Jerusalem, University of Oxford, University of Edinburgh, Central European University, Stanford University, University of Virginia, the Logos Research Group at the University of Barcelona, and the Arché research center at University of St. Andrews. Thanks to everyone who participated in those sessions. Thanks also to Mark Budolfson, Alexis Burgess, Sarah Buss, Herman Cappelen, Matthew Chrisman, David Enoch, Max Etchemendy, David Faraci, Chris Heathwood, Nadeem Hussain, Andrej Kristan, Frank Jackson, Barry Maguire, Eliot Michaelson, Daniel Nolan, Bob Pasnau, Peter Railton, Raul Saucedo, Karl Schafer, Kevin Scharp, Sharon Street, Tim Sundell, Amie Thomasson, and Daniel Wodak.

References

Anderson, Elizabeth. 1999. What Is the Point of Equality? *Ethics* 109 (2):287–337.

Anderson, Elizabeth. 2001. Unstrapping the Straitjacket of 'Preference': A Comment on Amartya Sen's Contributions to Philosophy and Economics. *Economics and Philosophy* 17 (1):21–38.

Appiah, Anthony. 2005. *The Ethics of Identity*. Princeton, NJ: Princeton University Press.

Aristotle. 2002. *Nicomachean Ethics*. Oxford: Oxford University Press.

Baker, Derek. 2018. Skepticism about Ought Simpliciter. In R. Shafer-Landau (ed.), *Oxford Studes in Metaethics*, vol 13. Oxford: Oxford University Press.

Berker, Selim. 2013. Epistemic Teleology and the Separateness of Propositions. *Philosophical Review* 122 (3):337–93.

Blackburn, Simon. 1998. *Ruling Passions*. Oxford: Clarendon Press.

Boyd, Richard. 1997. How to be a Moral Realist. In S. Darwall, A. Gibbard, and P. Railton (eds.), *Moral Discourse and Practice*. New York: Oxford University Press.

Braddon-Mitchell, David. Chapter 4, this volume. Reactive Concepts: Engineering the Concept CONCEPT. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Brandt, Richard B. 1979/1998. *A Theory of the Good and the Right*. New York City: Prometheus Books.

Broome, John. 2015. Reason versus Ought. *Philosophical Issues* 25 (1):80–97.

Burge, Tyler. 1979. Individualism and the Mental. *Midwest Studies in Philosophy* 4 (1):73–121.

Burgess, Alexis. Forthcoming. Never Say 'Never Say "Never" '? In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Burgess, Alexis, and Plunkett, David. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, Alexis, and Plunkett, David. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Burke, Edmund. 1790/1982. *Reflections on the Revolution in France*. New York City: Penguin.

Cappelen, Herman. 2013. Nonsense and Illusions of Thought. *Philosophical Perspectives* 27 (1):22–50.

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Carnap, Rudolf. 1947/1956. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

Carnap, Rudolf. 1950/1962. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Chalmers, David J. 2011. Verbal Disputes. *Philosophical Review* 120 (4):515–66.

Cohen, Stewart. 1984. Justification and Truth. *Philosophical Studies* 46 (3):279–95.

Cohen, Stewart. 2016. Theorizing about the Epistemic. *Inquiry* 59 (7–8):839–57.

Copp, David. 1997. The Ring of Gyges: Overridingness and the Unity of Reason. *Social Philosophy and Policy* 14 (1):86–106.

Copp, David. 2001. Realist-Expressivism: A Neglected Option for Moral Realism. *Social Philosophy and Policy* 18 (02):1–43.

Darwall, Stephen L. 2006. *The Second-person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Darwall, Stephen L., Gibbard, Allan, and Railton, Peter Albert. 1997. Toward Fin de Siecle Ethics: Some Trends. *Moral Discourse and Practice: Some Philosophical Approaches*. New York: Oxford University Press.

Dworkin, Ronald. 2000. *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, MA: Harvard University Press.

Dworkin, Ronald. 2011. *Justice for Hedgehogs*. Cambridge, MA: Harvard University Press.

Eklund, Matti. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.

Enoch, David. 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.

Enoch, David, and Schechter, Joshua. 2008. How Are Basic Belief-forming Methods Justified? *Philosophy and Phenomenological Research* 76 (3):547–79.

Finlay, Stephen. 2014. *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

Finlay, Stephen. 2019. Defining Normativity. In D. Plunkett, S. Shapiro and K. Toh (eds.), *Dimensions of Normativity: New Essays on Metaethics and Jurisprudence*. Oxford: Oxford University Press.

Fitzpatrick, William. 2008. Robust Ethical Realism, Non-naturalism, and Normativity. *Oxford Studies in Metaethics* 3:159–205.

Foot, Philippa. 1972. Morality as a System of Hypothetical Imperatives. *Philosophical Review* 81 (3):305–16.

Foucault, Michel. 1966/2000. *The Order of Things: An Archeology of the Human Sciences*. London: Routledge.

Foucault, Michel. 1980. *Power/Knowledge: Selected Interviews and Other Writings 1972–1977*. New York: Pantheon.

Gaus, Gerald. 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*: Cambridge University Press.

Gettier, Edmund L. 1963. Is Justified True Belief Knowledge? *Analysis* 23:121–3.

Geuss, Raymond. 1981. *The Idea of a Critical Theory: Habermas and the Frankfurt School*. Cambridge: Cambridge University Press.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Gilovich, Thomas. 2001. *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. New York: The Free Press.

Greenberg, Mark. 2006. How Facts Make Law. In S. Hershovitz (ed.), *Exploring Law's Empire: The Jurisprudence of Ronald Dworkin*. New York: Oxford University Press.

Greenough, Patrick. Chapter 11, this volume. Neutralism and Conceptual Engineering. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Habermas, Jürgen. 1968/1987. *Knowledge and Human Interests*. Trans. J. J. Shapiro. Cambridge: Polity.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Nous* 34 (1):31–55.

Haslanger, Sally. Chapter 12, this volume. Going On, Not in the Same Way. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Hills, Alison. 2009. Moral Testimony and Moral Epistemology. *Ethics* 120 (1):94–127.

Horkheimer, Max. 1937/1975. Traditional and Critical Theory. In *Critical Theory: Selected Essays*. New York City: Continuum

Huemer, Michael. 2001. *Skepticism and the Veil of Perception*. Lanham: Rowman & Littleeld.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon.

Jaggar, Alison M., and Tobin, Theresa W. 2013. Situating Moral Justification: Rethinking the Mission of Moral Epistemology. *Metaphilosophy* 44 (4):383–408.

Jenkins, C. S. I. 2014. Merely Verbal Disputes. *Erkenntnis* 79 (1):11–30.

Jenkins, C. S. 2006. The Rules of Flirtation. *The Philosophers Magazine* 36:37–40.

Jones, Karen. 2005. Moral Epistemology. In F. Jackson and M. Smith (eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. New York: Cambridge University Press.

Kratzer, Angelika. 2012. *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.

Littlejohn, Clayton. forthcoming. A Plea for Epistemic Excuses. In J. Dutant and F. Dorsch (eds.), *The New Evil Demon Problem*. Oxford: Oxford University Press.

Ludlow, Peter. 2014. *Living Words: Meaning Undetermination and the Dynamic Lexicon*. Oxford: Oxford University Press.

McPherson, Tristram. 2012. Unifying Moral Methodology. *Pacific Philosophical Quarterly* 93 (4):523–49.

McPherson, Tristram. 2018. Authoritatively Normative Concepts. In R. Shafer-Landau (eds.), *Oxford Studies in Metaethics*, vol 13. Oxford: Oxford University Press.

McPherson, Tristram, and Plunkett, David. 2015. Deliberative Indispensability and Epistemic Justification. In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*, vol. 10. Oxford: Oxford University Press.

McPherson, Tristram, and Plunkett, David. 2017. The Nature and Explanatory Ambitions of Metaethics. In T. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics*. New York: Routledge.

Moore, G. E. 1903/1993. *Principia Ethica*. Cambridge: Cambridge University Press.

Nolfi, Kate. 2014. Why is Epistemic Evaluation Prescriptive? *Inquiry* 57 (1): 97–121.

Nolfi, Kate. MS. Epistemically Flawless False Beliefs.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Pettit, Philip. 2012. *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge: Cambridge University Press.

Plunkett, David. 2015. Which Concepts Should We Use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 58 (7–8):828–74.

Plunkett, David. 2016. Conceptual History, Conceptual Ethics, and the Aims of Inquiry: A Framework for Thinking about the Relevance of the History/Genealogy of Concepts to Normative Inquiry. *Ergo* 3 (2):27–64.

Plunkett, David, and Shapiro, Scott. 2017. Law, Morality, and Everything Else: General Jurisprudence as a Branch of Metanormative Inquiry. *Ethics* 128 (1):37–68.

Plunkett, David, and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Portmore, Douglas W. 2013. Perform Your Best Option. *Journal of Philosophy* 110 (8):436–59.

Prior, A. N. 1960. The Runabout Inference-Ticket. *Analysis* 21 (2):38.

Quong, Jonathan. 2013. Public Reason. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (summer edn). https://plato.stanford.edu/archives/sum2013/entries/public-reason/

Railton, Peter. 1986. Moral Realism. *The Philosophical Review* 95:163–207.

Railton, Peter. 2003. Morality, Ideology, and Reflection, or the Duck Sits Yet. *Facts, Values, and Norms: Essays Toward a Morality of Consequence*. New York: Cambridge University Press.

Rawls, John. 1971/1999. *A Theory of Justice* (revised edn). Cambridge, MA: Harvard University Press.

Rawls, John. 1996. *Political Liberalism* (expanded edn). New York: Columbia University Press.

Roberts, Debbie. 2017. Thick Concepts. In T. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics*. New York: Routledge.

Satz, Debra. 2010. *Why Some Things Should Not Be for Sale: The Moral Limits of Markets*. New York: Oxford University Press.

Scanlon, T. M. 1998. *What We Owe To Each Other*. Cambridge, MA: Harvard University Press.

Scharp, Kevin. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.

Schroeder, Mark. 2018. The Unity of Reasons. In D. Star (ed.), *The Oxford Handbook to Reasons and Normativity*. Oxford: Oxford University Press.

Schroeter, Laura. 2008. Why Be an Anti-Individualist? *Philosophy and Phenomenological Research* 77 (1):105–41.

Schroeter, Laura, and Schroeter, Francois. 2014. Normative Concepts: A Connectedness Model. *Philosophers' Imprint* 14 (25):1–26.

Sen, Amartya. 2009. *The Idea of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Seuss, Dr. 1965. *I Had Trouble in Getting to Solla Sollew*. New York: Random House.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*: Oxford University Press.

Shah, Nishi, and Velleman, David J. 2005. Doxastic Deliberation *Philosophical Review* 114 (4):497–534.

Shapin, Steven. 1994. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Chicago: University of Chicago Press.

Sider, Theodore. 2012. *Writing the Book of the World*. Oxford: Oxford University Press.

Sidgwick, Henry. 1874. The Methods of Ethics.

Sinnott-Armstrong, Walter. 2006. *Moral Skepticisms*. Oxford: Oxford University Press.

Smith, Michael. 1994. *The Moral Problem*. Cambridge: Blackwell.

Smithies, Declan. 2012. Moore's Paradox and the Accessibility of Justification. *Philosophy and Phenomenological Research* 85 (2):273–300.

Strawson, Peter F. 1963. Carnap's Views on Conceptual Systems versus Natural Languages in Analytic Philosophy. In P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*. La Salle: Open Court.

Sylvan, Kurt. MS. On Divorcing the Rational and the Justified in Epistemology.

Thomasson, Amie L. 2016. Metaphysical Disputes and Metalinguistic Negotiation. *Analytic Philosophy* 57 (4).

Thomasson, Amie L. Chapter 21, this volume. A Pragmatic Method for Normative Conceptual Work. In A. Burgess, H. Cappelen and D. Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Tiffany, Evan. 2007. Deflationary Normative Pluralism. *Canadian Journal of Philosophy* 37 (5):231–62.

Väyrynen, Pekka. 2013. *The Lewd, the Rude and the Nasty: A Study of Thick Concepts in Ethics*. Oxford: Oxford University Press.

Velleman, J. David. 1988. Brandt's Definition of "Good". *Philosophical Review* 97 (3):353–71.

Weatherson, Brian. 2008. Deontology and Descartes's Demon. *Journal of Philosophy* 105 (9):540–69.

Wedgwood, Ralph. 2007. *The Nature of Normativity*. Oxford: Oxford University Press.

Wedgwood, Ralph. 2012. Justified Inference. *Synthese* 189 (2):1–23.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Wood, Allen W. 1999. *Karl Marx*. London: Routledge.

15

Conceptual Evaluation
Epistemic

Alejandro Pérez Carballo

My topic here is the question of conceptual evaluation along an epistemic dimension. My question is: what makes for good concepts, epistemically?

Thus formulated, this question cries out for clarification in a number of ways. And I will turn to that in due course. But it will help to get some sense of where we're going before getting started.

The distinction between epistemic and non-epistemic evaluations of belief is not only familiar, but rather intuitive. In part, I suppose, that is because we have a reasonable grasp of ways of evaluating beliefs—in terms of their truth, or their warrant—that are plausibly classified as 'epistemic'. There is no similar, pre-theoretically available way of evaluating concepts that is uncontroversially classified as epistemic. Still, it seems uncontroversial that scientific progress often involves introducing new conceptual tools. And this suggests that scientific progress sometimes involves conceptual progress—that better science incorporates better concepts.

What could make some concepts epistemically better than others? On what is perhaps the orthodox answer to this question—the answer that most philosophers I have encountered are inclined to give, once they get past their reluctance to take talk of concepts at face value without some heated preliminary discussion—the issue turns on what properties correspond to the concepts in question. A concept is better than another, epistemically, if and to the extent that the property corresponding to one is 'more natural' than the one corresponding to the other. This line is often accompanied by a quick reference to Goodman—'you mean like how "blue" is a better concept than "grue"?'—or to some discussion of natural kinds.

My proximate aim in this chapter is to argue that this answer cannot be right. More generally, I want to argue against a particular way of addressing this question—an approach that leans heavily on metaphysical, non-epistemic notions like 'structure' or 'naturalness'. But the ultimate goal is to make some progress in clarifying what a theory of conceptual evaluation should hope to accomplish.

As it happens, I am skeptical that there is a theoretically useful, non-epistemic notion of 'naturalness' or 'structure' out there. But here I will simply take for granted that the relevant metaphysical notions are in good order. In short, I want to grant for the purposes of this chapter that there is a theoretically useful notion of 'naturalness'

Alejandro Pérez Carballo, *Conceptual Evaluation: Epistemic* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Alejandro Pérez Carballo.
DOI: 10.1093/oso/9780198801856.003.0015

according to which some properties are natural and some are not—or better yet, that some properties are more natural than others. (I will assume, however, that the relevant notion of 'naturalness' is either a primitive notion or can be defined in purely non-epistemic terms.) My claim will be that there is no straightforward connection between a theory of epistemic evaluation of concepts and a theory of 'natural' properties thus understood.

1 Preliminaries

A number of authors have recently made explicit claims in what we can call 'applied conceptual ethics':[1] claims to the effect that some concept is better—in some sense or other—than another. Sally Haslanger, to take a well-known example, has defended a number of theses about "how we might usefully revise [our concepts of race and gender] for certain political and theoretical purposes" (Haslanger 2000: 34). On her view, there are legitimate purposes that will be better served by our using 'woman' so that (very roughly) *x is a woman* conceptually entails *x is subordinated (along some dimension or other)* (p. 39ff). Similarly, Kevin Scharp has argued that we should 'replace' our concept of truth with two concepts which, together, are better than our (allegedly inconsistent) ordinary concept for the purposes of 'serious theorizing' (Scharp 2013: 134).

These are not isolated cases, mind you. Peppered throughout the literature on the so-called species problem[2] in philosophy of biology one often finds claims to the effect that a particular species concept is better than another for certain theoretical purposes.[3] And one only needs to look back at different applications of the Carnapian method of explication to see that evaluative claims about concepts and vocabularies are a fairly common occurrence in contemporary philosophy.[4]

Often, as in the examples above, such comparative claims are relativized to specific purposes. The presumption then is that a concept cannot be good or bad *simpliciter.* But this is by no means the only game in town. Ingo Brigandt, for example, has argued that many concepts should be understood as having a built-in *epistemic goal,* which "consists of those theoretical, explanatory, or investigative aims that are pursued by this concept's use" (Brigandt 2010: 36). For example, according to Brigandt, the explanation of patterns of inheritance is an epistemic goal *constitutive* of the classical gene concept—the gene concept associated with Mendel's work and to the research program led by Thomas Hunt Morgan (Morgan et al. 1915). Note that the claim here is not just that the concept was introduced in order to predict inheritance patterns. Rather, the claim is that a concept that is not put to use for the purposes of predicting inheritance patterns would not *be* the classical gene concept of Mendelian genetics.[5] If Brigandt is right—and here I do not take a

[1] The terminology here is borrowed from A. Burgess and Plunkett (2013a).

[2] Kitcher (1984); Mayr (1996); Ereshefsky (2008); *inter alia.*

[3] See, for example, Franklin-Hall (2007) and references therein.

[4] For a more systematic overview of some of the relevant literature, see A. Burgess and Plunkett (2013a,b). For illuminating discussion of Carnap's ideal of explication and its relation to some of the work on the species problem in philosophy of biology, see Kitcher (2008).

[5] This at least is suggested by some of his remarks. For example, Brigandt (2010: 36): "each theoretical concept, at least each central biological concept, consists of three components of content: (1) the concept's

stand on that—then one can make sense of a non-instrumental way of evaluating concepts: we can evaluate how good a concept is relative to its constitutive epistemic goal.[6]

In this chapter, however, I do not aim to advance or take issue with specific claims about which concepts are better than others. Nor do I wish to discuss whether conceptual evaluation must always take place against the backdrop of a specific purpose. My concern, rather, is with an explanatory question that, as far as I can tell, has received little attention in the literature.[7] In brief, the question is: what makes for good concepts, epistemically? But let me say something about each of the key terms that figure in it, to get a better sense of what an answer to our question will look like.

1.1 'Makes'

Take any specific claim about some concept being better than another (whether or not *better than* is relativized to a specific purpose). To a first approximation, the question I'm interested in takes the following form: in virtue of what is the one better than the other? Or: what is it *because of which* one is better than the other?[8]

Now, it might appear as if the answer to this question is straightforward, at least if we are interested in comparisons among concepts relative to a specific purpose. For if what we want is to explain why a concept is better than another relative to a given purpose, the answer should be because the one is a better *means* to achieving said purpose than the other. In other words, if conceptual evaluation is always relative to a goal or purpose—whether or not that goal is constitutive of the relevant concepts—it is merely a matter of instrumental evaluation. And what explains how good something is relative to a given purpose is just that it contributes, in the right sort of way, to that purpose.

True, on this way of thinking there would still be room for disagreement as to what 'the right sort of way' is. Should we think that what makes a concept better than another for a given purpose is that using the one is more likely to causally bring about

reference, (2) its inferential role, and (3) the epistemic goal pursued with the concept's use". See also the discussion in Brigandt (2003), where he argues that different branches of biology are best understood as having different homology concepts, partly because the homology concepts used in different branches are put to distinct theoretical uses.

[6] See Thomasson (Chapter 21, this volume) for an alternative take on this strategy—if concepts have constitutive functions, one can evaluate concepts in terms of how well they serve the relevant functions.

[7] Admittedly, it is not entirely clear whether the explanatory question I will be interested in is entirely neutral on the more 'first-order' questions about which concepts are better than others. The analogous question about the relationship between meta-ethics and normative ethics has been the subject of some interesting discussion recently (see, e.g., Enoch 2011; Fantl 2006). I do not know what the outcome of that debate will be. But I expect whatever considerations bear against (or in favor) the claim that our answers to first-order normative questions can be constrained by answers to questions in metaethics can also be brought to bear on the question whether the explanatory question about conceptual evaluation puts any constraints on the first-order debate. That said, I expect that, whether or not these projects constrain one another, many first-order claims about some concept being better than another will be compatible with different answers to the explanatory question I will be interested in.

[8] Cf. Fine (2012); Rosen (2010). For illuminating discussion of the role of such *because* claims in normative theory and their relationship to the more 'metaphysical' discussion of grounding and the *in virtue of* relation, see Berker (2018).

the achievement of that purpose than using the other one? Or should we instead think of the relevant notion of contribution in non-causal terms?

But it would be a mistake to think that in relying on a given purpose in order to compare different concepts we are thereby presupposing some type of teleological story of what makes a concept better than another. After all, one could think of the relevant purposes as determining a particular dimension of evaluation, rather than a goal or something to be promoted. This would leave it open whether the relevant purposes are being used to pick out the particular ranking of concepts or whether they are doing something more—for example, whether they are part of what explains why the concepts are ranked in that particular way.[9]

This becomes especially apparent when we focus, as I will, on questions about conceptual evaluation for *theoretical purposes*. I would be surprised if authors writing about a concept being better than another for theoretical purposes were resistant to being paraphrased as writing about evaluation of concepts from an epistemic perspective. And we can clearly make sense of the possibility of evaluating beliefs, say, from an epistemic perspective without taking on a consequentialist or teleological view on epistemic rationality.

Our question—what makes for good concepts, epistemically—can then be reformulated as: what is it in virtue of which some concepts are epistemically better than others? Or: what explains why some concepts are better than others, epistemically?

1.2 'Concepts'

It should not surprise you that philosophers can mean very different things by 'concept'. And it should surprise you even less that cognitive scientists and philosophers also use the term in importantly different ways—the former use it to refer to so-called mental representations, the latter (often) to mean what gives such representations their representational import.[10] It is hard to expect much agreement, then, on questions of conceptual evaluation unless we specify at the outset what we're taking concepts to be.[11]

Now I have said little about what I take concepts to be. This is by design—I intend to remain as neutral as I can on the nature of concepts. The terrain surrounding these issues is slippery: one cannot enter it without getting entangled with questions about modularity, the 'grain' of content, and so on.[12] Still, I cannot avoid making some assumptions about what concepts are. My hope is that the assumptions I will make are rather uncontroversial. Fortunately, as we will see, for the purposes of this chapter

[9] Here, it is helpful to make an analogy with the debate over whether all moral theories admit of a 'consequentialist' counterpart. One could grant that any deontological theory can be formulated as a version of consequentialism without thinking that the consequences of a given action play a role in explaining why it has the moral properties that it does. See Portmore (2009) and references therein. Cf. also the discussion in Hammond (1988); Stalnaker (2002).

[10] See the Introduction to Margolis and Laurence (1999) for a guided tour of the issues.

[11] Things would be far worse, I suspect, if our interest were in questions of conceptual *engineering*: of creating and changing concepts. For those questions, it is hard to see how to make much progress without substantive assumptions about what concepts are. After all, on some views (e.g. Fodor 1975), changing a concept is entirely out of the question. Here, however, I will have nothing to say about conceptual change—my focus will be exclusively on conceptual evaluation.

[12] For a helpful overview, see Margolis and Laurence (2014).

it will turn out not to matter how we answer many of the questions about the nature of concepts.

First, I will be assuming that concepts are the sorts of things that words mean. This need not imply, of course, that every concept corresponds to the meaning of some word in English, nor that the meaning of every word in English is given by a concept. But I intend talk of concepts to be substitutable throughout with talk of meanings, so that, for example, talking about the epistemic merits of the concept *species*[13] is tantamount to talking about the epistemic merits of what the word 'species' means. Second, and partly as a consequence of our first assumption, I will presuppose that concepts are not mental representations but rather what would give such putative representations their representational import. In talking about concepts, then, I need not presuppose anything like the Language of Thought Hypothesis: as long as it makes sense to talk about the meaning of a word, everything I will say is neutral on questions about cognitive architecture. Third, I will assume that concepts have non-trivial representational content, and that the epistemic merits of a given concept are at least in part determined by their representational content. I am thus ruling out for present purposes views on which, for example, some of our 'predicative' concepts do not correspond to a property (except perhaps in a pleonastic sense), at least if the properties in question are supposed to do any explanatory work.[14]

1.3 'Good, Epistemically'

I'm interested in a particular type of evaluation of concepts, what I've been calling *epistemic* evaluation. I want to say a bit more about what I have in mind by that.

I think we can get a handle on what epistemic evaluation of concepts amounts to by taking seriously the idea that better science by and large incorporates better concepts. The hope is that any grip we have on the notion of *epistemically* better science will allow us to get a handle on a way of evaluating concepts that is uncontroversially epistemic in nature. For our purposes, then, a broadly consequentialist strategy will suffice. I will assume that concepts that are conducive to achieving what are clearly epistemic aims and purposes are better, epistemically, than those that are not.

What are the relevant goals and purposes? I want to leave that open. I will only assume that our epistemic goals include forming theories that have the familiar theoretical virtues: accuracy, explanatoriness, fruitfulness, etc. How exactly each of these should be cashed out is a good question, but one I will set aside for present purposes.

At any rate, remember that my question is not: *which* are the good concepts, epistemically? Here I will take for granted that we have a reasonably good sense of

[13] Throughout, I will use italics to name concepts, relying on context to disambiguate from other uses. I reserve the use of small caps for labeling theses, like SIMPLE below.

[14] I have in mind expressivist or non-cognitivist treatments of many 'predicative' concepts (views like those in Gibbard 1990; Blackburn 1998; Yalcin 2011; Pérez Carballo 2016; among others). Presupposing these views away only makes my work here harder: after all, if we hoped for a unified theory of conceptual evaluation, and some concepts that should fall under the scope of such a theory cannot be understood as determining or corresponding to a property, it would be a mistake to build our theory of conceptual evaluation on facts about the corresponding properties.

which concepts best help us achieve our epistemic goals. My question is rather: what is it about those concepts that makes them good concepts, epistemically? What, in other words, is it that explains the fact that they are the most conducive to our achieving our epistemic goals? (What is it *in virtue of which* good concepts are good?)

2 The Simple Hypothesis

What I will call the *simple hypothesis* is perhaps the first answer to our question that comes to mind. The hypothesis starts out with a metaphor. The rough idea is that some properties, unlike others, correspond to true 'joints' of nature. This can be motivated in more than one way, but here is one that is relatively easy to state.

Suppose you think that for any set of things—including mere *possibilia*, if you like—there is a property corresponding to that set: elements of the set are exactly those things that instantiate the relevant property. (You might think that properties *just are* such sets, but you need not.) Then you will have to agree that any two things—really, *any* two things—have uncountably many properties in common, and thus (or so it seems) are as alike to one another as any two other things. (A few missing premises are left implicit in the preceding—not because they are unimpeachable, but because I do not intend to argue that the solution to this problem I sketch below is the only way out.) And that, we can agree, would be madness: staplers and raccoons, to pick just one pair, have little in common, or at any rate much less in common with one another than two peas in a pod.

As is well known, you can avoid this conclusion if you insist that, among the uncountably many ways of grouping things together, a relatively small class of them has special status—and correspondingly, a small class of properties has special status. It is the sharing of one or more of *those* properties that makes for genuine similarity. Peas in a pod share many more of those special-status properties—and are thus much more like one another—than a stapler and a raccoon do.

The idea of a metaphysical distinction among properties—'natural/non-natural', 'sparse/abundant', 'joint-carving/gerrymandered', and so on—can be fleshed out in multiple ways.[15] But at this point we have enough materials to formulate a schematic version of the simple hypothesis (I will use 'elite' as a stand-in for any one of 'natural', 'sparse', 'joint-carving' and so on):

> SIMPLE (FIRST PASS): Concepts are better, epistemically, because and to the extent that they correspond to elite properties.

Again, it is worth highlighting the fact that the notion of 'eliteness' is supposed to be non-epistemic in nature.[16]

[15] Armstrong (1978); Lewis (1983); Schaffer (2004); Sider (2013), See also Dorr and Hawthorne (2013) and references therein.

[16] The claim thus is not merely that what makes concepts better than others, epistemically, is a purely 'objective' matter. (Exactly what this means is a good question, but glossing it with the familiar cluster of metaphors will have to do for now. Cf. Rosen 1994.) Rather, the presumption is that the notion of 'eliteness' is, in a sense that would need to be made precise, a purely metaphysical one.

This simple hypothesis has at least two considerable virtues (beyond simplicity, that is). First, it seems to get something right about our pretheoretic judgments about the comparative merits of some concepts. Intuitively, there is something epistemically better about our concept *green* over the concept *grue.* And it certainly *seems* as if this is due to the fact that whereas all green things genuinely have something in common, not all grue things do. Second, the simple hypothesis offers a reduction of one vexed question—which concepts are epistemically better?—to a question which, if no less vexed, is at least supposed to be independently pressing. We need, the story goes, the elite/non-elite distinction for other purposes—for giving an account of Laws of Nature and for giving a theory of intentionality, to name a few. It would be beautifully economical if we could also appeal to it in order to give an account of conceptual evaluation.

Of course, the simple hypothesis, as stated, cannot be right. The notion we are after is a comparative notion—that of a concept being epistemically better than another. And as long as we think, as I think we must, that not all good concepts are equally good, we will want to make finer distinctions than those afforded by a binary elite/non-elite distinction.

On the face of it, this is not an insurmountable obstacle. For one might in principle define a suitable relative of 'eliteness' that is comparative in nature. This would essentially involve specifying a 'more elite than' relation that applies to properties, and replacing SIMPLE (FIRST PASS) with:

> SIMPLE: A concept is epistemically better than another if and because the property corresponding to the former is more elite than the one corresponding to the latter.

(Again, on this picture, the 'more elite than' notion would need to be specified in purely non-epistemic terms. The goal, recall, is to reduce the question of epistemic evaluation of concepts to questions about the purely metaphysical merits of the corresponding properties.)

But already at this level of abstraction, we run into difficulties.[17] The first and most straightforward one is this. Elite properties, we are sometimes told, make up a minimal supervenience base for all facts.[18] As a result, high-level properties will not be elite properties. Now, this need not mean that all high-level concepts are equally bad. The 'more elite than' relation should presumably make distinctions among non-elite properties. But it does mean that the concepts of fundamental physics are epistemically better than the concepts of, say, psychology. And while

[17] Admittedly, many of the problems I will flag in the remainder of this section arise because of the decidedly Lewisian understanding of eliteness I am working with. An alternative gloss on the relevant notions—one that takes seriously the lessons from Fodor (1974) (see also Antony 2003: esp. p. 12f)—would allow for perfectly elite properties not only at the most fundamental level (say, the level of fundamental physics), but also at the level of chemistry, biology, etc. Cf. Schaffer (2004), and what he labels the 'scientific conception' of elite properties: "On the scientific conception, the properties invoked by total science are ontologically on par. All carve out joints of nature. Muons, molecules, minds, and mountains are in every sense equally basic" (p. 94). Ultimately, I will grant the proponent of SIMPLE that there is some conception of eliteness that is in reasonably good order, so I will set aside views like Fodor's for the sake of brevity. As should become clear, my arguments in §3 apply equally to both a Lewisian and a Fodorian view of eliteness.

[18] See, for example, Lewis (1984, 1983). For an argument to the contrary, see Eddon (2013). Cf. also Dorr and Hawthorne (2013: 10–13).

one might be willing to grant that there are some—or even *many*—epistemic dimensions along which the concepts of fundamental physics are epistemically better than those of psychology, it is hard to believe that there is no epistemic dimension along which psychological concepts are not worse than those of fundamental physics.

The second, related difficulty is a bit less straightforward, for it can only be formulated against particular assumptions about the fine structure of the 'more elite than' relation. At a very abstract level, the worry is that even if we grant that concepts corresponding to perfectly elite properties are better than the rest, it isn't clear that a plausible way of defining 'more elite than' for non-elite properties will mesh with reasonable assumptions about the relative merits of high-level concepts.

It is easier to see this with a specific proposal for how to understand the 'more elite than' relation. Suppose that we think, as some do, that a property is more elite than another iff its simplest definition in terms of fundamental properties is less complex than the simplest definition of the other (for some suitable notion of simplicity).[19] Consider now your favorite example of a concept of a multiply realizable property—say, the concept of pain. Presumably, the corresponding property will admit of a characterization as a disjunction, each of whose disjuncts corresponds to a way in which the property could be realized. It follows now from our toy theory of 'more elite than' that a concept for the property obtained by removing one or more disjuncts from the initial property will be epistemically better than the initial concept. So, the concept of pain-in-carbon-based-beings, for example, would turn out to be epistemically better than the concept of pain. And this, surely, should cast doubt on the current proposal.[20]

Now, all of this might be avoided if instead of starting with a primitive elite/non-elite distinction we start with a comparative notion specified in non-epistemic terms.[21] We could, for example, start with the assumption that there are objective facts about genuine similarity—for example, the fact that a horse and a rhino are more similar to one another than a raccoon and a stapler are to one another. And we could use these in turn to get both an elite/non-elite distinction *and* a notion of 'more elite than' which may well avoid some of the earlier difficulties.

For example, once we're allowed access to facts about objective similarity, we can define a 'more elite than' relation as follows: property *F* is more elite than *G* iff the most dissimilar *F*s are more similar to one another than the most dissimilar *G*s.[22] Perfectly elite properties, on this view, would just be the properties that are maximally elite.[23]

[19] Lewis made something like this suggestion, in passing, in Lewis (1984: 228). For a discussion of some of the difficulties involved in carrying this out in reasonable detail, see Sider (2013: section 7.11.1).

[20] A more subtle example, borrowed from Hall (2011): *being methane* is a much less useful concept, epistemically, than *being a saturated hydrocarbon,* but the 'canonical' definition of the latter will be much more complicated than the one of the former. Cf. Hall (2011: 21ff).

[21] Or if, following some speculative remarks in Sider (2013), we also allow for primitives other than 'elite' in our definition of 'more elite than', for example, *law-like.*

[22] Cf. Rosen (2015: 191).

[23] A notion of similarity could also be appealed to in giving a definition of eliteness directly, by letting elite properties be *convex*, in the sense that whenever *x* and *y* are more like one another than *y* and *z* are, then if *x* and *z* have property, so does *y*. Cf. Gärdenfors (2000: 70f), as well as Oddie (2005: 152ff). For discussion of the issues arising from putative connections between naturalness and similarity, see Dorr and Hawthorne (2013: 21–7).

Still, to the extent that we think that, by and large, a property is more elite than another iff it is more fundamental, we will be hard pressed to vindicate the thought that, at least along some epistemic dimension, concepts from special sciences are better for some purposes.

But let us suppose, for the sake of argument, that we can get an account of eliteness that is not compromised from the outset. The question is whether we can build a plausible theory of conceptual evaluation on top of it.

3 Against the Simple Hypothesis

I think SIMPLE cannot be right. I want to rehearse a few arguments to that effect. Ultimately, there are things a believer in SIMPLE can say in response to each of them. I will mention some of them. Thus I do not presume that any of the arguments here are decisive. The hope rather is that all of them, together, can be taken as motivation to seek an alternative story—one that might too be sensitive to the metaphysical contours of our objects of inquiry, but which will not be determined by it.

3.1 Initial Skepticism

To a naive reader, the strategy behind SIMPLE might not seem very promising. For it seems eminently reasonable to insist that not all concepts can be said to correspond to a property or relation.[24] Singular concepts, quantificational concepts, logical concepts—to name a few—are not naturally construed as corresponding to some property we could go on to evaluate for eliteness. And yet we would be giving up the game if we were to rule out all such concepts as being candidates for conceptual evaluation, or if instead we gave up on giving a unified theory of conceptual evaluation.

Yes, we cannot rule out the possibility of generalizing a theory of eliteness so that it encompasses more than properties and relations.[25] But we should acknowledge—*pace* Sider—the awkwardness of thinking that the epistemic benefits afforded by the logical vocabulary our best theories rely on is best explained by appealing to comparisons between the metaphysical counterparts of (say) conjunction and the Sheffer stroke.[26]

None of this is to say that the burden is squarely on the side of SIMPLE and its ilk. But it might be enough to give us pause before embracing optimism about the prospects of SIMPLE.

Another reason for being skeptical of the strategy recommended by SIMPLE comes from general considerations about the possibility of explaining normative phenomena in purely non-normative terms.[27] The view that concepts corresponding to elite properties are better epistemically than those corresponding to less elite ones gives

[24] Not that this cannot be done, in principle. For a sense of how such a story could go, see J. P. Burgess (2005); Dasgupta (2009).

[25] See Sider (2013) for what is perhaps the most sophisticated attempt at doing that.

[26] Cf. Taylor (1993: 99): "Even granted that the *physical* world might come jointed, the notion that what makes adding more natural than quadding is the prior jointedness of mathematical reality, rather than the way we think about it, is peculiarly unappealing."

[27] See Greco (2015) for a helpful discussion of the extent to which 'open question'-style arguments carry over to debates in meta-epistemology.

rise to the question: why is it better, epistemically, to use concepts that correspond to elite properties? And while there may well be an answer to that question that does not implicitly rely on other normative or evaluative notions, I think it would not be surprising if such an answer is not forthcoming.[28]

Building a case against SIMPLE on the basis of these two considerations would require a fair amount of work.[29] But I do not intend to pursue that here. Rather, I want to focus on a few additional considerations which, together, should at least call into question the strategy recommended by SIMPLE.

3.2 Variety of Explanatory Purposes

Consider the following claim:

> EXPLANATION. All else equal, a concept is epistemically good to the extent that it figures in good explanations.

Plausibly, this should be a consequence of any reasonable theory of conceptual evaluation. Is it a consequence of SIMPLE?

If 'explanation' in SIMPLE is read as *metaphysical* explanation, then EXPLANATION may seem to follow from SIMPLE, at least given certain reasonable assumptions.[30] Assume, first, that metaphysical explanation is a transitive relation among facts. Next assume that if x explains y, then x is *more fundamental* than y. Finally assume that x is more fundamental than y to the extent that properties 'involved' in x are more elite than those involved in y.[31] Then we might be able to conclude that the more explanations appeal to fact x the more elite the properties involved in x are.

This is still not quite what we want. At best we could conclude that the concepts involved in good explanations are likely to correspond to more elite properties, and thus (according to SIMPLE) that they are likely to be good concepts. We could not yet conclude that that all good concepts figure in some good explanations.[32] But at least something along these lines might work.

[28] Sider (2013) anticipates some of these concerns and claims without argument that aiming to have 'joint-carving beliefs' is "a constitutive aim of the practice of forming beliefs, as constitutive as the more commonly recognized aim of truth" (p. 61). But as Hazlett (2017) points out, few if any of the arguments that have been offered for the claim that truth is a 'constitutive aim' of belief carry over to support the claim that having 'joint-carving beliefs' is a constitutive aim of (the practice of forming) belief(s).

[29] See Dasgupta (2018) for a careful attempt at carrying some of this out.

[30] What, on a metaphysical conception of explanation, makes for *good* explanations? That is an excellent question, but one I have nothing to say about here.

[31] This is at best merely a gesture in the right direction. For one, because the claim that a property is involved in a fact needs to be cashed out in more detail—perhaps by assuming that facts are structured entities. But more importantly, because most facts will presumably involve properties of different degrees of eliteness, and we need an argument for thinking that x is more fundamental than y only if the least elite property involved in x is more elite than the most elite one involved in y. After all, a no less plausible hypothesis is that x is more fundamental than y only if the most elite property involved in x is more elite than the most elite property involved in y. (Note too that things get much more complicated very quickly if we allow for facts involving infinitely many properties, or if we allow for the possibility that the 'more elite than' relation is not a well-ordering.) Thanks here to Maya Eddon.

[32] If we think of facts as comprising a structure partially ordered by 'is more fundamental than', we could in principle have facts that are only 'one-step' above perfectly fundamental facts but on top of which we will find no other facts they explain. This would give rise to properties that are almost perfectly elite but

We run into more significant difficulties, though, once we recall that good explanations are not explanations tout *court,* but rather explanations *of* something. For, presumably, not everything is equally in need of explanation.[33] To take a simple example, consider the contrast between the following two events:

(1) Nancy, Tom, and William—three complete strangers—each win one of the three prizes on the raffle at the town fair.
(2) Alice, Barbara, and Carol—three sisters that came together to the fair—each win one of the three prizes on the raffle at the town fair.

The second of these two cries out for explanation in a way that the first one does not. And, plausibly, the more a fact cries out for explanation the more valuable an explanation of that fact will be. So we should expect that good concepts will figure in good explanations *of what is in need of explanation.* Or, better perhaps: better concepts will figure in good explanations of explananda in need of explanation.[34]

This by itself may not be enough of a concern. A proponent of SIMPLE might, after all, have a story about what makes a fact cry out for explanation that does not in turn rely on epistemic considerations. Trouble is, as far as I can tell no such story has ever been told. And it is not clear how such a story could be told. For whether something cries out for explanation or not seems to be highly context-sensitive—whether some event cries out for explanation seems to depend in part on something like an epistemic state or a body of theory.

To my knowledge, not much has been written on the question what makes something cry out for explanation.[35] But three families of proposals seem to have emerged from the relatively small discussion.

which do not figure in any metaphysical explanations. Concepts corresponding to that property would thus be good concepts, according to SIMPLE, but not good according to EXPLANATION.

[33] This claim is not entirely uncontroversial. See, for example, Friedman (1974: 13): "All phenomena, from the commonest everyday event to the most abstract processes of modern physics, are equally in need of explanation—although it is impossible, of course, that they all be explained at once." Friedman does not offer an argument for this claim. Instead, he claims he "cannot see any reason but prejudice for regarding some phenomena as somehow more natural, intelligible, or self-explanatory than others" (p. 13).

[34] For examples of the way in which the choice of concepts is sensitive to what is deemed worthy of explanation, see the discussion the 'category influence hypothesis' in Franklin-(2015: 933)—see especially the discussion of how classificatory practices in chemistry, based on atomic numbers, are partly the result of chemists' interest in explaining and accounting for certain material transformations, as opposed to the behavior of material in centrifuges. Cf. also Hendry (2010: 147): "eighteenth-century chemistry and its nomenclature were shaped by interests in the qualitative patterns of particular kinds of chemical behaviour (combustion, calcination, acid-base reactions) and explaining them in terms of a particular conception of elemental composition. These patterns are determined by sameness and difference in nuclear charge, and quite insensitive to sameness and difference in atomic weight." Another cluster of examples can be found in the discussion of the species problem in philosophy of biology. Kitcher (1984), for example, identifies nine different species concepts, each of which is best suited to a particular explanatory purpose—the presumption then is that different explanatory goals result in different assessments as to the epistemic benefits of a particular concept. Cf. also Stanford (1995).

[35] For a helpful overview of the rather small literature on this topic, see Grimm (2008). See also Wong and Yudell (2015). Related issues are sometimes discussed in the context of finetuning arguments—there, the question is whether the fact that the universe is capable of sustaining life cries out for explanation—in debates over whether the laws of nature or the universe's 'initial conditions' themselves stand in need of explanation, and in the debates over physicalism in the philosophy of mind, where the question is whether

First, there's what Grimm (2008) calls the 'surprisingness' account.[36] A fact cries out for explanation, on this view, to the extent that it is unexpected, or surprising. On the face of it, this seems like a promising view. But it is hard to cash it out so that it isn't vulnerable to obvious counterexamples.[37] We are all too familiar with the fact that the universe appears to be perfectly fine-tuned for the emergence of life. And yet that alone would not convince those who think that this fact cries out for explanation that it does not.[38]

Next, there's a view we can extract from the writings of Sylvain Bromberger, what we can call the *p-predicament* account.[39] On this view, a fact cries out for explanation to the extent that (a) we believe there is a true answer to the question *why* it obtained, but (b) any answer we can (currently) think of is ruled out by what we know.[40]

On each of these two views, however, whether something cries out for explanation is highly dependent on features of our epistemic situation. And they make it hard to see how something could cry out for explanation *tout court,* as opposed to crying our for explanation for a particular agent at a time, or relative to a particular body of background beliefs.[41]

Not all views on what it takes for something to be in need of explanation carry their sensitivity to epistemic considerations on their sleeves. For example, according to Stephen Grimm, a fact cries out for explanation just in case it is not necessary, where the relevant notion of necessity is narrower than broadly logical or even metaphysical necessity—Grimm does not quite say what the relevant notion is, other than that it is "tied to our sense of what sort of capacities a thing has, relative to the *kind* of thing it is."[42] The idea, I take it, is that objects (contingently) have

there are any 'brute necessities'. See, for example, Levine (1983); Callender (2004); White (2007). White (2005: 3) offers a plausible necessary condition on what it takes for something to cry out for explanation. Unfortunately, it does not help our discussion, for it appeals to the very notion of being in need of explanation that we are trying to illuminate.

[36] Grimm attributes this view to Hempel and Peirce. See p. 485 and references therein.

[37] Cf. Grimm (2008: 458, n6). A familiar objection due to Peter Lipton is that many things that are not surprising nonetheless cry out for explanation—the rattle of my car might cry out for explanation, even though there is nothing surprising about it at this point (see, for example, Lipton 2004: 26). The proposal in Wong and Yudell (2015) could be understood as a sophisticated attempt at reviving the surprising account—on their view, a fact cries out for explanation when we are surprised by it not 'fitting' with a background theory (p. 2884).

[38] To be sure, there is a way of reading 'surprising' on which it is still surprising that the universe we live in fine-tuned for life. But that reading smells remarkably like a metaphor for being in need of explanation.

[39] Bromberger (1962, 1971, 1988, *inter alia*).

[40] Bromberger characterizes a *p*-predicament as an all-or-nothing affair—see, for example, the Introduction to Bromberger (1992). But one could imagine generalizing his definition so one can be in a *p*-predicament to a greater or lesser extent. Presumably, the relevant sense in which we cannot think of an answer will need to be specified more explicitly (in this regard, see Bromberger 1971: 117).

[41] The point here is independent of the suggestion, most famously made in van Fraassen (1980: chapter 5), that whether something is a good explanation of a given explanandum is also highly context-dependent.

[42] Grimm (2008: 484). Cf. also p. 494: "a situation stands in need of explanation for someone in virtue of the person's sense that there are various alternative ways the subject of the situation (a system, say, or a substance that constitutes the '*A*' in a fact such as *A* is *F*) might have been."

certain capacities, which themselves determine the sorts of states they might be in.[43] And it is because *x* could have failed to be *F* that, according to Grimm, *x*'s being *F* is in need of explanation.

Whatever its merits, though, this view fails to do justice to the guiding contrast between, say, three strangers winning the three prizes at the town raffle and three siblings winning the three prizes. On Grimm's view, each of these is in need of explanation. Moreover, there isn't anything to be said for the rather plausible claim that the second one of these is more in need of explanation than the first one, even if both are in some sense in need of explanation.

There is yet another view worth mentioning, inspired by a passing remark of Cian Dorr's.[44] A fact cries out for explanation just in case it is unlikely, given that it obtains, that it is does not have a good explanation. On this view, the fact that three siblings won the three prizes cries out for explanation because it is unlikely, given that they did win, that there is no good explanation of it.

Now, on a plausible way of spelling this view out, whether something cries out for explanation will depend on some background epistemic state—a credence function, say, that could be used to measure the relevant conditional probabilities. But one could also spell it out so that the relevant conditional probabilities are perfectly objective.[45] We would still have to worry about which body of evidence is the relevant one to determine how unlikely it is—relative to those conditional probabilities and that body of evidence—that there is no good explanation of some particular fact. But setting that aside, as far as this view is concerned, there could be many concepts that correspond to elite properties which have none of the epistemic good-making features that appeal to eliteness was supposed to explain. Suppose, to illustrate, that there is some concept that correspond to a perfectly elite property. As far as this view is concerned, it could be that this concept never figures in explaining things that cry out for explanation. After all, facts about what cries out for explanation are not determined by facts about what properties are elite. It would thus be hard to see how the claim that a concept corresponds to a perfectly elite property could be what alone explains why it has the epistemic merits that it has.

[43] Presumably, Grimm is not thinking of those natures and capacities as being essential to a given thing. Otherwise, the fact that it is metaphysically possible for *A* to be *F* would imply that it is possible for *A* to be *F even if* we hold fixed *A*'s nature and capacities, so that the corresponding notion of necessity would be as strong as metaphysical necessity, contrary to what Grimm suggests.

[44] The view is discussed in a footnote, modified by a 'roughly', and not obviously wide enough in scope to apply to anything beyond whole theories. See Dorr (2010: 154, n29): "Roughly, for a claim to cry out for explanation is for it to be unlikely *conditional on its being true* that its truth is not explained by that of any better theory." There are multiple ways of making sense of this claim. On one, what cries out for explanation is a complete theory: *T* cries out for explanation just in case it is unlikely, given *T*, that *T* is not explained by some other theory T' that is better (in the relevant sense) than *T*. The downside of this understanding is that the scope is limited to whole theories. Another interpretation is this: a claim *C* cries out for explanation just in case it is unlikely, given *C*, that *C* is not explained by some theory T' that is better (in the relevant sense) than *C*. The downside of this way of thinking about it is that it requires that we find a suitable way of comparing particular claims about some event (say) with complete theories. Perhaps the view is that T' is better than *C* in the relevant sense if it is a better explanation of *C* than *C* itself. (Though this would make it hard to see how there could be any claim that does not cry out for explanation, save perhaps for 'self-explanatory' claims, whatever those turn out to be.) Or perhaps the idea is that T' is better (in the relevant sense) if T' is better than some other salient, available putative explanation of *C*. I cannot tell.

[45] As in, for example, Williamson (2000: chapter 10).

Incidentally, it is worth highlighting a general upshot of this last observation. The objection to SIMPLE is *not* that it would imply, in contrast to all plausible theories of what it takes to be in need of explanation, that it is an objective matter whether something cries out for explanation. The objection, rather, is that SIMPLE would imply that whether something is in need of explanation has nothing to do with *epistemic* considerations. It could be a perfectly objective matter whether something is in need of explanation relative to a given epistemic situation. So unless we have some sense as to how being in need of explanation could have nothing to do with epistemic considerations, we have some reason to think that SIMPLE cannot be right.

3.3 Trouble with Uniformity

A theory of conceptual evaluation should, if possible, be sufficiently general in scope. We want an answer to the question, what makes for good concepts, epistemically, that is relatively discipline neutral. To illustrate: an answer that applies uniformly to both mathematical concepts and concepts from the natural sciences is preferable, all else equal, than one that does not.

How does SIMPLE fare with respect to this desideratum?

On the one hand, SIMPLE generalizes quite nicely to different domains of inquiry. As long as it makes sense to talk of elite properties in a given domain, we can use SIMPLE to answer the question what makes for good concepts in that domain. And while not all proponents of the elite/non-elite distinction might be willing to countenance that such a distinction carries over to mathematical properties, for example,[46] I will simply assume that the proponent of SIMPLE is not among them.

On the other hand, digging a little deeper reveals that things are not as straightforward as they appear.[47] According to SIMPLE, what explains why the concept *electron* is a good concept is that the property of being an electron is suitably elite. And what explains why the concept *determined game* is a good concept is that the property of being a determined game is also elite.[48] Ask now: how are these two explanations supposed to go?

[46] If elite properties constitute a minimal supervenience base for all of reality, then there is no hope for a non-trivial theory of elite mathematical properties. And if elite properties are qualitative properties, there is even less hope for a non-trivial theory of elite mathematical properties. (Cf. Lewis 1991: section 2.6.) None of this need imply that one cannot give a theory of naturalness for mathematical properties: it's just that whatever that theory turns out to be it will have a rather different look than any of the familiar theories of eliteness. For relevant discussion, see Bricker (n.d.).

[47] I'm indebted to Tappenden (2008a) for raising the question whether there is a plausible theory of the natural/non-natural distinction that encompasses both mathematical and non-mathematical properties. Cf. also the discussion in Tappenden (2008b) of an argument in Sider (1996) for the claim that preference for one over another way of defining ordered pairs in set theory can only be based on merely pragmatic considerations.

[48] A (two player, perfect information) game which does not allow for draws is *determined* iff one of the players has a winning strategy: a strategy that will allow her to win no matter what the other player does. Alternatively, if we allow for games where players can draw, we say that a game is determined if one of the players has a strategy that guarantees she will not lose. A fundamental theorem attributed to Zermelo (but see Schwalbe and Walker 2001 for a more accurate account of the history behind the role of Zermelo's work in early game theory) is that every finite game—a game that ends after a finite number of moves—is determined. Not all infinite games are determined. Characterizing the class of infinite games that are determined has proven to be a surprisingly fruitful research program. More on this below.

Consider one particular theoretical benefit we might expect to accrue from epistemically good concepts: fruitfulness. A concept is fruitful, as I will understand the term, when it makes unexpected appearances in dealing with a variety of questions.[49] One particularly important way a concept can be fruitful is in lending itself to successful prediction: one can typically predict features of an unexamined object falling under that concept on the basis of features of other, examined objects falling under that same concept. So the question is: why would the fact that a concept corresponds to an elite property explain that concept's contributing to successful prediction?

If we zoom in on a particular class of arguably elite properties—the so-called natural kinds—we can get some sense of how to answer that question. Kinds, one might think,[50] correspond to property clusters that are sustained by certain causal mechanisms. It is the fact that certain causal mechanisms sustain the co-instantiation of certain features that explains why concepts corresponding to natural kinds are projectable.[51]

The idea, very roughly, is that properties come in clusters: for example, thermal and electrical conductivity, ductility, malleability, and having a shiny appearance. And this clustering is due to the presence of underlying mechanisms (something having to do with the presence of free electrons, in our example) which ensure that those properties tend to be coinstantiated.[52] As a result, the property *being metal* is a property that supports inductive generalizations: from the fact that the temperature of a given metal is inversely proportional to its conductivity, one is warranted in inferring that the temperature of other metals is too. This at least gives us a sketch of a story as to how the metaphysics of the property corresponding to our concept *metal* can account for the role that this concept plays in our epistemic practices.

Now, much like the concept *being metal,* the concept *determined game* is a fruitful concept. To see what I mean, let me take a moment to introduce some definitions. Any given set of real numbers can be identified with a unique set of sequences of natural numbers. For any set *A* of sequences of natural numbers, we can define the

[49] Note here that at least some of the problems raised by Nolan (1999) against taking the fruitfulness of a *theory* to be an epistemic virtue do not arise for thinking of fruitfulness as an epistemically good-making feature of concepts. As Nolan points out, if the fruitfulness of a theory is, as the term might suggest, a matter of it giving rise to new problems or opening up new lines of inquiry, it is hard to see how fruitfulness (or 'fertility', to use Nolan's term) would be an epistemic virtue. To think so would be like thinking that " 'Faces many problems' or 'Could do better' or 'Much room for improvement' are high praise on the report card of a theory" (p. 267). Similar concerns do not carry over to thinking of fruitfulness of *concepts* as an epistemically good thing.

[50] See, for example, Boyd (1988); Kornblith 1993); Millikan (1999).

[51] It is telling that such homeostatic property clusters can in principle be sustained by the intentional activities of human beings. See, for example, Mallon (2003). Presumably, such properties would not count as elite on any reasonable theory of eliteness. Otherwise, whether a property is elite would depend on contingent facts about human interests at a given point in time. (In this respect, the discussion of the concept of the GDP in Coyle (2014), and the way in which changes in the economy make it a less than useful concept can serve as an illustration of this point.) That suggests that the explanation for why concepts corresponding to these properties turn out to be epistemically valuable will have little to do with the eliteness of the relevant properties. But this is a line of argument I do not intend to pursue here. (Thanks to Hilary Kornblith for discussion on this point.)

[52] Cf. Boyd (1999: 82f).

two-player game G_A as follows: players *I* and *II* alternate picking a natural number; player *I* wins iff the resulting sequence is an element of A—else, player *II* wins. We say that a set of real numbers A is determined iff G_A is determined.

It is well-known that all open sets (countable unions of disjoint open intervals) and all Borel sets (sets obtained from open sets by closing under countable unions, countable intersections, and relative complements) are determined. The *Axioms of Definable Determinacy* state that all 'definable' sets of reals are determined—I write 'axioms' because there are different ways of making the notion of definability precise, each of which corresponds to a particular axiom of definable determinacy.[53] Axioms of Definable Determinacy have been a driving force behind much work in set theory over the last thirty years.[54] The study of determinacy did not just result in a number of general results about how well-behaved 'reasonably definable' sets of real numbers are—for instance, it is known that all determined sets of reals are Lebesgue measurable, have the Baire property, and contain a perfect subset if uncountable—but also served to establish deep connections between two seemingly unrelated branches of set theory: the study of large cardinals and descriptive set theory.[55]

Indeed, it would not be implausible to talk of the notion of determinacy as being projectible: it seems reasonable to conclude, from the assumption that all sets of reals we have been able to define are determined, that all definable sets of reals are determined, even though strong versions of definable determinacy are independent of the standard axioms of set theory.[56] Moreover, axioms of definable determinacy have been crucial in proving results in a range of different areas which later turned out to be provable in much weaker theories.[57] Thus, the notion of determinacy seems to have as good a claim at being an epistemically good concept as any natural kind term.

Of course, we cannot explain the fruitfulness of the notion of determinacy in the same way we explained the fruitfulness of the species concept. It would be quite a stretch to posit something analogous to a causal mechanism which sustains (in a non-causal way) the clustering of so-called regularity properties (Lebesgue measurability, etc.) around determined sets of reals. So a very different type of explanation will be needed to account for the fact that concepts corresponding to elite mathematical properties turn out to be fruitful. Such a story might well be forthcoming—perhaps mathematical properties have their own, distinctive way of being elite, one which explains the fruitfulness of the corresponding concepts. But as far as I can tell, no such story has ever been told. (The view that for a mathematical property to be

[53] Perhaps the best-known of these axioms states that all sets of reals in $L(\mathbb{R})$—the smallest transitive model of set theory that contains all reals and all the ordinals—are determined.

[54] For an overview of some of the main results, see Koellner (2014); Welch (2015); as well as Maddy (2011: chapter II). For some of the mathematical details, see, for example, Kanamori (2009: chapter 6).

[55] The canonical reference here is Shelah and Woodin (1990). See Larson (2012); Koellner (2014) for a more comprehensive list of references.

[56] This is not an isolated example. See, for example, the discussion in Tappenden (2008b) of the quadratic reciprocity theorem and the role that 'projecting' on the predicate 'is a quadratic residue'—and relatedly, the introduction of the Legendre symbol—played in the discovery of the theorem. Unlike in the case of axioms of determinacy, which have been accepted on 'mere' non-deductive grounds, the quadratic reciprocity theorem admits of a direct proof. See also the discussion of the concept of field, and the way it contrasts with Frege's concept of a 'quantitative domain', in Tappenden (2005: 20f).

[57] For further discussion, see Martin (1998).

elite is in part for the corresponding concepts to be fruitful would certainly not do.) And without such a story, a proponent of SIMPLE will be unable to give a unified answer to the question, what makes for good concepts, epistemically?

(Note that what I say here is perfectly compatible with the claim that fruitfulness of mathematical concepts is evidence that they correspond to elite mathematical properties.[58] What I'm calling into question is the existence of a suitable notion of eliteness for mathematical properties that might explain why concepts corresponding to such properties are fruitful.[59])

3.4 Concept and Property

I have been presupposing, for the sake of compliance with the presuppositions of SIMPLE, that to each concept corresponds a unique property. It is only given this assumption that SIMPLE can be formulated: if a concept corresponds to multiple properties, then we are not guaranteed that concepts can be compared in terms of the eliteness of *the* corresponding properties.

Now, for reasons all too familiar, this presupposition is not quite right—at least given plausible assumptions about how to individuate concepts. For instance, consider the concept *local.* Intuitively, there is no one property—the property of *being local* as it were—that corresponds to that concept. Rather, in different contexts, different properties are predicated of an object that is judged to fall under that concept.[60]

A proponent of SIMPLE could now restrict her view so that it only applies to concepts that do meet this presupposition. So let us banish 'local' from among the candidates to good concepts—banish, too, concepts like *leftmost, tall,* and *currently fashionable.* In principle, there is nothing objectionable to this strategy. The worry is that banishing context-sensitive concepts—concepts that correspond to different properties in different contexts—will leave us with little to apply SIMPLE to. In particular, the concern is that by banishing context-sensitive concepts we may end up banishing the very concepts we want a theory of conceptual evaluation to apply to.

What motivates this idea is the observation that many of our seemingly non-context-sensitive terms appear upon reflection to be context-sensitive after all.[61] Take a familiar example, from (Chomsky 1976):[62]

[58] Cf. Maddy (2011), whose 'thin realism' is based on the idea that the success of the set-theoretic methods is evidence that those methods are tracking the 'underlying contours of mathematical depth' (p. 82).

[59] For a nice summary of different attempts at cashing out a notion of mathematical depth—as well as their limitations—see Arana (2015). Arana's discussion is focused on a notion of depth as applied to mathematical theorems. But many of the issues he raises could be transposed into a discussion of depth as applied to mathematical properties.

[60] Even proponents of so-called semantic minimalism would grant that there is no such thing as the property of being local *simpliciter.* See Cappelen and Lepore (2005: 1). (Admittedly, Cappelen and Lepore express some skepticism that we should treat so-called contextuals as genuinely context-sensitive expressions—see fn. 1, p. 1.).

[61] The relevance of this brief foray into the debate over the extent of context-sensitivity in natural languages to our discussion about concepts should become clearer shortly.

[62] See also Chomsky (1995, 2000), and many other places. For illuminating discussion of these and related observations, see Pietroski (2003); Collins (2009).

(3) a. John wrote a book.
b. The book weighs five pounds.

As Chomsky points out (p. 48), the term *book* behaves differently in the two sentences in (3). In (3a), it must pick out an abstract entity: the book that John wrote can survive the destruction of the original file it was stored in. In (3b), however, it must pick out a concrete entity, capable of having mass. Interestingly, what is going on in (3) cannot obviously be explained by claiming that the two occurrences of book correspond to different lexical items. After all,

(4) John wrote a book that weighs five pounds.

is perfectly acceptable, in contrast with (e.g.)

(5) Jumbo waved his trunk, which was full of clothes.

which is to

(6) a. Jumbo waved his trunk.
b. The trunk was full of clothes.

like (4) is to (3).[63]

And *book* is not unique in this regard. Consider (from Chomsky 2000: 37):

(7) a. London is unhappy.
b. London is ugly.
c. London is polluted.
d. London is so unhappy, ugly, and polluted that it should be destroyed and rebuilt 100 miles away

Again, it seems as if *London* denotes slightly different things in (7a)–(7c), and yet the acceptability of (7d) suggests that we are not dealing with multiple lexical items here. Also consider (cf. Chomsky 1970: n. 7 and 1976: 49f):

(8) a. The temperature is rising.
b. The temperature is 70°.
c. The temperature, which was 70°, is rising.

(9) a. Mary wrote the proof in pen and paper.
b. The proof will have a lasting impact on the history of mathematics.
c. Mary wrote a proof, in pen and paper, that will have a lasting impact on the history of mathematics.

To be sure the arguments here are not decisive.[64] But they strongly suggests that terms like *book, proof,* and *temperature,* display a certain amount of unexpected

[63] Example (5) is example (19ii) in Chomsky (1976).

[64] See for example the discussion of Partee's so-called temperature paradox in Montague (1973), as well as the more recent discussion in Lasersohn (2018).

context-sensitivity, so that the property corresponding to each such predicate varies from context to context.[65]

All that said, these examples might appear somewhat irrelevant to the more interesting questions surrounding conceptual evaluation from an epistemic perspective. After all, they all concern what we could call 'folk', or 'common-sense' concepts. And it is tempting to think that part of what scientific progress consists in is the abandonment of such concepts. Once we narrow our attention to the conceptual tools actually used in mature sciences, we will no longer be faced with the kind of context-sensitivity that might otherwise get in the way of something like SIMPLE.[66]

Or so one might think. But unfortunately for the proponent of SIMPLE, it looks like many of the concepts that play a significant role in (non-fundamental) sciences can't quite be assigned to a single property. The phenomenon—which, following Wilson (2006) we can call the 'multi-valuedness' of the relevant concepts—is best understood by looking at some examples.

Consider the concept *hardness*.[67] What is the property corresponding to it? Of course, one *could* say that the property corresponding to the concept *hardness* is, well, the property of hardness. But in trying to find anything illuminating to say about what that property is—something that we might undoubtedly need to do in order to figure out its position on the eliteness scale—we quickly find ourselves at a loss.

There are different more or less familiar tests for hardness—scratching tests, indentation tests, rapping tests, among others.[68] Which test is deemed appropriate depends on a variety of factors, including the type of material being tested for hardness. As Wilson writes (2006: 336):

> [I]n everyday contexts we adjudicate the "hardnesses" of various materials, both comparatively and absolutely, through a wide variety of comparatively easy to apply tests—we might squeeze the material or indent it with a hammer; attempt to scratch it or rap upon it; and so on. In most cases, we will be scarcely aware of the exact technique we will have employed for this appraisal: "Did I rap, squeeze or scratch that piece of wood? I can't really remember." In fact, our choice of tests is likely to have been suggested by the material in question: we instinctively appraise a

[65] It is of course a non-trivial question to explain why these terms get to behave the way they seem to in sentences like (9c), and I have nothing to offer by way of explanation here.

[66] Indeed, Chomsky himself—insisting as he does that developing 'referential' semantics is the wrong way to go about studying meaning—seems to leave it open whether terms introduced in the natural sciences are well-behaved enough so that each term or concept corresponds to a unique property (see Chomsky 2000: 42f). On Chomsky's view, part of what prevents terms like *London* or *house* from being part of a 'mature' science (which Chomsky takes to mean that clauses like '*London* refers to London' cannot be part of any serious scientific study of meaning) is that they are "used to refer to concrete objects, but from the standpoint of special human interests and goals and with curious properties". And the evidence for this is that, for example, "[a] house can be destroyed and rebuilt, like a city; London could be completely destroyed and rebuilt up the Thames in 1,000 years and still be London, under some circumstances" (p. 21). No term belonging to mature 'naturalistic inquiry' should thus display the characteristics of terms like *London* or *house.* If the relevant characteristics include the seeming context-sensitivity of such terms, Chomsky must then think that highly context-sensitive items should be banished from good scientific vocabularies. Cf. Stoljar (2015).

[67] As should become clear, my discussion here is deeply indebted to Wilson (1982 as well as 2006).

[68] Wilson (2006: chapter 6, ix).

wood by rapping upon it, a rubber by squeezing, a metal by attempting to make a small imprint; a glass or ceramic by rapping lightly or scratching (not by trying to make a small imprint!).

But the type of material alone does not determine which particular test is appropriate. A hunk of steel, for example, may sometimes be tested using an indentation test—measuring the diameter of an indentation left behind by a steel ball that is placed on its surface for a fixed amount of time, as in the familiar Brinell tests—but also using an abrasion test. And the two tests can yield conflicting verdicts as to the comparative hardness of different types of steel.[69] It is partly because of this that one finds in the literature on hardness tests claims to the effect that *hardness* is not a 'fundamental' property, or that there may be no such thing as the property of hardness.[70]

At this point, it might be tempting to conclude that the concept *hardness* is somehow defective. And indeed, there may be some sense in which it is. But it would be unreasonable to conclude from this that there is nothing to be said for it, epistemically, given the way it is embedded in a network involving other macroscopic concepts like *solid, rigid body,* and *force* which seem essential to our understanding of the macro world.[71]

Can the proponent of SIMPLE explain what makes *hardness* an epistemically good concept? One possibility would be to insist that there really are multiple concepts in place: *Brinell hardness, scratch hardness,* and so on. The mistake, according to this line of thinking, is to think that there is a single concept whose epistemic credentials we need to account for. Rather, there are multiple concepts, each one of which is epistemically good in its own right. It is merely an accident of the way English developed that we have homonyms corresponding to different concepts.

There are at least two problems with this strategy, however. First, and perhaps least significantly in this context, this strategy would fail to do justice to the way that the

[69] Cf. Wilson (2006: 338). Of course, in discussing these issues authors sometimes appeal to qualifiers to avoid confusion. Thus, Walley (2012: 1033) disambiguates and talks about "the relation between Brinell hardness and resistance to abrasion" not being as straightforward as one might have expected. But sometimes, authors expect readers to disambiguate according to context, as in, for example, Malzbender et al. (2002: 52): "hardness is not a fundamental property of materials. Hardness is related to material properties, in particular to the yield strength and the elastic modulus, but this relationship depends on the indenter geometry. We should, therefore, be careful when comparing hardness values from different sources." Additional examples of this way of exploiting the ambiguity of 'hardness' are easy to find in the literature.

[70] For example, Newey and Weaver (2013: 13): "A hard material is difficult to scratch, wear away by abrasion or to indent. Hardness is not a fundamental property of a material: for each method of measuring it, it is some combination of elastic, plastic, and (in some cases) fracture properties. Hardness can be measured only by comparison with a material used as a scratcher or indenter and has objective meaning only in terms of a specific type of test. For example, glass will scratch steel but fractures more readily under indentation; nylon has a high resistance to wear but not to indentation." And Cahn and Lifshin (1993: 183): "Although there is probably no such property as 'hardness', it is a convenient generic test." See also the quotes from Samuel Williams in Wilson (2006: 350).

[71] Cf. the discussion of solidity in Wilson (2006: 352–3). As Wilson points out, the notion of solidity that is usefully applied to macroscopic objects—solid objects are those that "preserve their volume under all interactions, as long as they remain integral" (p. 354)—is not the one we would rely on when discussing the microscopic interactions between objects, for it is precisely by "not acting like a solid at the molecular level" that a piece of steel does not crumble into dust under mild applied stress.

term 'hardness' is actually used. The apparent ambiguity or polysemy of 'hardness' does not seem to pattern with lexically ambiguous terms like 'bank'. It would be at best an unimaginative joke to say: "there are two different kinds of banks: river banks and financial institutions". But talk of different kinds of hardness is not out place in the literature on materials science,[72] as is the claim that hardness is the capacity to resist penetration or abrasion, when there is no property that accounts both for resistance to penetration and resistance to abrasion.[73]

The second problem is that there seems to be no explanation, on this way of thinking, for why these allegedly different concepts are 'linked' together in the way they are. It certainly seems as if part of what lets concepts like *hardness* play the role they do in understanding macro phenomena is that they adjust themselves almost unnoticeably so as to pick out different properties in different contexts, or, if we prefer, so as to take each other's place (the different concepts) in different contexts. And this seems to be something over and above the putative eliteness of the properties corresponding to each of the ways in which *hardness* is used.

To see what I have in mind, consider a different example, one where the 'switch' from property to property is induced not by focusing on a different material but by 'shifting scales'.[74] When considering a three-dimensional object at the macro-level, we think of its surface as consisting of a two-dimensional 'skin' that surrounds it. Now, there are well-known puzzles that arise from applying these topological notions to macroscopic objects.[75] But there's no denying the immense fruitfulness of doing so. This is not just in the sense that our folk-theoretic understanding of macroscopic objects in some sense presupposes that surfaces are aptly thought of in topological terms. These topological notions are essential to continuum mechanics, for example, even if doing so requires a certain amount of theoretical acrobatics.[76]

Consider now what it takes for something to be a surface when looked at from a 'micro' point of view. There, it makes sense to think of surfaces as lattice structures

[72] Cf. the quoted passage in fn. 69.

[73] Cf. also Wilson (1982: 575f): "We can meaningfully ascribe an extension (with certain implicit parameters) to the everyday use of 'is 90°F' which differs from that appropriate to the temperature experts in our society. There is a natural temptation to claim that the experts assign a somewhat different *meaning* (or sense) to 'is 90°F' than the rest of us. The fact that 'is 90°F' corresponds to two fairly clearcut *ranges of application* represents an interesting aspect of English linguistics, but this common phenomenon should be regarded as *sui generis* and not lumped with ambiguities of 'meaning' such as the word 'bank' displays. Intuitive talk of the 'varying senses' of 'is 90°F' is unexceptionable if understood merely as a description of the double range of application, but it should not mislead us into supposing that the experts 'grasp' a distinct concept (especially when physicists employ the term in the same way as the rest of us in ordinary conversation with no sense of disharmony with their laboratory practice)."

[74] What follows owes much to Wilson (2006); Batterman (2013); Bursten (2018).

[75] For a nice summary of some of the issues, see Arntzenius (2008); Varzi (2015).

[76] See, for example, the discussion of how spray can form on the the surface of water in Wilson (2006: 210). As Wilson points out, if we start with a smooth water surface and play out the differential equations governing the flow of wind and water, we will never get the change in topology required in order for a bit of the surface to break off into a water droplet. As a result, some acrobatics are necessary in order to model the generation of the spray. The way Wilson describes it, "[w]hen a change in the fluid's topology looks imminent, practitioners begin investigating two fluid configurations that run in parallel, one containing the still attached drop and the other describing a drop of similar shape detached from its ocean."

made up of atoms 'connected' to one another by electromagnetic forces. It is by looking at the details of these structures that we can explain, say, why a given metal spoon will bend more easily than a seemingly identical one made of the same material. It is because of slight imperfections on the surface of the spoon—so-called *dislocations,* which are essentially irregularities in the crystalline structure of the material—that a steel spoon can be bent, the quantity and distribution of dislocations determining how easy it is for it to bend. And it is because of the presence of 'gaps' between the atoms that make up the surface of a metal beam that oxidation occurs[77]—which in turn explains why so-called cold-welding does not happen in everyday situations.[78]

Our concept *surface,* then, seems to pick out different properties depending on something like a background choice of scale.[79] A proponent of SIMPLE might thus be inclined to think that there are different concepts here, one for each choice of scale, and offer independent explanations for why each of them is epistemically good.

Now, on the face of it, there is something appealing about the view on which we have different *surface* concepts—a 'continuum' surface concept and an 'atomistic' surface concept. Yet there are many ways in which these concepts are related: for instance, changes in the microstructure of the surface of the object will in turn affect how the surface should be modeled in continuum mechanics.

Suppose, for instance, we are interested in explaining the way in which hammering a piece of metal might affect its macroscopic properties (through so-called work-hardening, one can strengthen a hunk of steel by making it less ductile). We may need to start by modeling its surface at the microlevel (perhaps because a blow resulted on some cleavage cracks along the surface, introducing dislocations) and model the rest of the steel using a continuum model. But as the hammering of the steel goes on, and the surface dislocations start moving down to the bulk of the steel, our model of the surface may need to switch: extending the atomistic model along the path that the dislocation takes will quickly become prohibitive.

[77] See Cadavid and Cabot (2017) for an accessible discussion of this point.

[78] Cf. Wilson (2008: 285f): "The outer molecular layers of an iron bar (which are anti-averaged into the usual two-dimensional boundary conditions of elasticity theory) actually comprise an extremely complex region when viewed up close."

I was alerted to the possibility of cold-welding in the absence of 'greasy atmospheric crud' by Wilson (2006: 373). A particularly dramatic illustration of what can happen to metal once oxides are removed is the well-known anomaly in the deployment of the antenna of the Galileo spacecraft: two metal parts were essentially 'welded' to one another, preventing the antenna from fully deploying. See, for example, Miyoshi (1999) for some of the details. Further discussion of the more general issues can be found in the works cited in fn. 92 of Wilson (2006, *loc. cit.*).

[79] Another example of this type of scale-dependence, also from Mark Wilson, is the concept *force.* As Wilson puts it (Wilson 2013: 54): "The term *force* has a notorious tendency to alter its exact significance as characteristic scale lengths are adjusted. At a macroscopic level, the 'rolling friction' that slows a ball upon a rigid track is a simple Newton-style force opposing the onward motion. But at a lower scale length, the seemingly 'rigid' tracks are not so firm after all: they elongate under the weight of the sphere to a nontrivial degree. So part of the work required to move our ball against friction consists in the fact that it must *travel further* than is apparent. But when we consider the 'forces' on our ball at a macrolevel, we instinctively treat the track length as fixed and allocate the effects of its actual elongation to a portion of the 'force of rolling friction' budget."

Or suppose we are trying to understand the way in which enough stress on a hunk of metal can lead to fracture.[80] If we rely simply on a continuum mechanical description of the phenomena, we will be unable to predict the emergence of cracks along the surface of our metal—cracks which will grow and result in fracture. This is because introducing cracks requires a change in the topology of our surface, which cannot be predicted by the continuous deformation allowed in a continuum mechanical setting. So in order to account for the emergence of cracks, we need to switch to modeling the surface at a microscale. Unfortunately, we cannot just model the phenomenon of interest entirely at the atomic level, for the number of atoms along our surface is computationally intractable. For these purposes, physicists have developed *multiscale* models, where different models of a surface (say) are 'linked' together, so that part of the surface where cracks emerge and grow are modeled at the atomic level and the rest of the surface—along and beyond which the applied stress will flow in a somewhat predictable manner—using a continuum model.[81]

And these relations are not clearly explicable in terms of law-like relations among the corresponding properties (for one, the property corresponding to the 'continuum' surface concept is one that no concrete object has in this world, so there could not be any law-like connection between instantiations of the property corresponding to the 'atomistic' *surface* concept and instantiations of the property corresponding to the 'continuum' one). Moreover, thinking of the difference between the two *surface* concepts as a simple case of lexical ambiguity would fail to do justice to the fact that both continuum and atomistic models of the effects of cleavage on a surface are models of the same phenomenon. We would be mischaracterizing multiscale modeling of materials if we thought of them as concerned with different phenomena at the same time, rather than as concerned with the same phenomenon at different scales.

Granted, the proponent of SIMPLE could instead insist that there is a single operative concept here, viz. the 'atomistic' *surface* concept. Talk of surfaces as continuous regions is, on this view, strictly speaking false. Still, the fact remains that the 'continuum' *surface* concept (assuming, for the sake of argument, that it is a different concept) plays a role in explanations which, according to EXPLANATION, should be deemed an epistemically good concept. Exactly what explains that it is an epistemically good concept is a difficult question. But it is hard to see how the answer will be compatible with SIMPLE.

[80] Perhaps a more pressing explanandum is the fact that the force required to slide a block over a flat surface is proportional to the block's weight but independent of the apparent area of contact—the so-called Amonton Law of Friction. The orthodox explanation appeals to the fact that, because of surface asperities, the true area of contact is much smaller than the apparent area of contact and it increases by way of deformation of these asperities in proportion to the applied compressive force. But alternative multiscale explanations—that do not appeal to surface roughness—are available, appealing instead, for instance, to the presence of self-healing cracks on the surface of the materials in proportion to the ratio of horizontal to compressive stress. See Gerde and Marder (2001) for discussion. See also Krim (1996) (and references therein) for a survey of the limitations of explanations appealing to surface roughness.

[81] See, for example, Zhang, Johnston, and Chattopadhyay (2014: 120): "The phenomenon of fatigue involves multiple length scales including crack nucleation in the microscale, coalescence of microcracks in the mesoscale and major crack propagation in the macroscale. Therefore, it is critical to develop a lengthscale-dependent and physics-based model to understand material performance and ultimately assess the survivability of complex structural systems." See also Rudd and Broughton (2000: esp. section 1).

4 Where to Next?

I wish I had an alternative story to offer, here, as to what makes for epistemically good concepts. Unfortunately, I do not. But I think there are a number of lessons to be drawn from our discussion so far, lessons that any such alternative story must take into account. In closing, I want to briefly go over what I take those lessons to be.

First, it may turn out to be impossible to give an informative answer to the question what makes for epistemically good concepts without making more substantive assumptions about what concepts are. The role that concepts—or meanings, if you would rather avoid talk of concepts altogether—play in our ability to make sense of the world around us cannot be fully accounted for in terms of a list of properties that the relevant concepts correspond to. We thus need a better grip on what concepts are—other than whatever mediates our representation of properties—in order to understand what makes some concepts better suited to our epistemic projects than others.

Second, we should be open to the possibility that no purely 'metaphysical' answer to our question will be forthcoming. Nothing in what I've said here rules out the possibility that some story that is not sensitive to features of our epistemic situation—one that goes beyond the relative eliteness of the properties corresponding to concepts—might be the best answer to the question what makes for epistemically good concepts. (Perhaps there are some relations among properties such that concepts are epistemically good to the extent that they are associated with a family of suitably related, sufficiently elite properties.) But unless it is roughly as simple as SIMPLE, we have no reason to favor an answer that is independent of our epistemic circumstances to one that is not—the one thing SIMPLE had going for it, which might have led to our thinking that something like it had better be right, was its simplicity. Any story about what makes for epistemically good concepts will, I suspect, turn out to be rather complex. Whether that complexity will come from the metaphysics of properties or from facts about the way in which creatures with our cognitive capacities interact with our rather complex environment is an open question.

Acknowledgements

For helpful comments, discussion, and advice, thanks to Phil Bricker, Maya Eddon, Nina Emery, Sophie Horowitz, Joe Levine, Chris Meacham, Katia Vavova, and participants at the 2016 workshop on Conceptual Engineering at the CSMN, University of Oslo. Special thanks to Alexi Burgess and Hilary Kornblith for extremely helpful comments on an earlier draft of this chapter.

References

Antony, Louise. 2003. Who's Afraid of Disjunctive Properties? *Philosophical Issues* 13:1–21.

Arana, Andrew. 2015. On the Depth of Szemerédi's Theorem. *Philosophia Mathematica* 23 (2):163–76.

Armstrong, David M. 1978. *Universals and Scientific Realism*, vol. 1. Cambridge: Cambridge University Press.

Arntzenius, Frank. 2008. Gunk, Topology and Measure. In Dean Zimmerman (ed.), *Oxford Studies in Metaphysics,* vol. 4 (pp. 225–47). Oxford: Oxford University Press.

Batterman, Robert W. 2013. The Tyranny of Scales. In Robert W. Batterman (ed.), *Oxford Handbook of the Philosophy of Physics* (pp. 255–86). Oxford: Oxford University Press.

Berker, Selim. 2018. The Unity of Grounding. *Mind* 127 (507):729–77.

Blackburn, Simon. 1998. *Ruling Passions.* Oxford: Oxford University Press.

Boyd, Richard. 1988. How to Be a Moral Realist. In Geoffrey Sayre-McCord (ed.), *Essays on Moral Realism* (pp. 181–228). Ithaca: Cornell University Press.

Boyd, Richard. 1999. Kinds, Complexity and Multiple Realization. *Philosophical Studies* 95 (1):67–98.

Bricker, Philip. 1992. Realism without Parochialism. Unpublished manuscript, University of Massachusetts, Amherst. Symposium paper at the Pacific APA meeting, 1992.

Brigandt, Ingo. 2003. Homology in Comparative, Molecular, and Evolutionary Developmental Biology: The Radiation of a Concept. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 299B (1):9–17.

Brigandt, Ingo. 2010. The Epistemic Goal of a Concept: Accounting for the Rationality of Semantic Change and Variation. *Synthese* 117 (1):19–40.

Bromberger, Sylvain. 1962. An Approach to Explanation. In R. J. Butler (ed.), *Analytical Philosophy,* vol. 2 (pp. 72–105). Oxford: Oxford University Press. Reprinted in Bromberger (1992: 18–51). Page numbers refer to the reprinted version.

Bromberger, Sylvain. 1966. Why-Questions. In Robert G. Colodny (ed.), Mind and Cosmos: Essays in Contemporary Science and Philosophy, vol. 3 (pp. 86–111). University of Pittsburgh Series in the Philosophy of Science. Pittsburgh, PA: University of Pittsburgh Press. Reprinted in Bromberger 1992: pp. 101–11. Page numbers refer to the reprinted version.

Bromberger, Sylvain. 1971. Science and the Forms of Ignorance. In Maurice Mandelbaum (ed.), *Observation and Theory in Science*. Baltimore: The Johns Hopkins Press. Reprinted in Bromberger (1992: 112–27). Page numbers refer to the reprinted version.

Bromberger, Sylvain. 1988. Rational Ignorance. *Synthese* 74 (1):47–64. Reprinted in Bromberger 1992: pp. 128–144. Page numbers refer to the reprinted version.

Bromberger, Sylvain. 1992. *On What We Know We Don't Know.* Chicago and Stanford: The University of Chicago Press and CSLI.

Burgess, Alexis, and Plunkett, David. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, Alexis, and Plunkett, David. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Burgess, John P. 2005. Being Explained Away. *Harvard Review of Philosophy* 13 (2):41–56.

Bursten, Julia R. 2018. Smaller Than a Breadbox: Scale and Natural Kinds. *British Journal for the Philosophy of Science* 69 (1):1–23.

Cadavid, Doris, and Cabot, Andreu. 2017. Oxidation at the Atomic Scale. *Science* 356 (6335):245.

Cahn, Robert W., and Lifshin, Eric. (eds.) 1993. *Concise Encyclopedia of Materials Characterization.* Oxford: Pergamon Press.

Callender, Craig. 2004. Measures, Explanations and the Past: Should 'Special' Initial Conditions be Explained? *British Journal for the Philosophy of Science* 55 (2):195–217.

Cappelen, Herman, and Lepore, Ernie. 2005. *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism.* Oxford: Blackwell.

Chomsky, Noam. 1970. Remarks on Nominalization. In Roderick A. Jacobs and Peter S. Rosenbaum (eds.), *Readings in English Transformational Grammar* (pp. 184–221). Waltham: Ginn & Co.

Chomsky, Noam. 1976. On the Nature of Language. *Annals of the New York Academy of Sciences* 280 (1):46–57.

Chomsky, Noam. 1995. Language and Nature. *Mind* 104 (413):1–61.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind.* Cambridge: Cambridge University Press.

Collins, John. 2009. Methodology, Not Metaphysics: Against Semantic Externalism. *Proceedings of the Aristotelian Society Supplementary Volume* 83 (1):53–69.

Coyle, Diane. 2014. *GDP: A Brief but Affectionate History.* Princeton: Princeton University Press.

Dasgupta, Shamik. 2009. Individuals: An Essay in Revisionary Metaphysics. *Philosophical Studies* 145 (1):35–67.

Dasgupta, Shamik. 2018. Realism and the Absence of Value. *Philosophical Review* 127 (3):279–322.

Dorr, Cian. 2010. Of Numbers and Electrons. *Proceedings of the Aristotelian Society* 90 (2):133–81.

Dorr, Cian, and Hawthorne, John. 2013. Naturalness. In Karen Bennett and Dean W. Zimmerman (eds.), *Oxford Studies in Metaphysics*, vol. 8 (pp. 3–77) Oxford: Oxford University Press.

Eddon, Maya. 2013. Fundamental Properties of Fundamental Properties. In Karen Bennett and Dean W. Zimmerman (eds.), *Oxford Studies in Metaphysics*, vol. 8 (pp. 78–104). Oxford: Oxford University Press.

Enoch, David. 2011. *Taking Morality Seriously: A Defense of Robust Realism.* Oxford: Oxford University Press.

Ereshefsky, Marc. 2008. Species. In Edward N. Zalta (ed.), The *Stanford Encyclopedia of Philosophy*, Fall. Stanford: Metaphysics Research Lab, Stanford University.

Fantl, Jeremy. 2006. Is Metaethics Morally Neutral? *Pacific Philosophical Quarterly* 87 (1):24–44.

Fine, Kit. 2012. Guide to Ground. In Fabrice Correia and Benjamin Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (37–80). Cambridge: Cambridge University Press.

Fodor, Jerry A. 1974. Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese* 28 (2):97–15.

Fodor, Jerry A. 1975. The *Language of Thought.* Cambridge, MA: Harvard University Press.

van Fraassen, Bas C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Franklin-Hall, Laura R. 2007. Bacteria, Sex, and Systematics. *Philosophy of Science* 74 (1):69–95. Originally published under the name "Laura R. Franklin".

Franklin-Hall, Laura R. 2015. Natural Kinds as Categorical Bottlenecks. *Philosophical Studies* 172 (4):925–48.

Friedman, Michael. 1974. Explanation and Scientific Understanding. *Journal of Philosophy* 71 (1):5–19.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought.* Cambridge, MA: MIT Press.

Gerde, Eric, and Marder, M. 2001. Friction and Fracture. *Nature* 413 (6853):285–8.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge, MA: Harvard University Press.

Greco, Daniel. 2015. Epistemological Open Questions. *Australasian Journal of Philosophy* 93 (3):509–23.

Grimm, Stephen R. 2008. Explanatory Inquiry and the Need for Explanation. *British Journal for the Philosophy of Science* 59 (3):481–97.

Hall, Ned. 2011. The Large-Scale Joints of the World. *Humana.Mente* 19:11–39.

Hammond, Peter J. 1988. Consequentialist Foundations for Expected Utility Theory. *Theory and Decision* 25 (1):25–78.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Noûs* 34 (1):31–55.

Hazlett, Allan. 2017. Understanding and Structure. In Stephen R. Grimm (ed.), *Making Sense of the World: New Essays on the Philosophy of Understanding.* Oxford: Oxford University Press.

Hendry, Robin Findlay. 2010. The Elements and Conceptual Change. In Helen Beebee and Nigel Sabbarton-Leary (eds.), *The Semantics and Metaphysics of Natural Kinds* (pp. 137–58). Oxford: Routledge.

Kanamori, Akihiro. 2009. *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*, 2nd edn. Berlin: Springer.

Kitcher, Philip. 1984. Species. *Philosophy of Science* 51:308–33.

Kitcher, Philip. 2008. Carnap and the Caterpillar. *Philosophical Topics* 36 (1):111–27.

Koellner, Peter. 2014. Large Cardinals and Determinacy. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* Spring. Stanford: Metaphysics Research Lab, Stanford University.

Kornblith, Hilary. 1993. *Inductive Inference and Its Natural Ground.* Cambridge, MA: MIT Press.

Krim, Jacqueline. 1996. Friction at the Atomic Scale. *Scientific American* 275 (4):74–80.

Larson, Paul B. 2012. A Brief History of Determinacy. In Akihiro Kanamori, Dov Gabbay, and John Woods (eds.), *Handbook of the History of Logic,* Volume 6: *Sets and Extensions in the Twentieth Century* (pp. 457–508). Amsterdam: North Holland.

Lasersohn, Paul. 2018. Common Nouns as Variables: Evidence from Conservativity and the Temperature Paradox. In Robert Truswell, Chris Cummins, Caroline Heycock, Brian Rabern, and Hannah Rohde (eds.), *Proceedings of Sinn und Bedeutung 21*, vol. 2 (pp. 731–46).

Levine, Joseph. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64:354–61.

Lewis, David. 1983. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61 (4):343–77. Reprinted in Lewis (1999: 8–55).

Lewis, David. 1984. Putnam's Paradox. *Australasian Journal of Philosophy* 62 (3):221–36. Reprinted in Lewis (1999: 56–77).

Lewis, David. 1991. *Parts of Classes.* Oxford: Basil Blackwell.

Lewis, David. 1999. *Papers in Metaphysics and Epistemology.* Cambridge: Cambridge University Press.

Lipton, Peter. 2004. *Inference to the Best Explanation*, 2nd edn. London: Routledge.

Maddy, Penelopea. 2011. *Defending the Axioms: On the Philosophical Foundations of Set Theory*. Oxford: Oxford University Press.

Mallon, Ron. 2003. Social Construction, Social Roles, and Stability. *Socializing Metaphysics: The Nature of Social Reality* (pp. 327–353). Lanham: Rowman & Littlefield.

Malzbender, J., den Toonder, J.M.J., Balkenende, A.R., de With, and G. 2002. Measuring Mechanical Properties of Coatings: A Methodology Applied to Nanoparticle-Filled Sol–Gel Coatings on Glass. *Materials Science and Engineering: R: Reports* 36 (2):47–103.

Margolis, Eric, and Laurence, Stephen. (eds.) 1999. *Concepts: Core Readings.* Cambridge, MA: The MIT Press.

Margolis, Eric, and Laurence, Stephen. 2014. Concepts. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* Spring. Stanford: Metaphysics Research Lab, Stanford University.

Martin, Donald A. 1998. Mathematical Evidence. In H. Garth Dales and Gianluigi Oliveri (eds.), *Truth in Mathematics* (pp. 215–231). Oxford: Clarendon Press.

Mayr, Ernst. 1996. What Is a Species, and What Is Not? *Philosophy of Science* 63 (2):262–77.

Millikan, Ruth Garrett. 1999. Historical Kinds and the "Special Sciences". *Philosophical Studies* 95 (1):45–65.

Miyoshi, Kazuhisa. 1999. Aerospace Mechanisms and Tribology Technology: Case Study. *Tribology International* 32 (11):673–85.

Montague, Richard. 1973. The Proper Treatment of Quantification in Ordinary English. In Jaakko Hintikka, Julius Moravcsik, and Patrick Suppes (eds.), *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics* (pp. 221–242). Dordrecht: Reidel.

Morgan, Thomas H., Sturtevant, Alfred H., Muller, Herman J., Bridges, Calvin B. 1915. *The Mechanism of Mendelian Heredity.* New York: Henry Holt.

Newey, Charles, and Weaver, Graham. (eds.) 2013. *Materials Principles and Practice: Electronic Materials Manufacturing with Materials Structural Materials.* Amsterdam: Elsevier.

Nolan, Daniel. 1999. Is Fertility Virtuous in its Own Right? *British Journal for the Philosophy of Science* 50 (2):265–82.

Oddie, Graham. 2005. *Value, Reality, and Desire.* Oxford: Oxford University Press.

Pérez Carballo, Alejandro. 2016. Structuring Logical Space. *Philosophy and Phenomenological Research.* 92 (2):460–91. Early View, 2014.

Pietroski, Paul. 2003. The Character of Natural Language Semantics. In Alex Barber (ed.), *Epistemology of Language* (pp. 217–56). Oxford: Oxford University Press.

Portmore, Douglas W. 2009. Consequentializing. *Philosophy Compass* 4 (2):329–47.

Rosen, Gideon. 1994. Objectivity and Modern Idealism: What is the Question? In Michaelis Michael and John O'Leary-Hawthorne (eds.), *Philosophy in Mind.* Dordrecht: Kluwer.

Rosen, Gideon. 2010. Metaphysical Dependence: Grounding and Reduction. In Bob Hale and Aviv Hoffmann (eds.), *Modality: Metaphysics, Logic, and Epistemology.* Oxford: Oxford University Press.

Rosen, Gideon. 2015. Real Definition. *Analytic Philosophy* 56 (3):189–209.

Rudd, R.E., and Broughton, J.Q. 2000. Concurrent Coupling of Length Scales in Solid State Systems. *physica status solidi (b)* 217 (1):251–91.

Schaffer, Jonathan. 2004. Two Conceptions of Sparse Properties. *Pacific Philosophical Quarterly* 85 (1):92–102.

Scharp, Kevin. 2013. *Replacing Truth.* Oxford: Oxford University Press.

Schwalbe, Ulrich, and Walker, Paul. 2001. Zermelo and the Early History of Game Theory. *Games and Economic Behavior* 34 (1):123–37.

Shelah, Saharon, and Woodin, Hugh. 1990. Large Cardinals Imply that Every Reasonably Definable Set of Reals is Lebesgue Measurable. *Israel Journal of Mathematics* 70 (3):381–94.

Sider, Theodore. 1996. Naturalness and Arbitrariness. *Philosophical Studies* 81 (2):283–301.

Sider, Theodore. 2013. *Writing the Book of the World.* Oxford: Oxford University Press.

Stalnaker, Robert C. 2002. Epistemic Consequentialism. *Proceedings of the Aristotelian Society Supplementary Volume* 76 (1):153–68.

Stanford, P. Kyle. 1995. For Pluralism and against Realism about Species. *Philosophy of Science* 62 (1):70–91.

Stoljar, Daniel. 2015. Chomsky, London and Lewis. *Analysis* 75 (1):16–22.

Tappenden, Jamie. 2005. Proof Style and Understanding in Mathematics I: Visualization, Unification and Axiom Choice. In Paolo Mancosu, Klaus Jørgensen, and Stig Pedersen (eds.), *Visualization, Explanation and Reasoning Styles in Mathematics* (pp. 147–214). Dordrecht: Springer.

Tappenden, Jamie. 2008a. Mathematical Concepts and Definitions. In Paolo Mancosu (ed.), *The Philosophy of Mathematical Practice* (pp. 256–75). Oxford: Oxford University Press.

Tappenden, Jamie. 2008b. Mathematical Concepts: Fruitfulness and Naturalness. In Paolo Mancosu (ed.), *The Philosophy of Mathematical Practice* (pp. 276–301). Oxford: Oxford University Press.

Taylor, Barry. 1993. On Natural Properties in Metaphysics. *Mind* 102 (405):81–100.

Thomasson, Amie L. Chapter 21, this volume. A Pragmatic Method for Conceptual Ethics. In Alexis Burgess, Herman Cappelen, and David Plunkett (eds.), *Conceptual Engineering and Conceptual Ethics.* Oxford: Oxford University Press.

Varzi, Achille. 2015. Boundary. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* Winter. Stanford: Metaphysics Research Lab, Stanford University.

Walley, Stephen M. 2012. Historical Origins of Indentation Hardness Testing. *Materials Science and Technology* 28 (9–10):1028–44.

Welch, P. D. 2015. Large Cardinals, Inner Models, and Determinacy: An Introductory Overview. *Notre Dame Journal of Formal Logic* 56 (1):213–42.

White, Roger. 2005. Explanation as a Guide to Induction. *Philosopher's Imprint* 5 (2):1–29.

White, Roger. 2007. Does Origins of Life Research Rest on a Mistake? *Noûs* 41 (3):453–77.

Williamson, Timothy. 2000. *Knowledge and its Limits.* Oxford: Oxford University Press.

Wilson, Mark. 1982. Predicate Meets Property. *Philosophical Review* 91 (4):549–89.

Wilson, Mark. 2006. *Wandering Significance: An Essay on Conceptual Behavior*. Oxford: Oxford University Press.

Wilson, Mark. 2008. Beware of the Blob: Cautions for Would-Be Metaphysicians. In Dean Zimmerman (ed.), *Oxford Studies in Metaphysics,* vol. 4 (pp. 275–320). Oxford: Oxford University Press.

Wilson, Mark. 2013. What is Classical Mechanics Anyway? In Robert W. Batterman (ed.), *Oxford Handbook of the Philosophy of Physics* (pp. 43–106). Oxford: Oxford University Press.

Wong, Wai-hung, and Yudell, Zanja. 2015. A Normative Account of the Need for Explanation. *Synthese* 192 (9):2863–85.

Yalcin, Seth. 2011. Nonfactualism about Epistemic Modality. In Andy Egan and Brian Weatherson (eds.), *Epistemic Modality* (pp. 295–334). Oxford: Oxford University Press.

Zhang, J., Johnston, J., Chattopadhyay, A. 2014. Physics-based Multiscale Damage Criterion for Fatigue Crack Prediction in Aluminium Alloy. *Fatigue and Fracture of Engineering Materials and Structures* 37 (2):119–31.

Zimmerman, Dean. (ed.) 2008. *Oxford Studies in Metaphysics*, vol. 4. Oxford: Oxford University Press.

16

Analyzing Concepts and Allocating Referents

Philip Pettit

What does philosophical theory seek in accounts of the familiar topics it addresses: time or change or causation, belief or intention or freewill, value or responsibility, justice or liberty? One broad family of responses suggests that it primarily tries to analyze the concepts that we express in such terms and their cognates, while another insists that the main aim is to explain what the patterns tracked by those concepts involve, assuming that the patterns are indeed real: it explores the properties or configurations of properties that constitute them. This chapter argues that those approaches highlight different aspects of the philosophical task and defends a more ecumenical perspective on the issue raised.

The analysis of the concept in any such case is an important part of the philosophical exercise but has to be complemented by an explanation of what constitutes the pattern it tracks, assuming again that it does track something real. Where the first part of the exercise involves conceptual analysis, the second involves allocating a referent to the concept analyzed. The analysis of the concept is a form of interpretation insofar as it teases out connotations of the term involved, the allocation of a property is a form of interpretation insofar as it identifies that which the term stands for.[1]

Thus, if the analysis of causation argues that every causal relationship has to take time and involve local connections between cause and effect, the allocation of a referent to that concept must identify in the world—inevitably, the world as seen from within a certain theory—a relation that explains those temporal and spatial features. If the analysis of freewill argues that every freely chosen act must be such that the agent could have done otherwise, the allocation of a referent to that concept must identify a feature of agents that explains why, in an appropriate sense, that sort of claim might be true. If the analysis of value argues that the judgment that

[1] Allocative interpretation is interpretation in the sense in which we speak of interpreting formal symbols in logic. Allocating a referent to a concept addresses the problem of locating that referent in a background picture of the universe—Frank Jackson (1998) describes this as the location problem in metaphysics—and this makes the term 'allocation' particularly appropriate to the way in which I use it here.

Philip Pettit, *Analyzing Concepts and Allocating Referents* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Philip Pettit.
DOI: 10.1093/oso/9780198801856.003.0016

something is valuable is non-inductively connected with a desire for its realization, the allocation of a property to the concept of value has to make sense of why that connection should hold. And so on.

This analytical-allocative pattern will only obtain, of course, in the case of concepts that are designed to track how things are in the world, not just to express feeling or emotion or whatever. And it will only obtain in the case of concepts that do actually succeed in tracking something objective about how the world is. But the focus of the chapter will be precisely on concepts that meet those conditions, being both cognitively oriented and actually satisfied. More specifically, it will be on concepts of that kind that are non-basic: that is, concepts such that mastery of them presupposes a mastery of distinct, simpler concepts. The concepts of causation, freewill, and value may be taken as exemplars of the targets addressed.[2]

To focus on cognitively oriented, actually satisfied concepts is not to suppose that there are no purely expressive concepts and no concepts that fail to track anything real; it is not to beg the question against expressivism or error theory. The attempt to analyze various concepts may reveal that they are purely expressive, the attempt to allocate properties to them that they embody an error about the world. Like the focus on non-basic concepts, the focus on cognitively oriented, actually satisfied concepts is adopted merely with a view to simplifying and sharpening the topic of discussion.

The chapter is in five sections. The first argues for a particular picture of what conceptual analysis involves, according to which it reveals various commitments made by those who use a term or concept. The second section discusses two models of how analytical and allocative interpretation might relate to one another; under one model analysis determinately constrains the allocation of referent, under the other it provides a limited license for allocating one of a number of candidate referents. The third section argues that the license model of how analysis relates to allocation applies in salient, philosophically interesting cases, distinguishes between two different grounds on which it may apply, and illustrates how it is used and to what effect. The fourth section considers the distinct genealogical approach to philosophical theory, arguing that it represents a way of serving the analytical and allocative purposes at once. And the final, fifth section illustrates the license model in action, documenting the extent to which the commitments surrounding the concept of political freedom leave room for many different accounts of what constitutes it as an ideal: that is, of what referent ought to be allocated to the concept, at least for the purposes pursued in political theory.

The license model defended argues that there is a great deal of room for negotiation on two fronts. First, in determining the right specification to link with a concept like that of causation or freewill or value: that is, in determining what counts as an instance of causation, an exercise of freewill, or a form of value. And, second, in identifying the property or other entity that satisfies the specification: in explaining what constitutes causation or freewill or value. It defends an image in which analytical and allocative materials impose a discipline on philosophical theory but

[2] There is a discussion in a footnote in section 4 on how bedrock concepts may lend themselves to what I describe as a genealogical treatment.

make room for imagination as well as method. In taking that line it connects with a variety of recent, more revisionary approaches to philosophy.[3]

1. The Place of Analysis in Philosophy

Two Inadequate Accounts of Analysis

By the comments offered so far, analysis teases out the connotations of a term or concept like that of causation or freewill or value. That very abstract and vague account of what it does can be articulated more concretely and exactly in a number of different ways. Before offering the articulation endorsed here, it may be useful to mention two alternative accounts that it opposes. One is defended by G. E. Moore (1903) and takes us back to the early days of analytical philosophy. The other is assumed by Mark Johnston and Sarah-Jane Leslie (2012) in the course of an attack on a particular school of philosophical analysis, which they describe as the Canberra plan.[4]

According to Moore's view, analysis of any term or concept, X, consists in spelling out the meaning of the word involved, where the meanings of their words are supposed to be immediately accessible to competent speakers. Competent speakers, roughly, are speakers who understand what they are saying. Although Moore does not make this qualification, they not only exclude those lacking expertise in the relevant language but also those who rely on the understanding of others to fix the meanings of their words. This is what I do when I use the word 'quark' or 'boson' or even 'force', intending that it be taken to have the meaning that expert physicists give it.

Moore's view of analysis leads him to propose an open-question test, as it has come to be known, for whether an analysis of a concept is successful. According to that test, the analysis of a concept, 'X', say one that equates it with 'Y+Z', must satisfy the following condition: that it should make no sense for the competent person to be able to ask with an open mind whether an instance of Y+Z really is an instance of X. That ought to be almost as pointless as asking whether an instance of X really is an instance of X. The analysis of a bachelor as an unmarried male might pass this test but few analyses of the kind that philosophers have traditionally canvassed would do so; it sets the bar for a successful analysis so high that it is doubtful if many concepts are analyzable by that criterion.

According to Johnston and Leslie, to turn to the other view, what analysis in the style of the Canberra plan seeks to do is to assemble those associated, allegedly general claims that guide competent speakers in their use of any term or concept, X. The idea is that those claims will spell out the 'application conditions' of the term in question, if

[3] The approaches I have in mind are well represented in this volume. They are associated with figures as different as Rudolf Carnap, W. V. O. Quine, Peter Strawson, Sally Haslanger, Peter Railton, David Chalmers, and indeed the editors of the volume, Alexis Burgess, Herman Cappelen, and David Plunkett. For a book-length development of one such approach, with a critical but fair overview of alternatives, see (Cappelen 2018).

[4] Their critique is focused particularly on the work of Frank Jackson (1998), including a joint paper with David Chalmers (2001).

not its Moorean meaning; they will determine when the term applies and when it does not. They take the Canberra plan to be inspired and implemented by David Lewis but focus in particular on the work of Frank Jackson (1998), including a joint paper with David Chalmers (2001).

As against this plan of analysis, however, Johnston and Leslie draw on empirical findings to argue that it is not general but generic claims that guide competent speakers in their mastery and employment of different terms. Generic claims about things that deserve to be described as X's are truths that hold only of some X's but are so salient to speakers that they serve to mark out the category effectively. Examples might be: 'lions have manes', when only mature males do, and 'mosquitoes carry malaria', when only some species of mosquito are carriers. If the goal of analysis is to identify the application conditions for a term like 'X', according to their critique, then looking to the claims that guide speakers will not deliver the end sought. It will leave analysts with truths that hold only of some X's or, worse, will lead them to misrepresent those truths as holding of all.

Where Moore's picture of analysis sets the bar of successful analysis so high that analysis is impossible in the case of most phenomena, this alternative picture sets it so low that analysis is of little or no philosophical interest. The first approach suggests that few terms or concepts in our vocabulary can be analyzable, since only a few tolerate an analysis that satisfies the open-question test. The second suggests that while it may be easy to identify the claims that guide us as competent speakers, there is no philosophical point in trying to isolate them; they are not general claims of the sort that a philosophical analysis would naturally seek.

But there is a mistaken assumption that is endorsed in common by these two representations of analysis and once we recognize the mistake involved, we can see our way to a very different account. The two accounts both assume that the claims that analysis seeks to isolate in relation to any term or concept 'X' must all lie within the easy grasp of competent users; they must all be platitudes that answer to the untutored intuition of speakers. But, as we shall now see, it is not true that analysis inevitably targets, or ought to target, only claims that are matters of common intuition.[5]

An Alternative Account of Analysis

A common theme among philosophers who give importance to conceptual analysis, including those associated with the Canberra plan, is precisely that the claims that analysis seeks to isolate may not be readily accessible to ordinary speakers. There may have to be a sense in which speakers acknowledge those claims, knowing them to be true, but this does not mean that they have to be able to recognize them readily as truths. As Frank Jackson and I (1990: 35) put the point in arguing for an analysis of certain psychological concepts, 'what people know is not limited to what they can write down on paper off the bat' (see too Pettit 1998).

[5] The mistaken assumption identified here may explain the now common belief that analysis should be supplemented, if not replaced, by empirical investigation of the claims that speakers actually find intuitive: that experimental philosophy, so called, should seek out a census of such intuitions.

But how does it come to be the case that speakers are committed to general truths of the kind that analysis seeks to identify, without being poised to spell out the claims involved as matters of immediate intuition? The answer is that exchanges between speakers can be informative and useful only to the extent that they make certain presuppositions in common, where those presuppositions may often be quite difficult to isolate.

In any exchanges, speakers must assume as a default—a defeasible default—that they are addressing the same subject-matter, not using the same words to mean different things; otherwise it would be hard for them to initiate or maintain conversations.[6] But the participants in a conversation will be unable to sustain that default if it turns out that they differ greatly in the things that they each presuppose about matters bearing on the terms they employ.

If I use the term 'X' in such a conversation, claiming for example that all X's are F, I will communicate nothing to you unless we each hold by various beliefs that help to identify the same things as X's and to identify the same property as F-ness. And I will have no ground for thinking that I can communicate something to you unless I assume, and assume that you assume, that we share those beliefs. At the limit, indeed, the beliefs may have to be a matter of common assumption or common ground, with each of us holding them, holding that each of us holds them, holding that each holds that each of us holds them, and so on (Lewis 1969).

The presuppositions we share as interlocutors with one another may vary, of course, depending on how far our backgrounds or interests coincide. But in order for conversations to connect with one another across different groups of interlocutors, there will have to be presuppositions that are shared quite widely in a linguistic community and shared as a matter of common assumption. Every conversation will start from a purportedly shared set of presuppositions and put them to the test. The presuppositions will pass that test and make a good claim to be shared as a matter of common assumption, just insofar as the conversation does not uncover and revise them or does not degenerate into a squabble.

The lesson of these observations is that speakers will be able to maintain the default assumption that they are talking about the same things, and considering the same issues about those things—that they are talking about the same X's, for example, and considering the same issue in asking whether they are all F—only insofar as they share a variety of presuppositions as a matter of common assumption. Those presuppositions are likely to include beliefs about the conditions under which their words apply, constituting true assertions; about when the truth of some assertions rules in or rules out the truth of others, or indeed rules neither way; about what such truths imply for the actions it is sensible to take in one or another

[6] For a defense of the default assumption of common address, see (Pettit 1993). It represents an important point of convergence—but only partial convergence—with the view defended by Laura and Francois Schroeter (2014). It is consistent with the line taken of course, that the community assumed in the conversation may vary, so that in one context a word is used on the assumption of common address with a wide group of speakers, in another on the assumption of common address only with a specialized group. Scientists may use a word like 'electron' or 'valence' or 'gene' with one assumption in speaking to ordinary folk, and with another in speaking with fellow specialists; the common ground may be wide and shallow in the one case, narrow and deep in the other.

situation; and so on. They will constitute more or less common ground on which any conversation is bound to build (Stalnaker 1978).

It should be clear that the presuppositions that constitute common ground in this sense may not be readily accessible to any participants in a conversation; few may be able to spell them out, let alone write them down off the bat. That they are common ground may only appear in the fact that speakers are likely to be surprised, even dumbfounded, if someone takes a line in assertion or objection or argument that is inconsistent with them. People's disposition to such surprise will testify to the fact that there is a sense in which they each accept the presuppositions and assume that others accept them; perhaps assume this, indeed, as a matter of common assumption.[7]

For an example of a presupposition that is widely shared among competent speakers, consider the claim that is part of almost every analysis of evaluative concepts: viz., that evaluative facts are supervenient on, and indeed fixed by, non-evaluative facts, however the divide is made between those two categories.[8] The thesis is that there can be no evaluative difference between any two items without a non-evaluative difference, because any such difference has to be explained in non-evaluative terms. That claim is not a platitude in the ordinary sense of a banal, immediately intuitive claim.[9] It becomes compelling only after reflection—typically, philosophical reflection—on the fact that it would make no sense to say that A and B differ in value while claiming that they do not differ in any other way. If you made such a claim in exchange with others, you would put yourself out of the conversation.

This example typifies the sort of fact that philosophical analysis claims to be able to uncover. Other examples would include purported facts, no doubt in need of qualification, like the following: that one event can be the cause of another only if it is locally connected to the other, whether directly or indirectly; that a system counts as an agent only if, in general, it has purposes, forms reliable representations, and pursues those purposes reliably according to those representations; that to hold a representation such as the belief that p is to be prepared to act and argue as if it were the case that p; and that to enjoy freewill in the exercise of a certain choice, performing it voluntarily, is to be fit to be held responsible for the choice.

The claims that philosophical analysis aspires to support, then, extend beyond the immediately intuitive and the banal. They constitute working assumptions to which competent speakers are committed by their shared argumentative and related habits and may go well beyond anything that is accessible without much thought. For that reason, I will describe them as conceptual commitments. How those commitments come to be generated and shared is not a topic that can be investigated here, although

[7] If this is so, then it would support a very different role for experimental philosophy from that associated in an earlier footnote with the assumption that analysis targets only matters of immediate intuition. Rather than seeking a census of intuitions, experimental philosophy would do better to identify the presuppositions of speakers that have the following character: if speakers begin to make contributions that are inconsistent with them, then that dumbfounds interlocutors.

[8] For an account that qualifies this claim, while describing it as 'the least controversial thesis in meta-ethics', see (Rosen 2014).

[9] It is unfortunate that the word 'platitude' is often used by defenders of the Canberra plan, as indeed Johnston and Leslie (2012) emphasize.

it ought to be clear that argument and exchange depends on their being present as a matter of common awareness among participants.

I said earlier that even philosophers broadly associated with the Canberra plan recognize that the commitments unearthed in analysis may be not be readily available to intuition. This is particularly salient from the fact that they expect analysis in some areas to uncover the sorts of widespread or holistic connections that can be spelled out only by resort to complicated Ramsey sentences, as they are known. The idea is that there are networks of related concepts such that to posit that one applies may be to require, and so implicitly to posit, a pattern of application across many other concepts in the set as well. Ramsey-style analyses have been routinely invoked in the analysis of mental and scientific concepts (D. Lewis 1983a) but have also been defended in other areas, as in the analysis of moral concepts (Jackson and Pettit 1995).

2. Interpretation, Analytical and Allocative

After Analysis

Taking analysis to uncover the commitments incurred by our use of certain terms or concepts, rather than recording just our unreflective intuitions, lets us see that it may be a significant exercise in philosophy. But analysis, of course, is incomplete, at least on our assumption that the concept analyzed is designed to characterize or track things—it is not just a device for expressing desire or feeling or whatever—and is successful in this role: there is something it actually tracks. Analysis opens the way to asking what the concept directs us to in the world, as we understand the world: what referent ought to be allocated to it, constituting what we conceptualize, for example, as causation or freewill or value.

Philosophical accounts of any phenomenon naturally take the form of first providing an analysis of the appropriate concept and then determining what referent, if any, should be allocated to it. The referent allocated to the concept may be significant in a manner that vindicates the concept and associated discourse, explaining the importance attached to it in our practice. But it may also fail, by one or another background set of criteria, to count as significant, so that pairing it with the concept may amount to deflating, even debunking the discourse. We would take a deflating or debunking view of moral discourse, by standard criteria, if we took the concept of goodness or rightness to direct us just to the fact that it attracts us as something to recommend. And we would take that sort of view of causal discourse, by similar criteria, if we took the concept of a cause to direct us merely to a regular antecedent of the would-be effect.

The set of commitments associated with a concept may be structured in different ways and this structure will affect the way they play a role in selecting a referent for the term. The commitments may all be relevant and necessary in selecting a referent, so that they constitute a regular, conjunctive set. Or it may be necessary only that a certain number of commitments are satisfied by any referent, in which case, the relevant set will be a disjunct of proper subsets. In either event, to complicate matters further, the relevant set may be fixed on a standing basis or it may be filtered, now in favor of one subset, now in favor of another, depending on the context of usage. In what follows, however, we may abstract from these complicating details; the claims made are defended on orthogonal grounds.

There are two pictures of how the analytical and allocative elements in this account of philosophical theory relate to one another. The first model takes the analysis of a concept to constrain the possible referent determinately, even uniquely; the second takes it to license the identification of any in a limited set of candidate referents: in effect, to license independent philosophical theorizing. We may describe these respectively as the constraint and the license models of the relationship between analysis and allocation, although strictly they should be described as the determinate-constraint model and the limited-license model.

The Constraint and License Models

The constraint model corresponds to what we described in the introduction as a standard picture of the relationship between analysis and allocation. According to this model or picture, the analysis of a concept puts us in a position to determine which among the properties or other items countenanced in a background, independently accepted view of the world should be allocated to it as its referent. The analysis takes the concept to constitute a more or less uniquely constraining template and the task on the allocative side is simply to identify the referent that satisfies the template.

The license model of how analysis and allocation connect is quite different. It suggests that despite our best efforts in analysis of any concept, it may well be that the analysis allows us to assign one or another type of referent to it: that it leaves that allocative choice open. The idea is that the purposes that shape the common language from which philosophical analysis generally begins are not very demanding, and that they may be perfectly well served by shared commitments that fail, by our view of the world, to identify a unique referent. Thus, analysis may give us a great deal of latitude as philosophers to say what in the world constitutes the phenomenon targeted by the concept: what, for example, constitutes causation or freewill or value.

That the commitments fail to identify a unique referent, however, does not mean that the discourse in which the concept figures must suffer from a damaging form of ambiguity; this would mean that interlocutors could not be clear about exactly what anyone had in mind in using the term. The differences between the distinct referents that our view of the world enables us to draw may not matter for the purposes of ordinary conversation; thus, for those purposes the concept may refer us just to an equivalence class of referents. Or while there is some ambiguity, with the concept being used in one context to pick out one referent, in another to pick out a different one, the differences of context may be well enough marked for this variation not to create a problem.

For an example of the first possibility, think about how we might speak of one event determining another, without drawing a finer philosophical distinction between causal determination and the sort of constitutive determination that holds, as we noted above, between the descriptive and the evaluative; that distinction may not matter in ordinary exchange. For an example of the second, we can turn to the concept of political freedom discussed in the final section. To anticipate that discussion, we may focus in some contexts on the way physical constraints remove someone's latitude of choice, and in another on legal constraints alone; we may say in one context that someone is unfree to vote because of being ill, in another that despite being ill they are free to vote because of having citizenship.

As analysis is necessary on the constraint model, so it remains necessary on the license model. A theory of X—a theory of causation or freewill or value—will only count as a theory of X insofar as it explains why any instance of X must broadly satisfy the commitments associated with the concept 'X': as we naturally, say, it must satisfy the connotations of the term. Let the theory not require any instance of X to satisfy those commitments, or not require that it at least come close to satisfying them, and the theory will change the subject: it will not be a theory of X, properly understood, but of something more or less related, X*.[10]

The necessity of analysis, in particular analysis of something close to what we describe as conceptual commitments, is defended by Mark Johnston (1992: 221) in his earlier work on color.

> There are many beliefs about color to which we are susceptible, beliefs resulting from our visual experience and our tendency to take that visual experience in certain ways. Some of these beliefs are "core" beliefs in this sense: were such beliefs to turn out not to be true we would then have trouble saying what they were false of, i.e., we would be deprived of a subject matter rather than having our views changed about a given subject matter.

Insofar as a phenomenon like color is conceptualized on a basis that involves holding by such core beliefs or commitments, any philosophical theory of color is required, on pain of changing the subject, to support at least a good many of them.

But while the license model resembles the constraint model in making analysis necessary for a philosophical account of any phenomenon, it generally leaves much more for philosophical theorizing to do. Suppose we are interested in a theory of causation or freewill or value, to stick with those examples. The fact that there is a choice to be made in allocating a referent to the corresponding concept means that there is work to be done in selecting the best referent to allocate: in developing the best view available, by independent criteria, of what it is that constitutes causation or value or freewill.

Thus suppose, plausibly, that it is an inescapable commitment associated with ascribing freewill that the agent could have done otherwise in any relevant choice. In that case, we must find a pattern in the choice in virtue of which that counterfactual, or something close to it, holds. And if we are naturalists, perhaps even determinists, we face a well-known challenge in identifying a pattern in the world, as we take the world to be, that would do this job. In facing that challenge, analysis as such will not

[10] These remarks on the license model presuppose a wish to maintain continuity with the way of thinking encoded in our ordinary usage of a term. What is also possible, of course, is that we might judge that while a term can serve in broadly the same role—say, it can preserve its extension—the commitments associated with it, as revealed in standard analysis, should be radically altered. This is the 'ameliorative' line that Sally Haslanger (2012) has defended with concepts of race and gender; for a broadly similar approach, see (Eklund 2017).

help; it will merely provide a condition of success: that our theory must make sense of the counterfactual.

Two Grounds for License

The most straightforward, if not the most common way, in which analysis may license independent philosophical theorizing arises when the commitments it links with a term or concept are relatively few and undemanding. For an example of this possibility consider John Rawls's approach to the theory of political justice.

Different political theorists, so Rawls (1971: 5–6) imagines, 'agree that institutions are just when no arbitrary distinctions are made between persons in the assigning of basic rights and duties and when the rules determine a proper balance between competing claims to the advantages of social life'. And with that base of agreement to determine the common topic they address, they generate conceptions of justice that divide on 'which differences among persons are relevant in determining rights and duties' and 'which division of advantages is appropriate'. Whether differences are relevant, and a division appropriate, will presumably be determined, on this approach, by independent judgments of moral merit.

Where Rawls uses the language of concept and conception, we would speak of concept and concept-property pairing. But his picture of what the philosophical theory of justice involves fits in a straightforward way with the license model. Conceptual analysis identifies a few, relatively vague commitments associated with the idea of justice: that there be no arbitrary distinctions made between persons and that a proper balance be established between people's competing claims. And those commitments then open up the question of what referent to allocate to the concept. They give us as philosophers the license to think about what constitutes justice at the level of institutional arrangements.

When the model gives a license for independent, non-analytical theorizing of the kind illustrated here, the analysis identifies a few determinate commitments, C1–Cn, that might be satisfied by any of a range of properties, P1 or P 2 . . . or Pn, that are available within our view of the world: within our view of the social institutions possible. And the job of philosophical theory is to argue in favor of one or another property, allocating it as referent to the concept: to argue on the basis of independent criteria that that property should be taken to constitute justice in the analyzed sense. Following that procedure, of course, Rawls himself argues for an institutional framework that satisfies his two principles of justice, one involving liberty, the other socio-economic equality.

Analysis may license independent philosophical theorizing on a second quite different ground as well. Rather than isolating so few commitments that any of a range of properties can satisfy them, it may isolate so many putative commitments that it licenses pruning them in this or that manner and, depending on what subset survives the pruning, allocating this or that referent to the concept. Where the license given in the too-few case is a license to pick and choose externally to analysis, the license given in this case is to pick and choose internally: that is, to pick and choose among which putative commitments to ratify as important.

In the work already cited, Mark Johnston (1992) defends a view of the commitments surrounding the concept of color that fits with this picture. We 'speak more inclusively about color', he says, 'as we underwrite more beliefs with some legitimate title to be included in the core' (p. 221). Thus, suggesting that we may privilege one or another subset of those commitments, an important question to address is how far we may strip down the commitments endorsed without depriving ourselves of the subject. 'How far short of speaking ever so inclusively do we have to fall in order to say truly that the external world is colored' (p. 222). The suggestion is that the commitments associated with the concept of color allow us, depending on whether we privilege the set as a whole, or some proper subset, to have one or another theory of what constitutes color, identifying one or another referent for a color concept. Johnston (1992) takes this suggestion, not just to bear on color, but to 'apply to many if not all concepts' (p. 222).[11]

In the first sort of case, as we saw, philosophical theorizing will have to be driven by independent, non-analytical criteria, such as the criteria of moral merit that Rawls would presumably invoke to support his theory of justice. In the second sort of case, independent, non-analytical criteria are also likely to play a role in supporting one or another theory. There may be analytically motivated reasons for preferring to prune relevant commitments in one direction or another. But there are also likely to be independent, non-analytical reasons for preferring one pruning to another: viz., that under the preferred pruning there is then a salient referent that can be allocated to the concept: a property that can be taken to constitute the phenomenon in question. Such considerations might plausibly operate in the case of color, arguing for ratifying one or another pruning on the grounds that there is something in the scientific image of the world—say, a spectral reflectance profile—that can then be allocated as the referent of a color concept.

The external and internal grounds on which conceptual analysis may license independent philosophical theorizing are not exclusive of one another. Suppose for example that we avail ourselves of the internal license to prune the putative commitments associated with color to a given set, C1, C2, and C3. It may still be the case that those commitments allow us to allocate one or another property to the concept; they may give us an external license to choose between those properties in explaining what constitutes color.

3. In Favor of the License Model of Analysis and Allocation

The Basic Case

The license picture of how analysis relates to allocation looks more plausible than the constraint model, since analysis often falls short of giving us one clear winner for the referent of a concept. First, to rehearse the abstract possibilities, the commitments that analysis of a concept picks out, as in the Rawls example, may allow for the allocation of one or another referent; this means that there is an

[11] Johnston (1992) suggests that the commitments encoded in our use of color concepts may not cohere with one another, in which case it will be obligatory, not just permissible, to prune the total set in favor of one or another coherent subset. I ignore this possibility in the current chapter.

external ground for license in allocating a referent. Second, as in the example from Johnston, analysis may allow for privileging one or another set of conceptual commitments, where different sets invite the allocation of different referents to the concepts; this means that there is an internal ground—a ground internal to analysis—for license in allocating a referent. And, third, analysis may fail on both counts, identifying a number of permissible sets of commitments to associate with a concept, where at least some of those sets may allow for allocating one of a number of referents to the concept.

We may return to the concepts of causation, freewill, and value to illustrate the failure of the constraint model, and the extent to which analysis provides a license to theorize in any of a limited number of ways about such phenomena.

Different accounts of causation differ on how far causation should be linked with counterfactual dependence or should be taken to presuppose a process like the transfer of energy; on whether it presupposes natural laws or the other way around; and on how far various plausible commitments among ordinary thinkers—say, the commitment to exclusively forward-looking causation—can be dropped. These differences may reflect different views on the commitments we make in positing causal connections and drawing inferences from them. Or they may reflect different views on the best referent to allocate to the concept of a causal connection, assuming a more or less agreed set of commitments associated with the concept. Or of course both problems may affect the attempt, on the basis of analysis alone, to construct a philosophically acceptable account of causation.

As the analysis of causation may provide a license to theorize in a number of different ways about it, so the same is true with the other two examples. Different accounts of freewill differ on how far it is tied to practices of responsibility, on whether any exercise of freewill is something the agent had the capacity not to trigger, and on how capacity in that sense is to be understood. Different accounts of value differ on issues like whether there is a non-inductive connection between something's being valuable and its being attractive, whether value is something we posit as a matter of determinable fact, and whether judgments of value can be based on direct intuition. And in each case the problem, plausibly, is that analysis does not provide a determinate constraint on how to think about freewill or value; it provides a license, whether of an external or internal kind, to identify a suitable referent for each concept on the basis of independent criteria of theory-choice.

These observations are meant to show that philosophical theory cannot live by analysis alone: that at least in cases like those illustrated, a candidate analysis needs to be paired with an account of what referents the analysis would allow us to allocate to the concept and of which candidate scores highest on suitable criteria. Analytical interpretation is certainly necessary in any philosophical theory of X; if it neglects analysis, the theory may change the subject and fail to be a theory of X. But, equally certainly, allocative interpretation is needed too and may have to be determined in the presence of a license to go in one direction or another. Any philosophical account of a phenomenon X is likely to have to satisfy two sets of criteria, then, one answering to reflection on the commitments we make in employing the concept of X, the other to an independent sense of what is best taken to constitute X: what referent it is best to allocate to the concept, in light of our overall theory of the world.[12]

The License Model and the Canberra Plan

The license model need not mark a departure from the methodology associated broadly with the Canberra plan. Frank Jackson and I (1995) follow that methodology in developing a meta-ethical theory that we describe, in the title of the paper, as 'Moral Functionalism' (p. 23). And in doing so we explicitly invoke the sort of license discussed here, once for analytically internal considerations and once for reasons that go beyond analysis.[13]

We go with the internal version of the license model when we list a comprehensive list of 'commonplaces'—purported commitments—that most of us will endorse on reflection about a concept like that of rightness or fairness, and then make the point that they leave us with a decision as to exactly which should count as a priori: that is, which should be ratified as part of the final analysis of the concept. They are 'candidates for a priori truths', we say, acknowledging that 'we may expect debates, of course, about which commonplaces are indeed a priori true': which should be suitably ratified (p. 23). The point made here is exactly like the point made by Johnston, when he suggests that we may analyze the concept of color more or less inclusively, as he puts it, depending on how close we come to privileging all of the core beliefs or commitments associated with the concept.

Jackson and I offer a network analysis in the paper on moral functionalism, arguing that concepts like those of rightness and fairness and the like are picked out by roles that they play in a network they form with other normative concepts. But we recognize that even after we have selected the subset of commonplaces to privilege, exercising the internal discretion that the license model allows, the analysis leaves us with external discretion, again in line with the license model, about exactly which sort of property to allocate to each of the concepts.

First, we note, the analysis 'leaves open whether rightness, say, is the ground-level (descriptive) property that occupies the rightness role, for example, the property of maximizing happiness, or whether it is the higher-order property of having a

[12] Timothy Williamson (2007: fn 121–2) takes the sort of analysis associated with the Canberra plan, as exemplified in (Jackson 1998), to be bedeviled by the problem that in many cases there will be a number of 'admissible candidates' for the role of referent for the analyzed term or concept—say, the property it ascribes; that these cannot reliably be conjoined or disjoined to constitute an admissible candidate; and that there may be no uniquely 'natural' candidate to select, even if 'natural' can be appropriately characterized. He would presumably see the line I take in this chapter as seeking vainly to make a virtue of necessity. For his own view of how we are guided in our application of concepts—in his terms, of what binds together the uses of the same word by different agents—see pp. 123–30.

[13] The approach has sometimes been described in a way that seems, rightly or wrongly, to jar with the license model. David Lewis (1983a) may suggest that the concepts analyzed in standard examples constrain the identification of a referent uniquely, for example, when he speaks of the best deserver of a name—the referent that best answers to the concept, under that analysis—or despairing of finding such a referent, speaks of identifying a near-enough deserver. And David Braddon-Mitchell (2003) may make a similar suggestion, when he argues that the analysis of our concept of *qualia* provides materials such that it might have one referent, W, under one conception of the universe, but a different referent in another, even another that also recognizes W; the referent in the first conception would be a near-enough deserver, the referent in the second a best or at least a better deserver.

property that occupies the rightness role' (pp. 27–8). And, second, if the analysis chosen 'identifies the moral properties with the ground-level, role-filling properties', we say that it leaves open whether 'to think of a role in folk moral theory as picking out a moral property in rigid or in non-rigid fashion' (p. 28).[14] A final, philosophical theory might exercise the license allowed on these fronts, either by finding independent grounds for choosing one or the other reading in each case or, in the absence of such grounds, by taking the referents to be equivalence classes of the possibilities that analysis leaves open.

The License Model and Natural Properties

Not only does the license model of the relation between analysis and allocation fit with the Canberra plan, it can also help to make sense of David Lewis's (1983b) views on the relevance of the naturalness or elite-ness of potential referents in interpreting concepts. The claim he makes is that if there is a choice about which of a number of properties to take as referent of a concept, we ought to opt for that which is, in his terminology, the most natural: intuitively, the one with the best claim to approximate cutting nature at its joints, in an established metaphor, not at the joints that happen to be privileged by our conventions.

This view has sometimes been taken to involve a belief in reference magnetism, as it has been called: a belief that the greater naturalness of the property attracts the concept towards it, as if by magic. In our terms, this might be cast as the view that the analysis of a concept is not determined only by the conceptual commitments of speakers but also by an independent fact: the fact that one of the referents that those commitments would allow is more natural than the others. But such a picture does not fit well with other aspects of Lewis's philosophy of language, in particular his emphasis on the role of conventions in fixing the meanings of our words (Schwartz 2014); this would suggest that the analysis of a concept should be guided only by the conventionally registered commitments of speakers. The tension can be resolved, however, on the license model. The point is worth noting, whether or not it fits with everything that Lewis says on the topic.

The license model allows that the allocation of a referent to a concept may be analytically underdetermined: that it may not be uniquely constrained by the conceptual commitments of speakers. And that suggests that the relative naturalness of candidate referents may play a role, not in shaping the analysis of the concept—that is what reference magnetism involves—but in deciding on which referent it is best, by independent, non-analytical criteria, to allocate to it; we look in a moment at what those criteria might be. The idea is that naturalness does not figure in determining what counts as an X, for any concept 'X', but that it may figure in selecting the best theory of what constitutes X.

Thus, consider whether a color word or concept like 'green' refers to a continuing, relatively natural color property or to the relatively non-natural, 'gruesome' property,

[14] As we think about what would be right in a possible, counterfactual world, the rigid option would have our thoughts target the actual filler of the role—say, the maximizing-happiness property—the non-rigid option would have them target whatever fills the role in that world, which may or may not be the happiness-related property.

invented by Nelson Goodman (1983), of being green if examined before time t, blue otherwise. The license model would allow us to take the referent of the concept to be the standing color, on the grounds that it is the more natural property, without relying on anything like reference magnetism. It might do this in either of two ways, one invoking an external reason, the other an internal reason, for choosing to allocate the more natural property to the concept.

The way of doing it that invokes an externally grounded license is the more straightforward. It would argue for pairing the color concept with the standing color as distinct from the gruesome counterpart, on the grounds that all the commitments associated with the concept of green are consistent with allocating one or the other referent, and that its greater naturalness argues in favor of allocating the standing color. The alternative approach, invoking an internally grounded license, would argue, first, that among the commitments governing the concept of green, one perfectly possible inclusion is a commitment to the effect that color properties have a sort of significance that is independent of whether and when they have been observed; and, second, that this commitment should be included in the analysis on the grounds that it would select the more natural, standing color as the referent to be allocated.[15]

Criteria of Allocation

The issue about naturalness raises the question, ignored so far, as to what criteria make one candidate referent more appealing than others, arguing on a non-analytical basis for preferring it in allocating the referent.

A general, plausible suggestion is that depending on the purpose associated with the use of a concept—ultimately the purpose of the sort of discourse in which it figures—a referent will be more suitable to the extent that it serves that purpose better. This would mean that with concepts that are part of an enterprise of mapping a given domain for explanatory ends, it will presumably be better that the concept should identify a property that pulls more explanatory weight, as it were, than one that pulls less. Assume, as Lewis does, that properties vary in their naturalness, where this higher-order property reflects the degree to which a property that bears it carves nature at the joints. In at least some explanatory enterprises, for example in the area of natural science, it will presumably be better to have a more natural rather than a less natural property allocated to a concept.

This thesis is implicitly endorsed in a common critique of ways of thinking that claim to explain—and even justify—existing, often oppressive social structures in by variations in people across class or gender or race. The critique is that those ways of thinking treat concepts like 'class' or 'gender' or 'race' as predicating natural properties, when they actually reflect variable modes of social organization. And the assumption behind the critique is that insofar as people mistakenly seek to explain social patterns in terms of such properties, they take the properties to cut nature at its joints, not around culturally variable divisions. The critique suggests that such

[15] Wolfgang Schwartz (2014), who supports something like this line about grue-ness, is insistent that it does not constitute reference magnetism: see pp. 10–12. For another, congenial critique of such magnetism, see (Sundell 2012).

concepts deserve a debunking analysis under which the properties they ascribe are not as substantive as assumed in the ways of thinking criticized.

Epistemic enterprises may serve very different interests, of course, in different areas of inquiry (Habermas 1971). And to the extent to which they do, the criteria governing the referents that we privilege in the allocative interpretation of relevant concepts will presumably vary as well. Suppose, for example, that the interest in explaining one another's doings consists in being able to represent our respective responses as suitably rational and reason-sensitive: being able to see one another as conversable creatures with whom we might deliberate about things and interact to our mutual gain (Davidson 1984). In that case, it will be important that the concepts we deploy ascribe attitudes of a psychological sort, not just neural states; attitudes with contents we can recognize, rather than attitudes with gruesome contents; and attitudes that conform to requirements of rationality, at least within familiar limitations. Such attitudes may not cut nature at the joints but they will cut nature where it matters to us as mutual interpreters and interactants.

This is not the place to expand on the variety of epistemic interests and the different criteria they may support to govern the allocation of referents to the concepts we analyze. The topic comes up again in the final section of this chapter, where we look at the philosophical theory of freedom, as that property figures in political theory: a normative enterprise with distinctive guiding interests.

Despite the differences between different philosophical areas of theory, it is worth noting that in all of them we may say that, insofar as the theory allocates a particular referent to a concept, indicating the respects in which it is superior to alternative candidates, it will provide an explanation of the phenomenon targeted. The explanation will not be causal but constitutive. It will give an account of the phenomenon, assuming it is not primitive, that identifies in terms of other concepts conditions such that their satisfaction ensures and explains the realization of that phenomenon. Thus, any account of what causation or freewill or value consists in—or indeed class or gender or race—will identify conditions in natural or social reality such that if they are satisfied, then the phenomenon is bound to be present. The point, as we shall see, applies also in the case of political freedom.[16]

4. Genealogy: Melding Analytical and Allocative Interpretation

Before illustrating the license model in the theory of freedom, however, it will be useful to make a further observation. There is one particular methodology in philosophy that goes beyond the two-step approach of looking first for analysis, then for the allocation of referent. This consists in developing a story as to how

[16] Daniel Stoljar (2017: chapters 5 and 6) provides a general account of the sort of philosophical explanation that provides information on what constitutes this or that phenomenon; he emphasizes rightly that depending on the level and kind of information it provides, the explanation may be more or less satisfying. Other accounts suggest that the sort of explanation needed at this point in philosophical theory should explore the ground of the phenomenon, as it is often called, see (Rosen 2010; Fine 2012).

various concepts could have arisen, in particular a vindicatory story that gives them suitable referents; if you like, it reverse-engineers the concepts it explains.

One clear example of such an approach is offered by H. L. A.Hart's (1961) classic study of the concept of law (Gardner 2013). Hart assumes, to begin with, that in almost any society certain norms like those against theft and violence, deception and infidelity, are bound to emerge and stabilize; to become internalized by members, as they prescribe conformity for themselves and others (p. 86); and, because of their importance in social life, to support a relatively authoritative assignment of corresponding rights and duties (pp. 84–5). He describes these norms as primary rules, arguing that they are regularities with which 'the general demand for conformity is insistent and the social pressure brought to bear on those who deviate or threaten to deviate is great' (p. 84); this 'social pressure may take only the form of a general diffused hostile or critical reaction', he says, stopping 'short of physical sanctions'.

Call this society Normitania, after the crucial and exclusive role of primary rules or social norms in securing social order. It does not matter from Hart's point of view that there may never have been a society of exactly the kind he postulates. All that he requires is that it should represent a real possibility, materializing under familiar social conditions in virtue of familiar human needs: this, at any rate, assuming the people are relatively equal in power (p. 195). He then argues for three claims: first, that our sense of human nature makes it intelligible, indeed predictable, that people would make certain unplanned adjustments in response to various problems arising in Normitania; second, that those adjustments would lead in aggregate to the emergence of a system of secondary rules, as he calls them, to regulate the operation of the primary rules; and, third, that those who live under that system—certainly those with official roles—would have a concept available to them, in particular a concept with a clear range of application, that corresponds to our concept of law.

Looking at the first step in his argument, there are a number of problems that primary rules would raise in Normitania, according to Hart. Being only informally established, they would often be uncertain or vague. In addition, they would be static, lacking the flexibility required for covering any novel circumstances that might come from changes in culture or technology. And finally, they would be inefficient in leaving it up to alleged offenders and would-be victims to agree on whether a norm was breached in a given case and on what recompense or retaliation should be implemented.

In response to these problems, Hart suggests in the second stage of his argument, the members of Normitania would be likely to improvise solutions, relying on the salience of certain procedures, the prominence of certain individuals, or the existence of certain precedents, to determine the approach to take. The responses might initially be ad hoc, with some individual or body being called upon to pass judgment in this or that case, or with some process or procedure being selected for that role. But they would be likely to consolidate over time, establishing patterns and procedures—secondary rules, in Hart's terminology—for dealing with the different issues arising.

Those rules would establish ways of determining what exactly any vague primary rules involve, how those rules can be amended to cope with changed circumstances, and the process to follow in passing judgment on whether someone has breached a rule and how the offense is to be rectified or sanctioned. He describes the secondary rules, respectively, as rules of recognition, rules of change, and rules of adjudication, arguing that they would be likely to be internalized in the same manner as primary rules, at least amongst those on whom they confer various roles.

At the third stage of argument, which is more or less implicit in the text, Hart makes two plausible claims: that people living under this regime—or at the very least its officials—would likely have a single concept available to characterize the primary rules that are brought under the regulation of the secondary rules, as well as some of the secondary rules themselves; and that it is plausible to equate that concept with our concept of law, so that Normitania may now be better dubbed Lexitania. Grant these claims, and the genealogy provided supports the hypothesis that our concept of law serves the same purpose as the Lexitanian concept, supporting presumably the same commitments, and that the property that our concept predicates is just the property that Lexitanians find in primary and secondary rules.

Assuming that the hypothesis is accepted—I say nothing here about what would make it acceptable—the genealogical method that Hart illustrates has much to be said for it. Relying on just our interpretive sense that we would naturally equate the concept employed in Lexitania with our concept of law, it directs us to functions that law is conceptually required to serve. Relying on the genealogy to identify the referent of the concept, it points us towards the sort of property it serves to predicate. And to that extent it achieves at least some of the purposes that analysis and allocation, pursued as such, might have hoped to achieve. It removes any mystery as to how we could get the concept of law going, it makes sense of the role of the concept without debunking it, and it directs us to a plausible property that constitutes its referent: this is what the concept serves to ascribe.

There may be other purposes that a direct form of analytical and allocative interpretation might hope to realize better: say, in developing an explicit definition of the concept, or a more exact sense of the property it ascribes, or in distinguishing laws from other rules. But while the genealogy does not achieve those purposes explicitly, it ought to make it easier to realize them. Thus, the genealogical method might be seen as part substitute, part preliminary, to the more demanding, two-stage exercise.[17]

Broadly understood, the genealogical method illustrated may be extended to a range of cases, as when Edward Craig (1990) uses it to explore the concept of knowledge or Bernard Williams (2002) employs it in articulating the notion of truth and truthfulness; in (Pettit 2018a), I rely on it in much the same way to cast light on the wide network of ethical concepts.

Extending the sense of a genealogy somewhat, there are other cases too that illustrate something close to the method illustrated. Think of Wilfred Sellars's (1997) myth of Jones, according to which we could have developed concepts of mental experience and attitude by seeking a theoretical explanation for our

[17] To add to its advantages, the genealogical method is capable in principle of teaching us important lessons about concepts that serve a bedrock role in our thinking, and are not analyzable in independent terms (Chalmers 2011); it can make sense of how we might find a use for those concepts, in particular a use that vindicates their employment. I try to put a broadly genealogical method to this sort of use in seeking to make sense of the concepts of rule and rule-following; see (Pettit 1993: part 1; 2002: part 1).

dispositions to make certain utterances and to take corresponding actions. Think of David Lewis's (1969) demonstration that as self-interested rational agents we could have coordinated with one another in familiar predicaments, and given rise to regularities of the kind that answer to the concept of a convention. Or think of Saul Kripke's (1980) story about how we could have introduced names as causally linked tags for the things we name, without sharing any single view of the descriptive character of the things named (Jackson 2004).

The genealogical method used in Hart and in these examples involves a story about how a whole society might develop a counterpart to the concept targeted, vindicating it in the application of the term. But it should be clear that this full-scale social element is strictly unnecessary from the point of view of the purpose served by the genealogy. And that directs us to another mode of genealogical explanation: the method of creature-construction described by H. P. Grice (1975). This involves telling a story about how even a single individual might construct in stages something complex like a robot; might develop concepts at each stage of construction to characterize the emerging entity; and might come to a point where we would say that the concept we were wanting to analyze—say, that of a belief or an intention—actually applies. The idea then, as in the social genealogies, is that the story would direct us to a likely referent to allocate to the concept: that which is realized in the robot.[18]

5. A Case Study: The Concept of Freedom

Since it has been a focus in my own recent work, I have chosen the concept of freedom in the social and political sense to illustrate in a little more detail how analytical and allocative interpretation work, and how they relate in this case as the license model suggests they would. The concept of freedom is ascribed to societies, people, choices, and actions but I shall focus in the first place on free choices. These are sets of options that are presented to an agent, where the agent can realize any option in the set, depending on his or her preferences.[19] Free actions may be taken to be exercises of free choice, free persons to be agents who are free to exercise suitable choices, and free societies to be communities where its members are free persons.[20]

Freedom is a good example with which to illustrate the license model, since the commitments associated with describing a choice as free are highly contestable and indeed contested. Thus, depending on which commitments are taken as significant,

[18] The method of creature-construction is foreshadowed by Jonathan Bennett (1964) in his study of rationality, invoked by Michael Bratman (2014) in analysis of shared agency, and by Peter Railton (2014) in analysis of belief. The approach taken by Christian List and me in analyzing the idea of group agency might also be cast in this mold; see (List and Pettit 2011).

[19] For more on the concept of an option, see (Pettit 2018b).

[20] This notion of a free action makes it richer than an action freely taken; an action may be freely taken when, unbeknownst to the agent, there is no other option available (Frankfurt 1969): the action is chosen in the presence of apparently acceptable alternatives and is, in that sense, voluntary (broadly following Olsaretti 2004). Equally, this notion of a free society is distinct from the more austere notion of a society that is self-determining: this might be controlled by an elite and contrast with the more traditional idea of a free state (Skinner 1998).

a different analysis is offered and a different property allocated as its referent. Perhaps the only core commitment that no analysis neglects is the assumption that you will enjoy freedom in a choice only to the extent that you are not exposed to hindrance with otherwise available options. Let hindrance be understood in a descriptive sense and not, as in some accounts (Nozick 1974), in a sense that deems something to be a hindrance only if it is wrongful. With that assumption in the background, a number of questions arise as to the further assumptions guiding the concept. And analyses of freedom of choice vary, depending on the answers given. I distinguish five questions that play a particularly crucial role in determining the analysis selected.

The questions all bear on the hindrance that is relevant to the freedom of a choice. To put them in the form of a list, they can be formulated as questions about what sort of hindrance is necessary for reducing your freedom:

1. Does that hindrance have to prevent you from taking an option? Or does it extend to the sort of hindrance involved in penalizing a choice of option, threatening to penalize it, deceiving you about the options, or manipulating you into misperceiving them?
2. Does the hindrance have to be imposed on you from without, whether by other agents or by nature? Or does it extend to include hindrances that derive from psychological disorders or conditions: hindrances like those associated with morbid fears, for example, or obsessions?
3. Does the hindrance have to represent a form of voluntary or at least intentional intervention on the part of an agent or agents, individual or corporate?[21] Or does it extend to the hindrances that non-intentional obstacles can impose, or that arise as an unforeseen consequence of others' actions?
4. Is it necessary for a reduction in your freedom that the hindrance affects the option you actually prefer? Or is it sufficient that you would have suffered hindrance if you had tried to take another option instead?
5. Is it necessary for a reduction in your freedom, assuming another person is involved, that that person actually imposes the hindrance? Or is it sufficient that if their will had been different—if they were not happy to let you choose as you wish—then they would have intervened and imposed a hindrance.

With these questions in focus, we can see that different analytical and allocative interpretations of the concept of freedom become readily identifiable. There are so many different conceptual commitments associated with talk of freedom that we have a choice about which to ratify in analysis and about which property to allocate to the concept. These internal grounds for why philosophy has a license in the theorizing about freedom have been fully exploited in the tradition, giving us different ways of thinking about the topic. These all begin from different analyses of what counts as enjoying freedom and go on to offer different accounts of what constitutes it.

[21] An action will be intentional insofar as it is generated appropriately by the agent's beliefs and desires; it will be voluntary, as suggested in the previous footnote, insofar as it is chosen in the presence of acceptable, or at least apparently acceptable, alternatives.

Thus, to take the sort of view associated broadly with Thomas Hobbes (1994a,b), he answers the questions as follows: yes, only prevention takes away freedom; yes, it has to be from outside; no, the hindrance does not have to be intentional; but yes, the hindrance must affect the actual option preferred; and yes, if another person is involved, they have to intervene as a matter of actual fact (Pettit 2008).

The sort of view associated with Isaiah Berlin (1969) breaks from Hobbes on the first, the third, and the fourth questions (Pettit 2011). He holds that, no, the hindrance that reduces freedom does not have to be preventive; yes, it has to be imposed from without; yes, only an intentional, perhaps only a voluntary intervention can reduce it; no, it may reduce freedom even when it affects an option you didn't actually prefer; but yes, if another person is involved, he or she has to intervene as a matter of actual fact.

What has come to be known as the republican view of freedom represents a salient alternative to these two.[22] It coincides with the Berlinian view on the first four issues. The intervention that reduces freedom, as Berlin holds, may extend beyond preventive hindrance; it must be imposed from without; it must originate in voluntary intervention; and it may reduce freedom by affecting any option in the relevant choice, whether or not it is actually preferred. But, as against Berlin, it holds that it is enough for suffering a reduction of freedom that someone who actually fails to intervene, would have done so, had they wished: they have the power required.

It is implausible to hold that any of these analyses of the concept of freedom in choice is clearly mistaken about the commitments associated with the term; they each take a permissible direction in selecting the commitments to ratify. Drawing on different analyses of freedom, they offer different accounts of the property that constitutes it. What constitutes it in the Hobbesian case, in more or less established language, is the property of non-frustration; in the Berlinian, the property of non-interference; and in the republican, the property of non-domination.

As I have argued elsewhere, the selection of which theory to endorse has to be made on independent, non-analytical grounds (Pettit 2014). The grounds, plausibly, are that the theory preferred satisfies the test of reflective equilibrium better than the alternatives; it helps to shape an overall political philosophy whose verdicts on particular issues, perhaps after some adjustments, constitute judgments that we are disposed on consideration to endorse. These are grounds of the sort that Rawls (1971) explicitly invokes in favor of his own two-principles theory of justice; indeed it is he who introduces the idea of a methodology of reflective equilibrium.[23]

[22] The republican literature is now voluminous. For my own contribution, see (Pettit 1997, 2012b, 2014). For monographs in English, see (Skinner 1998; Brugger 1999; Halldenius 2001; Honohan 2002; Viroli 2002; Maynor 2003; Lovett 2010; Marti and Pettit 2010; MacGilvray 2011; Gourevitch 2014; Taylor 2017; Thomas 2017). And for collections, see (Van Gelderen and Skinner 2002; Weinstock and Nadeau 2004; Honohan and Jennings 2006; Laborde and Maynor 2007; Besson and Marti 2008; Niederbeger and Schink 2012). For a recent review of work in the tradition, see (Lovett and Pettit 2009).

[23] The freedom example illustrates nicely the approach that has come to be described as conceptual ethics (Burgess and Plunkett 2013a,b; Plunkett and Sundell 2013). In terms of that approach, the issue in this case is whether to take the concept of freedom—the best concept, by relevant criteria—to be one that ratifies this or that set of answers to the five questions. The idea is that the analysis of the term 'freedom'

What holds of freedom holds of at least many of the topics that philosophy seeks to illuminate. Whether in providing an account of agency or action, belief or intention, causation, freewill, or value, philosophy does not advance by a consideration of conceptual connotations alone. It may start in any single case from an analysis of such connotations, seeking to satisfy them in some measure. But that starting point leaves it with the main work still to do: providing a constitutive account of the sort of phenomenon the concept targets; in particular, a constitutive account that makes it worthy of a place in the relevant background theory.

Acknowledgements

I was greatly helped in thinking about this topic by conversations with Daniel Stoljar and comments received from him, and by detailed criticisms and suggestions from Victoria McGeer, David Plunkett, and Sam Preston. My thanks to all. I am also grateful for discussion of an early version of the chapter at a conference on 'Deep Disagreements', at New York University, September 2017.

References

Bennett, J. 1964. *Rationality*. London: Routledge and Kegan Paul.

Berlin, I. 1969. *Four Essays on Liberty*. Oxford: Oxford University Press.

Besson, S., and Marti, J. L. 2008. *Law and Republicanism*. Oxford: Oxford University Press.

Braddon-Mitchell, D. 2003. Qualia and Analytical Conditionals. *Journal of Philosophy* 100:111–35.

Bratman, M. 2014. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press.

Brugger, W. 1999. *Republican Theory in Political Thought: Virtuous or Virtual*. New York: Macmillan.

Burgess, Alexis, and Plunkett, David. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, Alexis, and Plunkett, David. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Chalmers, D. 2011. Verbal Disputes. *Philosophical Review* 120:515–66.

Chalmers, D., and F. Jackson 2001. Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110:315–60.

Craig, E. 1990. *Knowledge and the State of Nature*. Oxford: Oxford University Press.

Davidson, D. 1984. *Inquiries into Truth & Interpretation*. Oxford: Oxford University Press.

Eklund, M. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.

Fine, K. 2012. Guide to Ground. In F. Correia and B. Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (pp. 37–80). Cambridge: Cambridge University Press.

directs us to different candidate concepts and the issue in philosophy, essentially normative in character, is to decide which concept the term should be taken to express. While the approach uses the word 'concept' where this chapter would speak of a concept-referent pairing, the difference is not significant and the line taken in this chapter offers support for conceptual ethics.

Frankfurt, H. 1969. Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66:829–39.

Gardner, J. 2013. Why Law Might Emerge: Hart's Problematic Fable. In D. d'Almeida, J. Edwards, and A. Dolcettit (eds.), *Readubg HLA Hart's The Concept of Law* (pp. 1–23). Oxford, Oxford University Press.

Goodman, N. 1983. *Fact, Fiction and Forecast.* Cambridge, MA: Harvard University Press.

Gourevitch, A. 2014. *From Slavery to the Cooperative Commonwealth: Labor and Republican Liberty in the Nineteenth Century*. Cambridge: Cambridge University Press.

Grice, H. P. 1975. Method in Philosophical Psychology. *Proceedings and Addresses of the American Philosophical Association* 68:23–53.

Habermas, J. 1971. *Knowledge and Human Interests*. Boston, Mass.: Beacon Press.

Halldenius, L. 2001. *Liberty Revisited*. Lund: Bokbox Publications.

Hart, H. L. A. 1961. *The Concept of Law*. Oxford: Oxford Unviersity Press.

Haslanger, S. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Hobbes, T. 1994a. *Human Nature and De Corpore Politico: The Elements of Law, Natural and Politic*. Oxford: Oxford University Press.

Hobbes, T. 1994b. *Leviathan*, ed. E. Curley. Indianapolis: Hackett.

Honohan, I. 2002. *Civic Republicanism*. London: Routledge.

Honohan, I., and Jennings, J. (eds.) 2006. *Republicanism in Theory and Practice*. London: Routledge.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jackson, F. 2004. What are Proper Names For? Experience and Analysis. In M. E. Reicher and J. C. Markek (eds.), *Kirchberg Proceedings* (pp. 257–69). Vienna: Kirchberg.

Jackson, F., and Pettit, P. 1990. In Defence of Folk Psychology. *Philosophical Studies* 57:7–30. Reprinted in F. Jackson, P. Pettit, and M. Smith. 2004. *Mind, Morality and Explanation*. Oxford: Oxford University Press.

Jackson, F., and Pettit, P. 1995. Moral Functionalism and Moral Motivation. *Philosophical Quarterly* 45:20–40. Reprinted in F. Jackson, P. Pettit, and M. Smith. 2004. *Mind, Morality and Explanation*. Oxford: Oxford University Press.

Johnston, M. 1992. How to Speak of the Colors. *Philosophical Studies* 68:221–63.

Johnston, M., and Leslie, S.-J. 2012. Concepts, Analysis, Generics, and the Canberra Plan. *Philosophical Perspectives* 26:113–71.

Kripke, S. A. 1980. *Naming and Necessity*. Oxford: Blackwell.

Kymlicka, W. 2002. *Contemporary Political Philosophy*. Oxford: Oxford University Press.

Laborde, C., and Maynor, J. (eds.) 2007. *Republicanism and Political Theory*. Oxford: Blackwell.

Lewis, D. 1969. *Convention*. Cambridge, Mass.: Harvard University Press.

Lewis, D. 1983a. Philosophical Papers Vol 1. Oxford: Oxford University Press.

Lewis, D. 1983b. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61:343–77.

List, C., and Pettit, P. 2011. *Group Agency: The Possibility, Design and Status of Corporate Agents*. Oxford: Oxford University Press.

Lovett, F. 2010. *Justice as Non-domination*. Oxford: Oxford University Press.

Lovett, F., and Pettit, P. 2009. Neo-Republicanism: A Normative and Institutional Research Program. *Annual Review of Political Science* 12:18–29.

MacGilvray, E. 2011. *The Invention of Market Freedom*. Cambridge: Cambridge University Press.

Marti, J. L., and Pettit, P. 2010. *A Political Philosophy in Public Life: Civic Republicanism in Zapatero's Spain*. Princeton: Princeton University Press.

Maynor, J. 2003. *Republicanism in the Modern World*. Cambridge: Polity Press.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Niederbeger, A., and Schink, P. (eds.) 2012. *Republican Democracy: Liberty, Law and Politics*. Edinburgh: Edinburgh University Press.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell.

Olsaretti, S. 2004. *Liberty, Desert and the Market*. Cambridge: Cambridge University Press.

Pettit, P. 1993. *The Common Mind: An Essay on Psychology, Society and Politics*. New York: Oxford University Press. Paperback edn, 1996.

Pettit, P. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.

Pettit, P. 1998. Practical Belief and Philosophical Theory. *Australasian Journal of Philosophy* 76:15–33.

Pettit, P. 2002. *Rules, Reasons, and Norms: Selected Essays*. Oxford: Oxford University Press.

Pettit, P. 2008. *Made with Words: Hobbes on Language, Mind and Politics*. Princeton: Princeton University Press.

Pettit, P. 2011. The Instability of Freedom as Non-Interference: The Case of Isaiah Berlin. *Ethics* 121:693–716.

Pettit, P. 2012a. Freedom in Hobbes's Ontology and Semantics: A Comment on Quentin Skinner. *Journal of the History of Ideas* 73 (1):111–26.

Pettit, P. 2012b. *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge: Cambridge University Press.

Pettit, P. 2014. *Just Freedom: A Moral Compass for a Complex World*. New York: W.W. Norton and Co.

Pettit, P. 2018a. *The Birth of Ethics: Reconstructing the Role and Nature of Morality*. Oxford: Oxford University Press.

Pettit, P. 2018b. Three Mistakes about Doing Good (and Bad). *Journal of Applied Philosophy* 35 (1):41–46.

Plunkett, David, and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Railton, P. 2014 Reliance, Trust, and Belief. *Inquiry* 57:122–50.

Rawls, J. 1971. *A Theory of Justice*. Oxford: Oxford University Press.

Rosen, G. 2010. Metaphysical Dependence: Grounding and Reduction. In B. Hale and A. Hoffman (eds.), *Modality: Metaphysics, Logic, and Epistemology* (pp. 109–35). New York: Oxford University Press.

Rosen, G. 2014. *Normative Necessity*. Princeton: Princeton University.

Schroeter, L., and Schroeter, F. 2014. Normative Concepts: A Connectedness Model. *Philosopher's Imprint* 14:1–25.

Schwartz, W. 2014. Against Magnetism. *Australasian Journal of Philosophy* 92 (1):17–36.

Sellars, W. 1997. *Empiricism and the Philosophy of Mind*. Cambridge, Mass.: Harvard University Press.

Skinner, Q. 1998. *Liberty before Liberalism*. Cambridge: Cambridge University Press.

Skinner, Q. 2008. *Hobbes and Republican Liberty*. Cambridge: Cambridge University Press.

Stalnaker, R. 1978. Assertion. In P. Cole (ed.), *Syntax and Semantics*, vol. 9 (pp. 315–23). New York: Academic Press.

Stoljar, D. 2017. *Philosophical Progress: In Defence of a Reasonable Optimism*. Oxford: Oxford University Press.

Strawson, P. 1962. *Freedom and Resentment and Other Essays*. London: Methuen.

Sundell, T. 2012. Disagreement, Error, and an Alternative to Reference Magnetism. *Australasian Journal of Philosophy* 90:743–59.

Taylor, R. S. 2017. *Exit Left: Markets and Mobility in Republican Thought*. Oxford: Oxford University Press.

Thomas, A. 2017. *Republic of Equals: Presdistribution and Property-owning Democracy*. Oxford: Oxford University Press.

Van Gelderen, M., and Skinner, Q. 2002. *Republicanism: A Shared European Heritage*, 2 vols. Cambridge: Cambridge University Press.

Viroli, M. 2002. *Republicanism*. New York: Hill and Wang.

Weinstock, D., and Nadeau, C. (eds.) 2004. *Republicanism: History, Theory and Practice*. London: Frank Cass.

Williams, B. 2002. *Truth and Truthfulness*. Princeton: Princeton University Press.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

17

The A-project and the B-project

Mark Richard

1

In a number of papers, Sally Haslanger has pursued what she calls an *ameliorative project*. This is a kind of conceptual analysis in which you begin by looking at the purposes behind the use of a concept—you ask what people are actually doing when they apply the concept and why they are doing it—and then go on to evaluate those purposes: you ask if those purposes are ones we ought have, or if there are different ones that should be assigned to the concept. One then considers whether some modification of the concept is called for, given the purposes it ought to have, and, if so, what they are. If the analyst thinks new purposes ought to be assigned to the concept, she may well propose a revisionary account of the concept, one on which it is suited for the purposes it ought to have.[1]

An ameliorative account of a concept is potentially quite different from other sorts of accounts one might give. It is not an attempt at conventional conceptual analysis, in which one "seek[s] an articulation" of a concept (employing, perhaps, a method of reflective equilibrium to arrive at an all things considered definition)—it is not, to

[1] Haslanger writes:

> The task is not simply to explicate the normal concept of X; nor is it to discover what things we normally take to fall under the concept we have in common; instead we ask what purpose is served in having the concept of X, whether this purpose is well-conceived and what concept (or concepts) would serve our well conceived purpose(s) . . . best. . . . this approach [to conceptual analysis] is quite comfortable with the result that that we must revise—perhaps even radically—our ordinary concepts. . . . (Haslanger (1999): 352)
>
> . . . we begin by considering more fully the pragmatics of our talk employing the terms in question. What is the point of having these concepts? What cognitive or practical task do they (or should they) enable us to accomplish? . . . In the limit case [of the project] the concept in question is introduced by stipulating the meaning of a new term. . . . But if we allow that our everyday vocabularies serve both cognitive and practical purposes, purposes that might also be served by our theorizing, then a theory offering an improved understanding of our (legitimate) purposes and/or improved conceptual resources for the tasks at hand might reasonably represent itself as providing a (possibly) revisionary account of the everyday concepts [that are the subject of ameliorative analysis]. (Haslanger 2000: 223–4)

(Here and below, all references to Haslanger are to the reprintings of Haslanger's work in Haslanger (2012).) I should note that in Haslanger (2000), she uses the phrase 'analytic account of a concept' instead of 'ameliorative analysis'.

Mark Richard, *The A-project and the B-project* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Mark Richard.
DOI: 10.1093/oso/9780198801856.003.0017

borrow Haslanger's terminology, an attempt to uncover the *manifest concept* associated with a term. Neither is it the quasi-empirical (descriptive, as Haslanger calls it) project of looking for a natural, physical, or social kind that (is the most plaussible candidate for what) our applications of the concept are in fact tracking (Haslanger 2000: 223). Nor is it the attempt to limn the set of things that it is our practice to apply the concept to, what Haslanger calls the *operative concept*.[2] But while different, these sorts of accounts of concepts as well as ameliorative accounts all seem to involve reference and extension. Conventional conceptual analysis seeks to identify, via conceptual articulation, what we on our best reflection take ourselves to be talking about when we apply a concept. A descriptive account identifies what (kind of) objects our practice 'tracks', and thus what (kind of) objects we should take the concept to contribute to truth conditions; an account of an operative concept computes a concept's "practical extension". And an ameliorative account aims, put roughly, to tell us what objects we *should* (given our "proper purposes") be talking about when we use the concept.

Haslanger herself offers 'ameliorative analyses' of the concepts of woman, man, and (roughly speaking) racial group. She begins by asking "what work the concepts of gender and race might do for us in a critical...social theory" (Haslanger 2000: 36–7). Her answer, put generally, is that what is needed are "accounts of gender and race that will be effective tools in the fight against injustice". As I understand her, in the case of the concept *woman* she reasons as follows. The, or an important, purpose of the concept *woman* is to subordinate people on the basis of their (perceived) female properties. We shouldn't be subordinating people on this basis; indeed, we should be fighting against such subordination. One way to do this is to reformulate the concept so that, so to speak, its noxious purpose is part of its definition. This will put the purposes for which the concept is actually being used front and center, allowing us to fight gender subordination. We should therefore understand what it is to be a woman as being someone who is systematically subordinated on the basis of (perceived) female properties. This makes the concept *woman* determine a response dependent property: women are people who are perceived in a particular way, and as a result are treated in a particular way. Since the behavioral response arises only in certain kinds of societies, on this account women exist only in societies whose ideology marks certain groups for oppression. The concept *man* is taken to be analogous, though men are privileged, not subordinated, on the basis of their perceived properties. Of course this analysis is strikingly out of synch with the accounts most philosophers and non-philosophers would give of the concepts.[3]

[2] The distinctions and terminology here are something of a mash up of the discussions in Haslanger (1999, 2000, and 2006).

A descriptive account of a concept is not an operative account of the concept—we might be clearly be tracking a natural kind with a word but widen or narrow its application due to theoretical confusion. That said, for simplicity, in what follows I ignore whatever differences there might be in a descriptive account of a concept and an attempt to find an operative concept.

[3] The text oversimplifies Haslanger's proposal. Her (penultimate) definition is

> S is a woman iff i) S is regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction; ii) that

I will call a project of offering and trying to get others to accept revisionary ameliorative accounts of a concept an *A-project*. An A-project is something that is carried out in a particular social situation at a particular time. It is undertaken as a reaction to a particular social and historical situation as conceptualized in a particular way with a particular vocabulary. It is focused on a particular concept C, expressed by particular vocabulary W, and the way that C functions in its historical milieu. It will be successfully carried off only if a large number of those who think about the world using C, expressing those thoughts with word W, come to do something we might call 'changing their concept C' in ways that reflect the revisionary analysis while continuing to use W to express their (revised) concept C. It will be successful only if these uses of W are accompanied by the intention that they should be understood as having the relevant meaning, and are indeed so understood.[4]

It is obvious, I think, that the A-project is often worthwhile. But there are cases—Haslanger's own version of the A-project is one—in which it can at least *seem* downright odd. Her project, after all, would be successfully carried off only if a large number of those who theorize about gender and race were to come to use 'woman', 'Latino', and so on with the conscious intention that they should be understood as using the words with the relevant meanings, and indeed are so understood.[5] But why should we want to pull *this* off, as opposed to the seemingly simpler task—the B-project—of getting theorists to agree that most members of the relevant classes—females, those of Hispanic descent—are indeed systematically subordinated on the basis of being members of those classes?

The answer I take it is that our (legitimate) purposes for having concepts of gender and race are much better served by pulling off the A-project than the B-project. These purposes, I take it, are ones like trying to get rid of the subordination of females and Latinos by theorizing about it in a fruitful way, a way that "cuts at the social joints"—that is, that displays social kinds and social forces that explain why the social world is as it is.

But this answer invites the question, Why should we think that these purposes are better served by the A-project than the B-project? A reason one *might* have for thinking this is that the notion *group identified in terms of marks M and subordinated on the basis of having M* is of considerable explanatory utility: unless and until you

> S has these features marks S within the dominant ideology of S's society as someone who ought to occupy certain kinds of social position that are in fact subordinate (and so motivates and justifies S's occupying such a position); and iii) the fact that S satisfies (i) and (ii) plays a role in S's systematic subordination, i.e., along some dimension, S's social position is oppressive, and S's satisfying (i) and (ii) plays a role in that dimension of subordination.

And this requires Chisholming in light of the fact that one's real or imagined traits may trigger, in different contexts and countries, quite different patterns of ideologically inspired behavior.

[4] Actually, this is not quite right as it stands. An A-project will be directed at a particular audience, and that audience may be considerably smaller than the set of all those (in the relevant historical milieu) who use C, expressing it with W. So the text needs to be restricted: success requires that a large number of those at whom the project is directed who think about the world using C, etc., etc.

[5] The concept *Latino* is an ethnic and not racial term. It's clear that Haslanger's intent is that her proposal generalize to such concepts.

see the relevant groups in that way, you will be unable to perceive or explain various social facts. I agree with the thought, but it doesn't seem to be a reason for preferring the A-project to the B-project. Why think that we are more likely to get people to recognize and be able to explain subordination by conceptual reformation than by . . . getting them to recognize social structures for what they are? Doesn't the A-project require fighting two battles—overcoming resistance to recognizing such things as implicit bias *and* overcoming resistance to using a word in a counter-intuitive way? Wouldn't the project's goals be achieved by fighting and winning just the first battle? For that matter, doesn't the A-project as Haslanger executes it cross the line between conceptual therapy and stipulative rebranding? Isn't the fact that people perceive it in just this way a reason to think that in *this* case we are better off pursuing a B-project instead of an A-project?

One might even argue that Haslanger's ameliorative project borders on incoherence. What I'm calling the A-project is, or at least often involves, a project of offering (certain) 'revisionary' accounts of a concept or of the meaning of a word. There are several ways in which an account of Xs might be revisionary: it might significantly depart from the manifest concept of Xs, from the operative concept of Xs, or from both. Cases in which one holds on to one of the two and brings the other in line with it are of course worthwhile—here theory and practice are brought into harmony. But the more revisionary cases—like Haslanger's own example of an A-project—where the proffered account is pretty strongly at variance with both manifest and practical concept might well be thought to be incoherent. For you might argue so: The semantics of a concept term H—its ex- and in-tension—is surely more or less determined by the manifest and operative concepts associated with it. But in a revisionary version of the A-project one expects that there will be substantial changes in both of these and thus substantial changes in the concept's semantics. But surely the referential semantics of a concept determines its identity—substantial changes in semantics mean abandoning the concept. So revisionary versions of the A-project look to be verging on incoherence.

Someone who so objects thinks of concepts and meanings as things with semantic essences: a concept or meaning, even if it lacks a definition, is essentially a concept with a particular referential semantics, perhaps more or less natural, perhaps more or less gerrymandered. As such, it owns an intension, a shadow across possible worlds. Or better, the intension owns it: change the intension associated with a word and you change the concept associated with it. Haslanger has no patience for this way of thinking of meanings and concepts and neither do I. I agree with her that a concept is something with a history; it is to a certain extent misleading to speak of *the* practical or manifest concept of water, Latinos, or women, for we want to be able to speak a single concept changing over time so that its articulation (which reflects the manifest concept at a particular time) or its *practicum* (reflecting the concept's application) may change while the concept persists.

The revisionary A-project makes no sense at all unless we think of concepts in this way. So it would be good to have some handle on what such a picture of concepts might look like. The next two sections sketch such a picture. Section 4 returns to the question of the prospects for success for A-projects and other attempts at conceptual engineering.

2

Word meanings are a kind of concept, and it is on this sort of concept that I propose to focus. Notoriously, some—internalists—take word meanings to be by and large idiosyncratic, a reflection of the individual's individual cognitive history, while others take such meanings to be part of a common store. For the moment, let us not quibble with the internalist: each speaker speaks her own language. Still, when speakers are in actual and potential communication there is typically an enormous amount of similarity in their languages. Communication and language acquisition conspire to insure this: enough evidence that others use a word in ways different than I tends *ceteris paribus* to reproduce the others' usage in my idiolect.

Speakers in actual and potential communication have similar idiolects; they tend to associate very similar inferential roles and presuppositions with their words. Of course for all the similarity there is diversity, not only in presuppositions but in phonology, morphology, syntax, and so on. Linguistic individuals interact, and these interactions produce changes in the idiolects of the interactors, with many of these changes being (more or less) permanent. The changes—that is, the acquired properties—can be and often are transmitted to others. Some such changes may spread aggressively across a population; others fizzle or disappear; yet others (think of slang) persist in a minority equilibrium. Over time changes in a population's linguistic behavior may lead to words' bearing meanings quite different from those they had originally.

To me—and, I hasten to say, to many linguists, as my analogy is hardly novel—it is striking how much this resembles the biological world. There we find populations of individuals who are very similar—they have similar genomes. The members of a population interact with one another, with the interactions resulting in individuals who tend to resemble the interactors. Over time individuals who make up a population lineage may, as changes in transmitted properties become fixed in the population, become so different from their ancestors that we say they are of a different species. I think we gain a certain amount of illumination if we think of linguistic entities—in particular word meanings and the concepts words express—as being like those segments of population lineages that we label species.[6]

If meanings are species-like, what exactly are they? What, for example, is the meaning of 'cousin'?

When a speaker speaks, she makes presuppositions that she expects her audience will recognize she is making, ones she expects the audience will have ready for use in making sense of what she says. Some such presuppositions are tied to particular words and accompany their use. When we speak of cousins using 'cousin', we expect to be recognized as talking about parents' siblings' progeny; we expect the audience can access this idea in interpretation. For some such presuppositions, it will be common ground in a linguistic community that speakers make them and expect that to be recognized by their audience. I call these sorts of presuppositions

[6] I argue for such a view of concepts and meanings in Richard (2019). This section and the next are a sort of *Readers' Digest* synopsis of sections of chapters 1 and 3; section 4 borrows some paragraphs from chapter 6.

interpretive common ground, ICG for short. I say the meaning of lexical items is, to a first approximation, interpretive common ground in the sense just limned.

You should say: What do you mean by 'meaning'? I could mean something like the determinant of reference and truth conditions—something like Kaplanian character. Or I could mean something that determines what proposition is expressed by a sentence's use. Or I could mean meaning in the sense of that with which one must be in cognitive contact in order to qualify as a competent speaker in a population.

I mean the last. ICG is relevant to reference and truth, but reference and truth can't be read off it, if only because what's common ground is often erroneous. There is, I think, a sense of 'what is said' in which what is said is determined by the ICG of the phrases in a sentence and referential semantic values; there are other senses of 'what is said' in which ICG does not determine what is said. ICG is meaning as the anchor of linguistic competence; it is what knits us together as beings who share a language and thus can communicate.

A phrase has an interpretive common ground in a particular population at a particular time; its ICG consists of those presuppositions that (it is common ground in the population at that time that) the phrase's users make and expect their audience to recognize as made and to use in understanding the phrase's use. But what is a presupposition? What is it for a presupposition to be in a population's common ground?

I intend something like Bob Stalnaker's notion of presupposition. One makes a presupposition for certain purposes; to presuppose that p for purposes R is, roughly, to be disposed, in situations in which R is among one's purposes, to behave as would a person whose R-relevant behavior was in part determined by her belief that p. To have as a conversational presupposition

Q: 'Cousin' is used to talk about parents' siblings' progeny

is in the first instance to be disposed to behave for the purposes of conversation like someone whose conversational behavior—in particular, her use of the word 'cousin'—is determined in part by belief in Q.[7]

Whether a presupposition is common ground in a population is a matter of what is the norm so far as assumptions go. Q and

P: Users of 'cousin' presuppose Q

are common ground in the community of adult speakers of English in greater Boston. Does this mean that *every* competent speaker of English in that community presupposes Q and P? Of course not. There might be a few "cousin fanatics" among Bostonian English speakers—people who think only males can be cousins and are on

[7] Thus, the notion of presupposition I'm working with is one that we might say is *quasi-cognitive*. To presuppose p for purpose X is to be disposed, when one has that purpose, to behave in ways that believers of p behave. One way to have such a presupposition is to explicitly believe p. But cognitively unsophisticated things like birds, baboons, and babies can be disposed to behave as if they believed p without having anything at all like an explicit belief that p. I think this is a virtue of the account I'm giving. It allows for meaning in cognitively unsophisticated populations, and it allows that some or even much of what we mean may be opaque to us, as we can be disposed to behave as if we thought p without being able to recognize that we are.

a tear about it, challenging anyone who has the temerity to suggest that there are female cousins. These wingnuts don't presuppose Q, but that doesn't mean that P and Q aren't common ground in the population they help constitute. To say that the presupposition that p is common ground in group G is *roughly* to say that

[G] Gs presuppose p
Gs presuppose that Gs presuppose p
Gs presuppose that Gs presuppose that Gs presuppose p

and so on up are all true, where the claims in [G] are generics: on a first pass, 'Gs presuppose p' says something along the lines of *all normal Gs presuppose p*; 'Gs presuppose that Gs presuppose p' says something to the effect that all normal Gs presuppose the first generic claim.[8]

Explaining normality is not something I propose to do today, but here is a crumb of elucidation. I assume that for normal As to be Bs is not a matter of statistics, but of there being some explanatory connection between being an A, the situations As (normally) find themselves in, and tending to be a B.[9] Put crudely, the generic claim is that there are mechanisms M (that are normal for As and their situation) which in propitious circumstances tend to lead to As being Bs. If we think of generics and common ground in this way, then to say that Q is common ground amongst adult Bostonians does not imply that every Bostonian presupposes it. Rather, it implies that there is a mechanism that in the normal course of things brings Bostonians to presuppose it, to presuppose that Bostonians presuppose it, and so on. And of course there *is* such a mechanism—it's the Boston public school system.

I trust it is clear that if meaning is something like ICG, then it is indeed something species like. The presuppositions that accompany a word's use in a population are relatively stable but they do change over time—sometimes such change is glacial, sometimes it is saltative. At the level of the individual speaker, presuppositions about how others use a word are very much like the genomes of the members of the species—there is typically a certain amount of allelic variation across speakers in the assumptions they make about how people expect a word to be understood. It is, of course, a large and difficult question as to how much change in a word's ICG percipitates a change in meaning on the order of a demise of an old concept and the birth of a new one. But it should be clear that on reasonable criteria of meaning persistence, change in what constitutes ICG is no more likely to drive a meaning out existence than changes in the distribution of a gene's alleles in a population trigger the death of a species. This is so, for example, if we say that a change in ICG marks the death of one meaning and the emergence of another only when those changes (significantly) impede fluid conversation or the role of term in inquiry.[10]

[8] P's being common ground in G is something like "it's being out in the open in G that Gs accept p". As is common, the text glosses this idea in terms of indefinitely iterated attitudes. I am not sure that this is the best way to understand the notion. (I'm not sure it's not, either.) So officially I'm taking the notion common ground as a primitive.

[9] I here help myself to the rudiments of the views of my colleague Bernhard Nickel. For a development of the view of both the semantics of generics and of their connection with explanation that I am presupposing here, see Nickel (2016).

[10] This issue is discussed at length in chapter 4 of Richard (2019).

It should also be clear that this is a picture of meanings and concepts that potentially makes sense of the idea that a concept can persist through changes in what it is a concept of. Reference is determined by a host of things—speaker presuppositions, to be sure, but also environmental relations, relations of deference, a term's role in theory. The latter three reference determinants can change—indeed, they can change quite a lot—while the common ground surrounding a term is more or less stable. Concept persistence doesn't require persistence of reference.[11]

3

Species are a population level phenomenon, one defined in good part by synchronic and diachronic relations among members of population lineages. Species undergo evolutionary change without going out of existence. The distribution of a gene's alleles within a population can change dramatically over generations, so that an allele—one that controls, say, beak size—that begins as a (statistical) outlier drives to fixation because of some advantage it gives to the possessor. But bigger beaks do not new species make.

To say that meanings are species-like is to suggest that they too are a population level phenomenon. It is to suggest that we should think of them as diachronic ensembles of individuals—ensembles of lexical entries of speakers in actual and potential communication—which undergo diachronic changes that resemble the changes in allele distribution just mentioned. Different speakers have different standing presuppositions which they expect their audience to recognize as being made when a word is used; such differences make a word's meaning in a population enjoy something like allelic variation. And meanings can be expected to undergo changes that look like the change a species undergoes when one allele of a gene is challenged for whatever reason by an upstart allele.

It is worth discussing an example. I'll focus on feminist attempts in nineteenth-century America to change the way that both law and society thought of rape. For historical details, I draw from an article by Jill Hasaday (2000).

The nineteenth-century American legal understanding of rape—and surely its dominant social understanding for at least the first two-thirds of the century—was that it occurred when a man had intercourse with a woman who was not his wife without her consent.[12] The nineteenth-century feminist movement in America began arguing publically in favor of the idea that a woman 'has a right to her own person'—and thus has a right to refuse her husband's demands for sex—in the mid 1850's; the radical 'Free Love' movement at about this time went so far as to apply the terms 'rape' and 'sexual slavery' to much of what happened in the marital bed (Hasaday

[11] ICG is meaning *cum* anchor of competence, that with which a speaker needs to be in cognitive contact in order to be competent. A complete discussion would at this point take up the question, What sort of relation must one bear to ICG in order to be competent? I won't do that here, save to observe that (a) ICG is a collection of 'first order' presuppositions (e.g., that 'cousin' refers to parents' siblings' progeny) and 'higher order' ones (e.g., that speakers presuppose that 'cousin' so refers); and (b) one can be competent in virtue of being disposed, in interpreting speech, to make both first and higher order presuppositions, or just the first order ones, or just the higher order ones. For discussion, see chapter 3 of Richard (2019).

[12] Throughout what follows I simplify the presuppositions that articulate concepts.

2000: 1415ff). These ideas were pressed fairly vigorously for the balance of the nineteenth century, but had little effect beyond making sexual abuse in some states a ground for divorce. That the idea that women have a right to their bodies was a cornerstone of the feminist movement in the U.S. in the 1800's seems to have been pretty much forgotten for much of the twentieth century.

The legal and cultural surround of the official definition of rape here is of some interest. The idea that a man literally could not rape his wife was tied to two things. The first was a legal view of marriage as involving a status that was permanent and non-negotiable. The state set certain parameters for the rights and duties of marriage partners that were not optional. Hasaday quotes the author of "one of the most influential family law treatises" of the time as writing

> [T]he idea, that any government could, consistently with the general weal, permit this institution to become merely [a] matter of bargain between men and women and not regulate it by its own power is . . . too absurd to require a word of refutation.[13]

Secondly, assumptions about marital rape depended on the idea that in agreeing to marry one gave irrevocable consent to having sex whenever one's partner requested it—an idea that, as is well known, traces back to the British jurist Matthew Hale, who wrote that marital rape was impossible "for by their mutual matrimonial consent and contract the wife hath given herself up in this kind [i.e., sexually] unto her husband, which she cannot retract".[14] If you don't question the bizarre view of consent in Hale's doctrine, the idea that marital rape is impossible might well seem correct.

At the beginning of the nineteenth century, there was an *official notion*—ON, call it—of rape, the notion of the moral and legal infraction of sexual intercourse forced upon a woman without her consent by someone other than her husband. Everyone would have recognized that ON was just that: it was the notion of rape that (it was commonly known) played a certain role in law and society so that for legal and most social purposes all and only those things described by its articulation counted as rape. By the mid 1860's there was a *competitor notion* CN, that of sexual intercourse forced upon a woman by anyone without her consent. Though ON continued throughout the nineteenth and much of the twentieth century to be the "official" notion of rape, it seems that by about 1875 it was commonly known—at least among the well educated who paid attention to these things—that ON had in some sense acquired a competitor—there were people who were pushing to have CN play the role of "the official notion" of rape.

What did the word 'rape' as used by tolerably educated adults in the U.S. in 1875 mean? The question could be construed in several ways. Thinking that reference is determined by meaning, one might take it as a question whose answer needs tell us something about reference and truth conditions. Thinking that "what is strictly and literally said" is determined by meaning, one might take it as a question whose answer needs tell us something about "the proposition expressed by a sentence in

[13] Joel Bishop in *Commentaries on the Law of Marriage and Divorce 11* (Little Brown, 1864); cited at Hasaday (2000: 1387).

[14] Matthew Hale, *The History of the Pleas of the Crown* (1736), quoted at Hasaday (2000: 1397).

which the word 'rape' is used". Thinking that meaning is a handy substantive for whatever it is whose grasp makes one a competent speaker, one might want the answer to the question to tell us what it was about the relations of the minds of adults who understood 'rape' to the linguistic and social world that constituted their competence. I take the question in this last sense.

In trying to answer it, I will make some assumptions. In line with the view that I have been developing, I assume that facts about meaning are in a fairly strong sense determined by the presuppositions that speakers make in speech, in particular by those they expect their auditors to bring to the task of interpretation. Many of these presuppositions will be part of what I've been calling interpretive common ground—they are the norm, so far as what speakers presuppose in speaking and their normality is common knowledge. But not all such assumptions need be the norm in this way. In particular, some of what nineteenth-century feminists presupposed about rape were not things that most people presupposed, and feminists would have been painfully aware of this fact. What feminists could and (I think) did assume was that educated speakers were aware that a fair number of people used 'rape' presupposing that all forced intercourse is rape; they could and did presuppose that when they spoke publically of rape, their audience would recognize that they were making this presupposition. Certainly speech in which the activities in the marital bed were called rape would be accompanied by this presupposition—if the audience didn't recognize that it was expected to recognize the presupposition, what the feminist said would appear more or less unintelligible.

I think there are five likely answers to our question, What did the word 'rape' mean in 1875? (a) It meant what it meant in 1800, ON. (b) It meant CN, or something very much like CN, which in turn is close to the modern meaning of the word. (c) It was ambiguous: in some mouths it expressed something like ON, in others, one or another variant of CN. (d) At least as it was used by the educated, it had a meaning that in a sense (which I will explain presently) combined both ON and CN. (e) The question is misconceived: in times in which a word threatens to undergo meaning change, it will often be impossible to come up with an account of what it means. According to this last answer in the case at hand, there was a distribution of variant uses of 'rape' in the population, much as there was a distribution of alleles that, in the normal course of things, led to brown or blue or green or grey eyes. In such a case, there is no more such a thing as the meaning of the word in the population than there is such a thing as the eye color of the population.

Answer (d) needs to be clarified. To do this, we need the notion of *competing ways of using a word* in a population. Speakers often associate several meanings with a word *cum* morpho-phonetic-syntactic object; homophony and polysemy are examples. Even abstracting from population-wide homophony and polysemy, for many words, speakers' lexicons often contain information that encodes distinct common ways of using a term in a population. If you've spent time in Alberta, you interpret the natives' use of 'toboggan' differently from American uses of the word. Someone who travels between Boston and New York and likes soup knows that what the speaker expects if she asks for clam chowder in Boston is very different from what the speaker expects if she asks for it in New York. Someone who knows me and knows my daughter knows that I assume that a concoction involving vodka and various liqueurs is not and could not be a 'martini'; not so my daughter.[15]

There are different ways of using 'toboggan' and 'martini'. Ways of using words—Ways, for short—populate a social and linguistic landscape constituted in part by the mentalities of individual users, in part by the roles Ways play in legal, religious, and other social structures. Ways realized by a particular individual's use of a word carry such things as inferential connections, behavioral dispositions cued to beliefs, and particular presuppositions. Social structures are more or less enduring ways people behave, ones involving standardized patterns of behavior in particular situations, expectations about such behavior, behavioral norms, and publically recognized ways of labelling such behaviors and norms. Information about such structures is typically carried by the (token) Ways used to describe and think about them.

A token Way—the understanding Vice President Schuyler Colfax or the feminist Elizabeth Cady Stanton voiced in 1870 with 'rape', for example—has a particular role, one determined by its properties in an individual's lexicon, and the connections it has to the social structures it describes and is used to think about. The types such tokens realize may come to 'compete' to occupy a particular niche. In the case at hand, the competition is (idealizing only a bit) one between ON and CN to occupy the niche in the population of English speakers defined by (a) the functional and inferential roles associated by users with the word; and (b) the legal and social role that the notion of rape had in the society. And this really is a competition: One can't, in using 'rape', presuppose both that marital intercourse is never rape and that it sometimes is; legally, forced intercourse in marriage can't be and not be rape.

That said, there is a way that the understanding one assigns to the word 'rape' can be constituted by both ON and CN, a way not altogether unlike the way a heterozygote combines two variants of a particular gene. It's common for someone to associate several overlapping sets of presuppositions with a word; 'toboggan', 'clam chowder', and 'martini' are examples. This is not like standard cases of homophony or polysemy. If Jane says "I went to the bank yesterday to make a deposit" and Jim says "I went to the bank yesterday to collect reeds", one cannot say things like "Jane and Jim each went to a bank yesterday" or "Jane and Jim both said that they went to a bank yesterday". If Jim says 'Let's smoke a blunt' and Jane says 'Let's smoke the salmon', one cannot say 'Jim and Jane each want to smoke something'. But if Jane says 'May I have a vodka, Chambord, and pineapple martini' and Jim says 'Gimme a dirty Hendricks martini', I can (and, to expedite communication, will) say 'Jane and Jim each asked for a martini' or 'Jane and Jim both had a martini', even though I reject Jane's presuppositions about martinis. I take it this indicates that, however reluctantly in the case of Jane, I interpret the Ways that lie behind Jane and Jim's utterances with the Way that I use 'martini' when I speak of martinis, even though I am aware of a difference in the meaning *cum* endorsed presuppositions that Jane and I assign to the word 'martini'.[16] Something similar will be true of Jane's interpretation of me if she knows my feelings about what counts as a martini: If

[15] I got the example of 'martini' from Ted Sider, who tells me that he got it from Karen Bennett.

[16] Whether I will say Jane had a martini actually depends in part on my audience.

I ask for a martini, she will interpret me as having asked for the same sort of drink as she did—she'll think I asked for a martini—though she knows that there is a difference in meaning *cum* endorsed presuppositions between us. All this suggests that Jane's and my uses of 'martini' stand in a relation of 'interpretive coordination': though we knowingly differ in the presuppositions we endorse, we discount that difference in communication, each interpreting the other's use of 'martini' with her own.[17]

In this situation there are two Ways of using 'martini', Ways defined by sets of presuppositions made by a user and expected by her to be recognized as being made. Furthermore, it is common knowledge that there are these two Ways of using the word: those who use it in one way know that others use it in the other. And these Ways stand in a curious relation. On the one hand, embedded in a single individual they are in some sense inconsistent; they are in some sense distinct concepts whose presuppositions aren't consistent. One can't (without hypocrisy, at least) adopt a policy of using the word in full voice sometimes in one Way, sometimes in another.[18] On the other hand, the two Ways of using the term are treated socially as if they expressed the same concept. My daughter and I interpret each other as speaking about "the same thing" with 'martini', as evidenced, for example, by our practices of interpreting and reporting each other's thinkings and sayings.

The structure of the 'rape' example is much the same as that of 'martini' example. A feminist and a conservative in the 1880s who recognized that they differed as to whether marital rape was possible would *of course* interpret uses of 'rape' in a way that indicated their lexical entries for the word were interpretively coordinated with one another and with those of other members of the society: They would report one another as disagreeing about rape.

We can now clarify the idea that 'rape' in late nineteenth-century America had a meaning that in some sense combined ON and CN. In the population we are discussing, Ways of using the word to pick out sexual violation were interpretively coordinated. This network of coordinated Ways was associated with a particular sort of individual functional role (everyone, for example, expected all to take rape to be a serious crime) and a particular social role (it was a felony). We call such a network of coordinated Ways with a tolerably well-defined individual and social role a *p-* (for public) *word* in the population. Note that a p-word need not be used with the same first order presuppositions by those who share it: Jane assumes a martini can be made from vanilla vodka; I don't. And that this is so may be common knowledge within a population: People in 2005 knew that there was a Way of using 'marriage' on which the user presupposed that marriage between those of the same sex is impossible (and

[17] There is another difference between the martini example and standard examples of homophony and polysemy. There is a sort of tension between Jane's and my uses of 'martini'—I can't endorse both uses in full voice. But I feel no tension in my uses of 'smoke' with its various transitive meanings—as something one does to a cigarette or a salmon filet or to another person. Those meanings do not strike me as being—they are not—in the kind of competition that differing uses of 'martini' are. Likewise for 'bank'.

[18] This needs qualification. I could, of course, adopt a policy of using the word in different ways as the way it is used by my audience shifts. The point in the text is that there is something very odd about a person who willfully (and without fairly elaborate signaling) uses the word now one way and several sentences down the conversation the other.

expected their audience to recognize this), and another Way of using it on which the user presupposed that same sex marriage is possible (and expected their audience to recognize this).

When it is common knowledge among users in the network that constitutes a p-word w that a particular Way S accompanies (some) uses of w, S is a *p-sense* of w in the population. The set of presuppositions about rape made by feminists that they expected their audience to recognize was one p-sense of 'rape' in 1875; the set of corresponding conservative presuppositions another. When there are multiple p-senses of w in a population, we say (with a nod to Gallie (1956)) that the sense (sometimes I say 'the meaning') of w is *contested* in the population.[19] The meanings of 'martini', 'rape', and 'marriage' are all contested in the relevant populations.

The idea of a word's having a contested meaning is a fairly straightforward generalization of the idea of a word's meaning being its ICG. When a word's meaning is contested, each competent user uses it with one of its p-senses: competent users make the presuppositions that are the basis of that sense, expect auditors to recognize that, and make whatever higher order assumptions make the fact that w is used with this p-sense common knowledge. And for each Way of using the term—for each of its p-senses—it is common knowledge that *that* is a way the term is used in the population.[20] If one insists on synchronic stand-ins for word meaning, *the* meaning of a word, when its meaning is contested, is (realized by) the collection of its p-senses. This collection, after all, is something that competent speakers are in cognitive contact with. They know that its member senses are ones that users employ in communication: each constituent sense is known to include the assumptions that some subset of the population makes and expects its audience to recognize. The collection of p-senses associated with the relevant word are publically acknowledged; they are known to be senses that are in a certain sense "co-interpretable". If Colfax presupposes the first order assumptions in ON and Cady Stanton those in CN in joint conversation, they will, even if they know this, interpret one another as "talking about the same thing" when they use 'rape'; each will say things like 'well, you and I just disagree as to whether that is a case of rape'.

I think this is the best way to think of meaning in the case at hand. Compare it to the nihilistic view that in 1875 there was no more such a thing as the public meaning of 'rape' than there was such a thing as the eye color of human beings. The nihilist, it seems to me, has got ahold of the wrong analogy. True, there's no such thing as the eye color of human beings. But there *is* such a thing as the human eye. Its realization varies from individual to individual, but there are commonalities enough that make it

[19] Gallie assumed that only words whose meaning was in some sense normative could express what he called contested concepts—his idea, as I understand it, was that it was essential to the phenomenon he was interested in that differences in conception resulted from the way one's values or other broadly normative commitments influenced one's conceptualizing. One might agree with Gallie about this by arguing that *every* concept has a normative element.

George Lakoff, influenced by the work of his student Alan Schwartz, has made much of the notion of contested concepts. For a summary of Lakoff's take on the notion, see chapter 12 of Lakoff (2008); see also the quite remarkable senior thesis, Schwartz (1992).

[20] This is oversimplified, since there are different ways to be competent; again, see the discussion in chapter 3 of Richard (2019).

sensible to talk of *the* human eye, a reification of what's common to all normal eyes. There is variation across tokens of the human eye in such things as color and pupil size, but the variation is relatively constrained. In describing the human eye, one describes both the commonalities—rods and cones are always present—and the variations—one finds a range of pigmentations. Likewise, there was a common structure—a common set of first order presuppositions—to conservative and feminist views of rape in 1875, as well as something like allelic variation. Both the common structure and the variations can be read off of the (contested) meaning of the term 'rape', and it justifies the reification involved in speaking of *the* meaning of the term at the time.

The human eye is a biological object that has a history. While its structure is currently stable, it's nonetheless a historical entity. New eyes tend to resemble the eyes of those responsible for the body in which they are situated; future distribution of properties like eye color is determined by the way current eye owners and their progeny interact with each other and the environment; the eye is in principle liable to historical change due to mutation, selective pressures, and drift. All of this is mirrored in the semantic case: meaning structure is heritable; variant distribution is determined by such things as interaction of variants with one another and the environment and forces analogous to drift and mutation. The nihilist's view, refusing to think of meaning in population terms, misses just the sort of thing we miss if we refuse to speak of the human eye—or for that matter, the human species.

What was said above about coordination and reported speech gives reason to reject the view that in 1875 'rape' was ambiguous, as people like Stanton meant one thing with it while people like Colfax meant something else. This is really a version of the nihilistic view that there was no such thing as *the* meaning of the term. Agreed, there is a clearly *a* sense in which it is correct to say that different people meant different things in using the term—this difference in meaning resided in the differences in presuppositions speakers made and expected the audience to recognize as being made. But from the fact that people mean different things in *this* sense with the word, it doesn't follow that there isn't such a thing as *the* meaning of the word as it is used by everyone. First of all, there is of course a good deal that the various ways of using the term had in common: the presuppositions involved in each way of using the term were in a clear sense an extension of a 'common core' of presuppositions. Second of all, I take it to be an upshot of the sort of interpretive coordination of uses mentioned above that in interpreting the use of 'rape', speakers proceeded in the way one proceeds when one takes another to mean what one does with a term though differing on some of the "theory" associated with it. Speakers proceeded as would speakers with at least some commitment to reaching a common understanding of how the term was to be used—they co-interpreted but reserved the right to insist that their own conception was the one that all should adopt. That they proceeded in this way, I would say, means that they understood one another as sharing a word which had the same public meaning whoever used it. We, I would say, should understand them in this way as well.

The drift of the last few pages is that we are best off if we take the word 'rape' to be univocal as Stanton and Colfax use it. I imagine that some will agree with the claim about univocality, but say that what the word meant in 1875 is pretty much what it means today. The argument is simple: When Stanton and Colfax used the word, they were talking about, they were referring to, what *we're* talking about when we use the word. For suppose Colfax uttered

(R) It's impossible for a man to rape his wife.

Stanton would probably have asserted both

(1) When Colfax uttered 'it's impossible for a man to rape his wife', he said that it was impossible for a man to rape his wife.
(2) If Colfax said that it was impossible for a man to rape his wife, he was wrong.

All of *us* will say these things. The truth of (1) and (2) when uttered by Stanton and endorsed by us suggests identity of reference of 'rape' as used by Stanton, Colfax, and us.

Reference supervenes on meaning. So, whatever the word meant in 1875 has to be pretty closely related to what the word means today, closely enough so that the reference of the word then is what it is now. The simplest account is that the word's meaning is what it contributed and still contributes to what is said when the word is used. But this is what's determined by an articulation of CN—it's an extension, or (structured) intension, or something of the sort that picks out any and all forced intercourse.

We can agree for the sake of argument with the argument's steps—the nineteenth- and twenty-first-century references were the same; reference supervenes on meaning in some sense of meaning (and on environmental relations and other things); this means there has to be a pretty significant similarity in meaning in some sense of meaning across the centuries—up to the last one. I observed above that there are number of different notions of meaning: meaning as what determines reference and truth conditions; meaning as what is contributed to what is said; meaning as the anchor of competence, as that with which one must be in cognitive contact in order to be competent. It is meaning in this last sense that we are discussing. Why should we suppose it to be "simpler" (in a sense of simplicity that governs choice of theory) to identify this last sort of meaning with (what determines) reference than to say that what a word means is something that is constituted by the evolving practice of speakers and auditors to attempt to describe, think about, condemn, and regulate certain aspects of the social world? The idea that meaning *cum* the anchor of competence is something like interpretive common ground is one way of working out such a picture of meaning. When we think of meaning in this way there is distance between meaning and reference. If a meaning or concept is something like interpretive common ground, there is nothing particularly odd about the idea that all of those who think with a concept at a particular time are radically misconceiving what they are thinking about—there is nothing odd about the idea that the meaning of a term misrepresents what it is a concept of. So there is no problem, if we think of concepts in this way, with thinking that both (1) and (2) are true. But their truth doesn't imply that we have said what needs to be said about the once and current meaning of 'rape' once we have trotted an articulation of CN.

4

The A-theorist introduces a novel allele into the conceptual gene pool with the hope that it will be driven to fixation by social and intellectual forces, its competitors driven, if not to extinction, then at least to the status of marginal conceptual alternatives. Under what conditions does such a project have any chance of success? A preliminary to giving an answer is considering concrete cases.

The nineteenth-century feminist A-project is not an example of a completely successful A-project. Insofar as the audience addressed was not simply women willing to listen to feminist arguments, but the society as a whole, the project failed; witness that it was necessary to launch the whole thing again a hundred years later. But it succeeded in the sense that a significant number of people accepted the competitor notion CN, as opposed to the official notion ON, as 'the correct' way to think about the topic.

Why is this? Well, for one thing, the feminist project was conceptually, though not socially, rather conservative. The purposes the feminists in effect were assigning to the concept of rape would likely appear to the target audience to be more or less continuous with the cognitive and social purposes the concept already served, even at the beginning of the nineteenth century. Even then, to call something rape was, first and foremost, to identify it as a sexual, not a property, violation, and to condemn it for that reason; one did not, after all, need to be married to be raped. Insisting that marital rape is possible preserves this aspect of the concept of rape, while attempting to undermine an ideology that constrained the way in which the concept could be applied. Arguing for the understanding of rape embodied in CN is in good part a matter of straightforward ideological critique of ideas that are in some sense independent of the concept: once one rejects the view of consent underlying Hale's doctrine it is natural and hardly surprising that the legal and cultural understanding of rape would be transformed from ON into something like CN. A consequence is that it would not feel like false advertising—it would not have *been* false advertising—for feminists to represent themselves as pointing out that the best way to understand the **existing** practice of labeling and prosecuting things as rape is as a practice whose rationale—rationale, not practical upshot—is to condemn sexual violation of women no matter who the agent is.

Consider, next, the example of that brilliant *provacateur* who began using 'queer' as a badge, if not of honor, then at least of defiance and pride. There is a sense in which her 'proposal' for reworking the concept expressed by the word is also conceptually conservative. For the proposal—to express approval or at least neutrality towards gay people in applying the word 'queer'—is not a proposal to change what we might call the 'practical extension' of the term—that is, it is not a proposal, the upshot of which is to add or subtract from the collection of those to whom the term would commonly be taken to apply. The 'proposal', as I see it, is to change the affective, expressive component in the concept—its common, mutually recognized pragmatic trappings, if you will. Classification itself—the circumscribing of a particular (albeit fuzzily defined) group of objects—remains the same.

Consider now Haslanger's own version of the A-project. As noted above, when one reads her proposed analyses of concepts like *woman* and *Latino*, one has a strong feeling that she is engaged in a subject changing maneuver. It is not difficult to see why one might feel this way.

Haslanger's idea is that the application of a concept like *Latino* is the first step in a systematic (though partially non-conscious) process of discrimination against the group to which the term is applied. Latinos are so-classified in good part in order to discriminate against them; indeed, Latinos *are* people with a certain ethnic heritage who are discriminated against on the basis of that heritage. It is important to recognize this social fact. One way to do this is to revise our understanding of 'Latino' to reflect it.

One feels that, unlike the two examples of 'concept engineering' just mentioned, Haslanger's proposal **imposes** understandings of and purposes upon talk and thought involving the concept *Latino* that are quite foreign to the ways such thought and talk can be understood and the purposes it in fact serves. The proposal is certainly not classificatorily conservative in the way the feminist or the appropriative projects are.[21] The new concept doesn't arise simply by removing ideological accretions from something that could be said to be a notion that was there all along. We do not hold constant the classification the concept effects in practice while flipping its emotive valence. Nor do we claim that there was a kind users of the word in some sense meant to be talking about with the term, a kind that is more clearly conceptualized once the alternative analysis is adopted. One feels that the proposal's analysans is pretty much discontinuous with the analysandum on all relevant dimensions.

Now, it is not altogether clear whether this feeling is correct. There are any number of stories one might tell about "the" purpose or purposes of our gender and race concepts. Focus on the concept *Latino*. My suspicion is that the story most people would on reflection tell about the meaning of 'Latino'—and thus the core of the presuppositions most people make and expect to be recognized as making in using the term—is something like

> P1: The concept *Latino* is the concept of a person whose heritage includes (a significant number of) ancestors from Latin American countries who were themselves of Hispanic descent. Thus, to think of someone as a Latino is to think of them in this way.

Certainly the way we actually proceed in classification seems to be captured by something like this. Because of this convergence of presupposition and practical application, one is inclined to say that the best way to understand our existing practice of classifying people as Latino is given by P1. And so, one might argue, an account of the concept *Latino* like Haslanger's that incorporates a notion of subordination that is absent from both what's presupposed and from classificatory behavior is simply changing the subject.

[21] If this is not clear: The proposal, for reasons discussed in note 3, is probably best understood as a proposal that makes terms like 'Latino' relational: one is a Latino only relative to a culture in which one is systematically subordinated in virtue of one's ancestry or "Latino appearance". This seems to imply that if I go to, say, Mexico and comment on how many Latinos there are there, I am speaking falsely, for (there being no pattern of subordination on the basis of the "marks of being Latino"), there are, if the word's reference is determined by the proposal, no Latinos in Mexico.

But one might also say that, whether we are conscious of it or not, the following is true:

> P2: An important function of the concept *Latino* is that it facilitates classifying people whose heritage includes (a significant number of) ancestors from Latin American countries who were Hispanic as having such a heritage *so that* they can be discriminated against and otherwise subordinated. Thus, "the", or a point of having the concept is to facilitate discrimination on the basis of ethnicity.

Let us agree that a good part of the upshot of classifying people as Latinos is captured by P2. Because of this, one might say, a good part of the purpose of the concept is to conceptualize people ethnically so as to subordinate them on that basis. And so an account of the concept like Halsanger's, an account that incorporates a notion of subordination, is one that simply brings the concept's "point" into focus. So it can't be said to be an account that is "changing the subject".

I have doubts about this argument. There is difference between what a thing is and what it gets used for: a screwdriver doesn't become a can opener by being used almost exclusively to pry the lids off paint cans. I worry that the argument just given blurs this sort of difference. To agree that the *upshot* of ethnic classification is subordination is not to agree that in classifying ethnically we are classifying (in part) on the *basis* of subordination.

The A-project is a project that seeks to change the meaning of a term. There are at least two things that are naturally labelled as changing the meaning of a predicate, a change in its ex- or possible worlds in-tension—*r-change*—and a change in the presuppositions that constitute the predicate's ICG—*c-change*. The version of the A-project we are discussing looks to involve both, since it is a matter of giving an extension shifting meaning to terms like 'woman', 'Latino', and the like, and a matter of getting a group to take a certain way of thinking of the extensions for granted.[22] Insofar as this particular version of the project involves extension shifting, it strikes me that it was never likely to be successful. Haslanger tells us that she wants to answer such questions as *What is it to be a man? What is it to be a Latino?*. The answers are to be 'critical analytical' ones, in the sense that the search for answers is to be guided by considering "what work the concepts of gender and race might do for us in a critical... social theory" (Haslanger 2000: 226). But of course we *begin* by using the concepts *man* and *Latino* in delimiting the project. An extension shifting answer strikes me as one very difficult to make stick—as very difficult to get people to accept—if it is not grounded in something about prior usage that can be adduced to make plausible that the answer "simply reveals what we were talking about all along", or that the answer is an apt response to an ambiguity, confusion, or inconsistency in prior use.

It is worth observing that r-change is in this regard quite different from c-change. Suppose, addressing feminists and race theorists, that I point long and loudly to the facts that women and people of color *as classes* are subordinated, and that this

[22] "way of thinking" has unfortunate Fregean connotations: Fregean ways of thinking (senses) are reference determining. I intend here ways of thinking in a more or less colloquial sense, on which (for example) stereotypes associated with racial and gender terms are ways of thinking of their references associated with the terms. Of course ways of thinking in *this* sense are not reference determining.

subordination is achieved on the basis of a classification in terms of "observed or imagined bodily features presumed to be evidence of their role in reproduction (women) or ancestral links to a certain geographic region (racial groups)". To say that these classes are subordinated on these bases is of course not to say that all the members of the classes are. Rather it is (in part) to say that there are mechanisms in place that tend to lead those who display the features to be subordinated; the claim is a generic, not a universal. Suppose I make it clear that the fact that I am pointing to is a fact about history and culture—it is a fact about women and minorities in particular historical and cultural contexts, significant in part because the relevant sort of subordination occurred and occurs in a startlingly wide swath of history. Suppose I go on to say that this fact about females and people who have the relevant racial heritage is significant enough that it should be at the forefront of our theorizing about gender and class. And suppose finally that I am heard: people accept what I say, recognize that others do, and come to expect others to know these facts. As a result, **generics** like *women are subordinated on observed or imagined bodily features presumed to be evidence of their role in reproduction* become **part of the ICG** of the term 'woman' (and so in a tolerably clear sense, that I have been trying to lay out, become part of the concept *woman*); as a result, such generics come to play a role in thought and theory about women and minorities.

All of this would effect a change in the meaning of 'women', 'Latino', and so on—not an r-change, but a c-change. It is not a change in what people are talking about or in what they think they are talking about with those terms. Rather, it is a change in the way they think about what they are talking about, a change in the assumptions and presuppositions they make when they use the terms. It is a change that is relatively easy to effect—indeed, it's plausible that some progress has already been made in getting people in general, not just activists and academic theorists, to think of the relevant groups in this way. Effecting this sort of change, it seems to me, achieves much, perhaps most, of what Haslanger's project was meant to achieve. And it does this without having to take on the burden of shifting the reference of anything. For changing what everyone takes for granted in using a word is not, in itself, shifting what anyone is talking about with a word.

Bringing about what I've been calling c-change is a sort of 'conceptual engineering'. One might engage in it with the intention that it will lead to r-change. But there's no need to have such an intention in order to try to change what is common ground about what users of a (term expressing a) concept presuppose. And of course this sort of 'conceptual engineering' is not particularly the province of philosophy or of the academy in general. It happens all the time.

Conceptual engineers and ameilorists often describe their projects in ways that imply that they will be successful only if the reference of the concept under their scalpel shifts. That seems to me a pretty narrow vision of what it could be to ameliorate our thinking. Certainly the arch conceptual engineers—propagandists, advertising copy writers, spinmeisters, cagey politicians—don't think of what they are doing in such terms. The trumpets of the Trump have pretty much succeeded in getting the generic idea that illegal immigrants are bad *hombres* into the common ground of certain groups, so that members of those groups presuppose and expect to be recognized as presupposing this when they use 'illegal immigrant'. Doing this, I would say, clearly changed the meaning of 'illegal immigrant' in those groups, but of course it didn't change its reference. Pretty obviously the goal was never to change the reference of the term: shifting the reference would have been the wrong outcome, since the goal was obviously to get people to think of *illegal immigrants*, not of illegal immigrants who are bad *hombres*, as rapists and murderers.

We ought to think of conceptual amelioration and engineering as an attempt to foster a kind of evolution within a population. The revisionary analyst drops a mutation into a population, hoping that it will "reproduce" and in one way or another establish itself, even replace all of its alternatives over time. The goal might be referential shift, but often enough such shift will be unnecessary for the project to achieve whatever goals are driving it. In order to think fruitfully about the prospects of success for such a project, we need to answer various questions. In what sense do meanings and concepts reproduce? Given a population into which a new use of the word is introduced, under what conditions can we expect the new use to establish itself? What sorts of conversational encounters make people adopt a new interpretive strategy, one changing the presuppositions they take to accompany a term's use? Do new meanings reproduce fastest if they are first firmly entrenched in small groups, or do they naturally spread like the flu? Etc., etc.

If you think of the A-project not as an ivory tower exercise—an *ex cathedra* philosophical pronouncement of what the little people should be meaning with their words—but as a genuine attempt to effect social change you should be thinking about these sorts of questions. You should be asking questions like: What are reasonable, what are unreasonable models of how conceptual change occurs in a population? Given that we think a model reasonable, which of its variables are open to manipulation by the revisionary analyst? How do we change the strategies that people bring to the game, that is, to the project of interpreting others?

These are the sorts of questions that we ought to be asking, not only about versions of the A-project, but about conceptual analysis in general. Conceptual analysis is generally not just descriptive but normative. In interesting cases—the analysis of knowledge, of free action, of truth—what we tend to find is evidence not of a single underlying albeit vague concept, but a profusion of more or less mutual, not altogether consistent, presuppositions and patterns of application that (with a bit of the philosopher's art) can be resolved into a collection of candidates for what we might mean by the terms we use. To arbitrate amongst them is at least in part a matter of asking, not what natural or gerrymandered kind we are trying to pick out, but asking what the point or points of having and applying the concept under study is. Philosophical analysis is pretty much *always* a (thinly veiled) version of the A-project. As such, philosophical analysis is not simply theory; it is practice. And as practice, it demands that its practitioners be practical.

References

Gallie, W. B. 1956. Essentially Contested Concepts. *Proceedings of the Aristotelian Society* New Series 56 (1955–6):167–98.

Hasaday, Jill. 2000. Contest and Consent: A Legal History of Marital Rape. *California* Law Revue 88:1373.

Haslanger, Sally. 1999. What Knowledge Is and What It Ought to Be: Feminist Values and Normative Epistemology. *Philosophical Perspectives* 13:459–80.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Nous 34*:31–55.

Haslanger, Sally. 2006. What Good Are Our Intuitions? Philosophical Analysis and Social Kinds. *Proceedings of the Aristotelian Society Supplemental* 80:89–118.

Haslanger, Sally. 2012. *Resisting Reality*. Oxford: Oxford University Press.

Lakoff, George. 2008. *The Political Mind*. London: Viking Penguin.

Nickel, Bernhard. 2016. *Between Logic and the World*. Oxford: Oxford University Press.

Richard, Mark. 2019. *Meanings as Species*. Oxford: Oxford University Press.

Schwartz, Alan. 1992. Contested Concepts in Cognitive Social Science. Senior Thesis, UC, Berkeley.

18

Talk and Thought

Sarah Sawyer

1. Introduction

Thought and language are uniformly acknowledged to be distinct phenomena requiring distinct philosophical treatment. Thought is one thing; language another. Nonetheless, they are intimately related. According to Frege, 'The thought, in itself immaterial, clothes itself in the material garment of a sentence and thereby becomes comprehensible to us. We say a sentence expresses a thought' (Frege 1918: 292). Abstracting away from Frege's commitment to the immateriality of thoughts, the assumption that sentences express thoughts is widely accepted. It has, however, led to a near-ubiquitous identification between the linguistic meanings of non-indexical sentences on the one hand and the contents of the thoughts expressed by those sentences on the other. The identification is embedded in the widespread practice of appealing to a single proposition to specify both the semantic content of a non-indexical sentence and the content of the thought thereby expressed. Thus the semantic content of the sentence 'Marriage is a legal union of two people' is taken to be the proposition that marriage is a legal union of two people, which proposition is also taken to be the content of the thought the sentence expresses. At the level of words rather than sentences, the terms 'linguistic meaning' and 'concept' are typically treated as synonyms, as evidenced by the widespread practice amongst philosophers of switching freely between the two, often within a single sentence. What is said and what is thought are thus typically treated as identical.[1]

But the identification of what is said and what is thought is, I will argue, a mistake. It involves a conflation of two distinct phenomena. In this chapter, I outline an externalist account of linguistic meaning and an externalist account of thought content that clearly distinguishes the two.[2] I advocate, instead, a dual-aspect theory of representation according to which sentences *have* semantic contents and *express* thoughts. Crucially, the two are not only theoretically distinct but diverge in actual

[1] The practice is so widespread that providing references would be an impossible task.

[2] The view and its implications for meaning-shift are presented in Sawyer (2018). In the current chapter I discuss a wider range of issues, focusing in particular on the nature and value of conceptual engineering. The distinction between linguistic meaning and thought content can be found in Burge (1986).

Sarah Sawyer, *Talk and Thought* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Sarah Sawyer.
DOI: 10.1093/oso/9780198801856.003.0018

cases. Indeed, for creatures like us, divergence is the norm.[3] Understood primarily as a thesis about terms rather than sentences, the externalist, dual-aspect theory of representation maintains that non-indexical linguistic terms such as 'whale', 'number', 'explanation', 'fairness', 'marriage', and 'gender' *have* linguistic meanings and *express* concepts. Linguistic meaning is here to be understood as linguistically encoded content rather than as reference or denotation; and concepts are to be understood as representational constituents of thoughts individuated at the level of sense rather than reference.[4] Here too, the crucial point is that the linguistic meaning of a non-indexical term and the concept expressed by that term are not only theoretically distinct but diverge in actual cases. This is because, on the account I offer, linguistic meanings and concepts track different phenomena and play different explanatory roles.

The distinction, understood along the lines I propose, brings theoretical gains in a cluster of related areas. It provides an adequate account of meaning change across time, it accounts for the possibility of substantive agreement and disagreement across different theoretical frameworks, it accommodates at face value the phenomenon of contested meanings, and it explains both the nature and the value of conceptual engineering, placing the phenomenon of conceptual engineering in a framework of theoretical change more broadly understood in a way that addresses recent prominent concerns.

A caveat before we begin. The current chapter is focused on non-indexical terms only. I set aside discussion of indexical terms, such as 'I', 'here', and 'now' as well as of proper names, such as 'London' and 'Sarah', which I have argued elsewhere should be understood as containing an indexical element in their singular use.[5] Indexical terms require separate treatment.[6] I focus instead on the claim that the linguistic meaning of a non-indexical term should be distinguished from the concept expressed by that term. This is where the theoretical gains lie.

2. The Background: Internalism and Externalism

The (problematic) identification of the linguistic meaning of a non-indexical term and the concept expressed by that term sits most naturally within a thorough-going internalist theory of both thought and language. Internalism about thought and language are defined by their commitment to a local supervenience thesis according to which what a subject thinks and what she means by her words are each determined by her intrinsic (typically, physical) states. Internalist theories of thought and

[3] As will become clear, narrowing the divergence between the two is the upshot of successful inquiry. See section 6 below.

[4] See Frege (1892).

[5] For my view on proper names, see Sawyer (2010), which has its roots in Burge (1973).

[6] For present purposes, I take as an indexical term any term which requires a contextual application in order to determine a referent. That is, roughly, any term that admits of a content/character distinction in the sense of Kaplan (1989). Indexical terms do not obviously express concepts—or, at least, are not guaranteed to express a single concept across different occasions of use; and the semantic content of a sentence containing an indexical term is not obviously identical to the content of the thought expressed. Which terms fall into this category is a matter of debate.

language are designed to capture the way the world seems from the individual's perspective narrowly construed.[7] Concepts and linguistic meanings, then, understood as the internalist understands them, are determined by the same set of intrinsic facts and are invoked to capture a single phenomenon, namely the subject's individual, perceptual, discriminatory capacities and inferential dispositions which together inform her deployment of a concept and her use of a term. This means that within a thorough-going internalist framework, not only is there no reason to distinguish the linguistic meaning of a term from the concept expressed by that term, there is also no means of doing so. It is precisely this feature of a thorough-going internalist theory that, I maintain, precludes it from providing an adequate account of phenomena such as meaning change, substantive agreement and disagreement across different theoretical frameworks, and the nature and value of conceptual engineering. I return to this issue in section 6 below.

A distinction can be drawn between the linguistic meaning of a term and the concept expressed by that term only within a framework that distinguishes between the facts that determine the former and the facts that determine the latter. A thorough-going internalist framework does not have the resources to do this, but a thorough-going externalist framework does.[8] In fact, Putnam's original introduction of externalism to the mainstream literature provides the means to distinguish the two, but his exclusive focus on language obscures this important insight. Putnam, concerned with language rather than thought, identifies two elements that he claims are missing from traditional semantic theories.[9] The first is the contribution of society, and the second is the contribution of the real world. The contribution of society is illustrated by Putnam's example involving his incomplete understanding of the terms 'beech' and 'elm', and brings in the core notion of linguistic deference, which he articulates as deference to 'experts'. The contribution of the real world is illustrated by Putnam's notorious Twin Earth thought experiment involving the term 'water', and brings in the core notion of causal relations to natural kinds with hidden essences. My suggestion, elaborated throughout the chapter, is that the contribution of society to representation is best understood as a contribution to language and the contribution of the real world to representation is best understood as a contribution to thought.[10]

It is reasonable to think not only that the contributions Putnam identifies can be understood as contributions to different phenomena, but, in addition, that they

[7] See for example Fodor (1980, 1987), Stalnaker (1990), Segal (2000), and Chalmers (2003). For an internalist account of linguistic meaning most relevant to the concerns of the present chapter, see Sundell (2011, 2012) and Plunkett and Sundell (2013).

[8] 'Mixed' theories which combine an internalist theory of thought and an externalist theory of language (or vice versa) also have the resources to do this; indeed, they necessarily entail a distinction between the linguistic meaning of a term and the concept expressed by that term, although this implication has not generally been explored. I do not discuss such theories in the chapter because the way in which they draw the distinction does not imply the theoretical advantages with which I'm concerned. For examples of mixed theories of the first kind see Putnam (1973, 1975) and Crane (1991). I know of no-one who holds a mixed theory of the second kind and can see little motivation for such a view.

[9] See Putnam (1973, 1975).

[10] As will become clear, I take the contribution of the real world to extend far beyond what is given by causal relations to natural kinds with hidden essences.

cannot be seen as contributions to a single phenomenon, whether that be to language, as Putnam maintains, or to thought. This is because there is a tension between the two contributions, a tension which is masked by an ambiguity in the notion of an 'expert'. It is standard in the externalist literature to assume that an expert is someone who is knowledgeable about the relevant subject matter. On this understanding, the contribution of society and the contribution of the real world would necessarily be aligned. On this view, the linguistic meaning of an individual's term is determined either by causal relations to the real world, or by deference to experts, the linguistic meaning of whose term is determined by causal relations to the real world. Either way, linguistic meaning is determined at root by relations to the real world; the only question is whether the determination relation is direct or indirect.[11]

But the claim that linguistic deference involves deference to those who are knowledgeable about the relevant subject matter is an idealization, and one which cannot play the central role required of it in an account of *actual* linguistic deference. Linguistic deference must be deference to other members of one's linguistic community, and these may not be experts in the sense of being knowledgeable about the relevant subject matter. The history of science is the history of experts who were wrong, sometimes significantly so, about the subject matter that fell into their area of expertise. It would be wrong to conclude that they were not experts after all; on the contrary, to be an expert is to be worthy of deference despite the possibility of error. This fits with our actual linguistic and sociological practice surrounding the term 'expert'.

We must reject the standard externalist assumption, then, that an expert, for the purposes of linguistic deference, is necessarily someone who is knowledgeable about the relevant subject matter. Rather, an expert is, roughly, someone to whom others defer not because they *are* knowledgeable but because they are *perceived* to be so. Being perceived to be knowledgeable does not, of course, preclude being knowledgeable; but being knowledgeable is neither necessary nor sufficient for being perceived to be so. Note, however, that on this more realistic, less idealized understanding of an expert, linguistic deference provides no guarantee that the linguistic meanings of our terms hook up with the real world in the way that is supposed to be secured by direct causal relations. This is because the way the experts take the world to be may be different from the way the world is.[12] The contribution of society and the contribution of the real world will therefore sometimes take us in different directions. To resolve the tension, the two contributions that Putnam identifies not only can, but should be seen as contributions to different phenomena.

The contribution of society to representation, I suggest, is to be understood as a contribution to language, and the contribution of the real world to representation is,

[11] It is a mistake to think that the contribution of the real world is relevant to natural kind terms whereas the contribution of society is relevant to non-natural kind terms. Drawing the distinction along these lines is clearly not true to Putnam's original discussion which uses examples of natural kind terms such as 'beech' and 'elm' to illustrate the phenomenon of linguistic deference. It also fails to do justice to the hidden depths of the natural world, the reality of the non-natural world and the fundamental role of linguistic deference in language-learning.

[12] Errors may be both theoretical, in the form of false beliefs, and practical, in the form of misapplications of words and incorrect deployment of concepts.

I suggest, to be understood as a contribution to thought. These distinct 'external' facts ground a theoretically significant distinction between the linguistic meaning of a term on the one hand and the concept expressed by that term on the other. I take the distinction to be general, applying to terms across a broad range of disciplines and areas of inquiry. The distinction is not restricted to natural kind terms, or to empirical terms more broadly construed, but extends beyond the empirical realm to mathematical, logical and philosophical terms as well as to normative terms, including social and ethical terms. In what follows, I provide an externalist account of linguistic meaning and an externalist account of concepts, and I demonstrate some of the theoretical gains of distinguishing the two.

3. Linguistic Meaning

I take linguistic meaning to supervene on use. I reject the internalist claim, however, that the linguistic meaning of a term as used by a given individual depends solely on her use of the term. There is often widespread variance in the use of a term by different individuals across a community, but this variance is consistent with a term's having a single meaning for all. Linguistic meaning in this communal sense is what dictionaries aim to record. They achieve this aim, in so far as they do, not by recording statistical averages across individual usage but by paying attention to patterns of usage and deference across the community. Patterns of deference reflect our general recognition of the fact that, for any given term, some individuals are more competent in its use than others. It is the use of the most competent together with patterns of deference amongst all (including the most competent) that determines linguistic meaning.

Linguistic deference is a much more general phenomenon than is sometimes recognized, applying to most, probably all, terms in the language. This includes, for example, very basic terms such as 'red', 'table', and 'sharp', and slang terms such as 'wicked', 'newb', and 'banterous'. It is the generality of the phenomenon of linguistic deference that makes talk of expertise misleading. Not only does the term 'expert' carry connotations of knowledgeable status, as noted in the previous section, but it also carries connotations of a restricted application to scientific, theoretical, or technical terms only. This restriction is artificial, since the fundamentality of linguistic deference is built into the very nature of language-learning. Talk of competence is preferable, even if not perfect.

What I have said so far is couched at a relatively general level. At a more specific level, the linguistic meaning of a term at a time can be understood as the characterization of the relevant subject matter that members of the linguistic community would settle on at that time were they to reach reflective equilibrium in the context of a dialectic.[13]

The dialectic, for these purposes, is an honest, open debate, shorn of all subjective elements, in which participants aim for a characterization of the subject matter

[13] For an account of linguistic meaning along these lines, see Burge (1986, 1989) and Sawyer (2007, 2018).

through reason and reflection on actual and hypothetical cases, deferring to the most competent as and when appropriate. The dialectic is not to be understood as involving maximal reflection on the subject matter, as this would inevitably extend the discussion beyond actual use to what was perceived to be ideal future use. Rather, the relevant notion is to be understood as full reflection within actual empirical and theoretical boundaries, where this is consistent with the participants agreeing that the subject matter has not yet been fully characterized. The primary focus of the dialectic is the characterization of a subject matter rather than the characterization of meaning *per se* because although the questions 'What does 'x' mean?' and 'What is an x?' are different questions, the former can be, and typically is, answered by answering the latter.[14]

Within the set empirical and theoretical boundaries, there may be terms for which no agreement would be reached. The account implies that for such terms there is no settled linguistic meaning at the time. I take the claim that at least some terms have no settled linguistic meaning at a time to be a natural implication of the fact that meaning depends on use, which is subject to flux and changes over time. I therefore see the implication that not every term will have a settled linguistic meaning as a virtue of the account. Language is organic, and this fact must be captured by an adequate theory of linguistic meaning.

Of particular significance in this context are terms with so-called 'contested meanings'.[15] The literature on conceptual engineering provides a set of staple examples, including 'race', 'gender', 'marriage', and 'rape', although terms with contested meanings occur across all realms of inquiry. The meanings of such terms are contested in the sense that different groups of people within the linguistic community—the conservative group and the progressive group, as we might call them—disagree in fundamental ways about how the term ought to be used, where this disagreement is reflected by the different actual uses of the relevant term by the different sub-sections of the community. In such cases, the community as a whole is aware of the disagreement in use, but the disagreement takes place against the common understanding that the matter would not be resolved by a stipulative disambiguation introduced to accommodate the different uses. The different parties to the dispute, each convinced that their own use is correct, do not see themselves as talking past each other, but as offering different views on a single subject matter. In the context of a dialectic, even if each sub-section of the community were to settle on an agreed characterization of the relevant subject matter, no overall agreement would be reached by the community as a whole. The fundamental disagreement over the correct characterization of the subject matter in such cases splits the use of the relevant term and renders its meaning contested. I return to these important cases in section 5 below.

The account of linguistic meaning offered accommodates the fact that the meaning of a term can change over time. Since meaning supervenes on use, a change in the linguistic meaning of a term depends on an underlying change in linguistic practice;

[14] This point is made in Burge (1986), where it is related to the claim made in Quine (1951) that there is no separating truths of meaning from matters of fact.

[15] See Gallie (1956) for an early discussion of 'essentially contested meanings'.

and if there is a change in linguistic practice, there will be a change in the characterization that would be settled on in the context of a dialectic. For example, the meaning of the term 'meat' has clearly changed between Shakespearean times and now. The term 'meat' used to mean something like *food in general*, whereas now it means something like *animal flesh that is eaten for food.* But this is precisely the result that would emerge in the context of dialectical reflection on use at the two times.

Given that the linguistic meaning of a term determines its extension, the account of linguistic meaning also accommodates the fact that the extension of a term can change over time. I take the extension of a term to be the class of entities that satisfy the term's (descriptive) linguistic meaning. Thus it is because an apple satisfies the description 'food in general' that it falls into the extension of the term 'meat' in Shakespearean times; and it is because an apple does not satisfy the description 'animal flesh that is eaten as food' that it does not fall into the extension of the term 'meat' now. Finally, the account also provides linguistic norms by establishing how the relevant term ought to be used, namely in accordance with the agreed characterization of the relevant subject matter. I return to questions of normativity and truth in section 7 below.

4. Concepts

A quick internet search provides an extensive list of words that have changed their meaning over time, including 'meat', 'spinster', 'bachelor', 'clue', 'awesome', 'awful', 'wicked', 'girl', 'egregious', 'pretty', and 'hussy'. In all of these cases, and many more besides, the change in meaning has been accompanied by a change in topic, or subject matter.[16] This is what makes such cases relatively unproblematic from a philosophical perspective.

The more interesting cases are those for which we want to say that the meaning of the relevant term has changed while the subject matter has not. Such cases are widely regarded as philosophically problematic. To illustrate the nature of the problem raised by this kind of case, I start with a puzzle articulated by Sainsbury, although a variation of the puzzle occurs across a wide range of philosophical literature.[17] Let us assume, as Sainsbury does, that the term 'whale' was embedded in a linguistic practice in ancient times, when people thought whales were fish, that is different from the linguistic practice in which it is embedded now, when people think whales are mammals. This, he says, raises a dilemma. Either the sentence 'Whales are fish' means the same in ancient times as now, or it doesn't. To say that it does fails to accommodate the fact that meaning is determined by use; but to say that it doesn't fails to accommodate the fact that there is substantive disagreement across the two times.

Substantive disagreement is to be contrasted with merely verbal disagreement, such as the kind of disagreement that might occur over whether an apple is a form of meat between someone from Shakespeare's time and someone from the present time,

[16] Unlike technical terms such as 'extension' and 'reference', I take the terms 'subject matter' and 'topic' to be non-technical and relatively intuitive.

[17] For Sainsbury's articulation of the puzzle, see his (2014).

the former insisting that an apple is a form of meat, the latter insisting that it isn't.[18] The disagreement would be merely verbal in the sense that the two parties would be talking past each other. One way to capture this is to note that there is no single content over which the parties disagree. We can stipulate that 'meat$_s$' is to mean what 'meat' means in Shakespeare's time, and 'meat$_P$' is to mean what 'meat' means in the present. It is plausible to assume that the disputants, once apprised of the difference in use, would accept the stipulations and agree that apples are a form of meat$_s$ but not a form of meat$_P$. The dispute would not persist after the stipulation because it would be clear that the subject matter of the term 'meat' in Shakespeare's time is not the same as the subject matter of the term 'meat' now. Disambiguation works in this case because it separates what are clearly two distinct subject matters.

Substantive disagreement, in contrast, is disagreement over a single subject matter. The disambiguation strategy that works in cases of merely verbal disagreement does not, therefore, work in cases of substantive disagreement. The disagreement over the truth of the sentence 'Whales are fish' persists even if we stipulate that 'whale$_a$' is to mean what 'whale' means in ancient times and that 'whale$_P$' is to mean what 'whale' means in the present. This is because the stipulation does not disambiguate two distinct subject matters; rather, it distinguishes two theories about a single subject matter. The question remains which theory, if either, is correct. The disagreement is fundamentally a disagreement over the nature of whales.[19]

Sainsbury's puzzle is puzzling, then, because, intuitively, we want to be able to say both that the linguistic meaning of the sentence 'Whales are fish' is different at the two times and that the subject matter of the sentence is the same at the two times, and it is unclear exactly how we can do both. The reason it is unclear how we can do both is that sameness of subject matter appears to require a single propositional content over the truth of which the parties to the dispute disagree, and this appears to require sameness of linguistic meaning which directly contradicts the claim that the linguistic meaning of the term has changed.

The diagnosis of the puzzle lies in the recognition that if linguistic meaning is the only representational element in our theory, it is subject to inconsistent constraints. Linguistic meaning cannot both supervene on use *and* determine a stable subject matter. This is because in order to supervene on use, linguistic meaning must change in accordance with a change in linguistic practice that comes about as a result of a change in the community's beliefs; but in order to determine a stable subject matter, linguistic meaning must be insensitive to at least some changes in the community's beliefs and hence insensitive to at least some changes in linguistic practice. No single element can do both.

[18] On the nature of verbal disputes, see for example Chalmers (2011). Although I agree with much of what Chalmers says, I think the account ultimately suffers from the conflation between linguistic meanings and concepts that I urge in the current chapter. I do not have the space to discuss Chalmers's views in detail here.

[19] The boundary between merely verbal disagreements and substantive disagreements may be vague and will certainly sometimes appear so, given that the distinction between subject matters is not necessarily transparent. This is consistent, however, with there being clear-cut cases on either side of the boundary.

The solution to the puzzle, then, begins with the recognition that two representational elements are required; one to supervene on use, and the other to determine a stable subject matter. Of the two roles, linguistic meaning is ideally suited for the former. The account of linguistic meaning offered above starts from the assumption that meaning supervenes on use, and explains how a change in linguistic meaning tracks changes in a community's linguistic practice. The second role is that of determining a stable subject matter. This, I maintain, is the function of concepts. Concepts are constituent, representational elements of thoughts that connect thinkers representationally to a subject matter about which individual or communal beliefs may vary.

In order for concepts to play the requisite role of securing a stable subject matter, they must be understood as externally individuated, fundamentally non-descriptive components of thought. They must be externally individuated if they are to determine a subject matter that can be stable across individuals with different individual beliefs; and they must be fundamentally non-descriptive if they are to determine a subject matter that can be stable across communities with different communal beliefs. Concepts are not individuated by individual conceptions—they are not individuated by the way the individual thinker takes the world to be. Nor are they individuated by communal conceptions—they are not individuated by the way the community as a whole takes the world to be. Concepts are individuated, at the fundamental level, by relations to objective properties. This is the sense in which the contribution of the real world to representation is best understood as a contribution to thought. It is the fact that the subject matter itself enters into the individuation conditions of the relevant concept that explains how the concept expressed by a term can determine a stable subject matter.[20]

Once two distinct representational elements are acknowledged, Sainsbury's puzzle is resolved. We wanted to be able to say both that the linguistic meaning of the sentence 'Whales are fish' is different at the two times, and that the subject matter of the sentence is the same at the two times. This we can now do. The linguistic meaning of the term 'whale' has changed over time, but it expresses the same concept at the two times, and hence concerns the same subject matter; the linguistic meaning of the sentence has changed over time, but it expresses the same thought at the two times, thereby securing a single propositional content over the truth of which the communities can be understood to have a substantive disagreement.

The assumption that substantive disagreement presupposes a single propositional content over the truth of which the parties to the dispute disagree is almost ubiquitous. It is questioned by Plunkett and Sundell, who argue that intuitions about substantive disagreement can be accommodated without appeal to a shared propositional content.[21] It is interesting to note that both sides to this particular dispute assume that there is only one propositional content in question, which is both the

[20] I say 'at the fundamental level' to allow that there may be some descriptive concepts and some empty concepts. By 'fundamental', I mean representationally fundamental, and I take a liberal view on the issue and include amongst the fundamental concepts ordinary concepts such as those expressed by the terms 'whale', 'marriage', 'race', 'gender', 'belief', 'rape', 'moral goodness', and 'justice'.

[21] See Sundell (2011, 2012) and Plunkett and Sundell (2013).

semantic content of the sentence and the thought expressed by it. The mainstream assumes that this propositional content must be identical in cases of substantive dispute; Plunkett and Sundell argue that it need not be. I offer a middle way. I agree with the mainstream that the best explanation of our intuitions about substantive disagreement is that there is a single propositional thought content over the truth of which the parties to the dispute disagree; I agree with Plunkett and Sundell that substantive disagreement does not require sameness of linguistic meaning.

5. Theoretical Frameworks and Contested Meanings

Distinguishing the linguistic meaning of a term from the concept expressed by that term in the way that I have suggested offers a solution to the question of how substantive disagreement can occur across theoretical divides. The linguistic meaning of a term at a time is determined by the received theory of the relevant subject matter at that time; but the concept expressed by a term is not determined by the theoretical framework within which it occurs; rather, the concept expressed by a term provides an anchor to a shared, objective world. This means that concepts may be shared across different communities who hold substantially different theories of the same subject matter. The account thus avoids a pernicious kind of conceptual relativism and overcomes the kind of incommensurability that Kuhn thought plagued theory change.[22]

For the same reason, it provides a proper understanding of the phenomenon of contested meanings. As noted in section 3 above, we can say that the meaning of a term is contested when there is disagreement in use between sub-sections of the community and the disagreement would not be resolved by a stipulative disambiguation introduced to accommodate the different uses because the disagreement concerns the correct characterization of a single subject matter. In this sense, contested meanings essentially involve substantive rather than merely verbal disagreements. The term 'marriage' has a contested meaning in this sense. Use of the term 'marriage' by conservatives differs from use of the term 'marriage' by progressives, the former insisting that marriage is necessarily a union of a man and a woman, the latter disagreeing.[23] But the disagreement would not be resolved by a stipulative disambiguation, with the meaning of 'marriage_C' determined by the conservative use and the meaning of 'marriage_P' determined by the progressive use. It would not be resolved because the progressives' very point is that the sex of the individuals concerned is irrelevant to the question of marriage—that there are not two kinds of marriage, but one. The disagreement, then, is a substantive disagreement about the nature of marriage. The stipulation does not disambiguate two distinct subject matters; rather, it distinguishes two theories about a single subject matter.

The important point to note here is that the characterization of contested meanings makes essential reference to the fact that they involve substantive disagreements about a single subject matter. This incurs an explanatory debt. We need an account of

[22] See Kuhn (1962).

[23] The relevant issue in this context is not about whether the law precludes same-sex couples from being married but about whether there is something about the nature of marriage that precludes it.

how it can be that a term with a contested meaning is about a single subject matter when the parties to the disagreement use the term in such different ways. The theory of concepts offered in the previous section provides the requisite explanation. What makes the disagreement between conservatives and progressives a disagreement about marriage is that the term 'marriage', despite having a contested meaning in the community at large, expresses a single concept that anchors the use of the term by both groups to a single subject matter. The distinction between linguistic meanings and concepts I have suggested, then, provides an adequate explanation of the phenomenon of contested meaning by explaining how a disagreement in use between sub-sections of society can be a substantive disagreement about a single subject matter. Which terms have contested meanings will naturally change over time as a settlement is reached for some terms and disagreement arises over others. Contested meanings are grounded in theoretical differences within a community at a time. The changing landscape of contested meanings reflects the forces of theoretical and social change, and is an important phenomenon for this very reason.

6. Conceptual Engineering

The phenomenon of contested meanings has an important role to play in the context of conceptual engineering, which also faces a number of problems that can be resolved by distinguishing between linguistic meanings and concepts in the way that I have suggested. Let us take as an example of a recent project in conceptual engineering Haslanger's proposed revisionary analysis of 'woman'.[24] According to Haslanger:

> S is a woman iff
>
> i. S is regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction;
> ii. that S has these features marks S within the dominant ideology of S's society as someone who ought to occupy certain kinds of social position that are in fact subordinate (and so motivates and justifies S's occupying such a position); and
> iii. the fact that S satisfies (i) and (ii) plays a role in S's systematic subordination, that is, *along some dimension*, S's social position is oppressive, and S's satisfying (i) and (ii) plays a role in that dimension of subordination (2012: 234, original emphasis).

Revisionary analyses of this kind must satisfy two constraints: they must be both revisionary and conservative. They must be revisionary in the sense of proposing a new meaning for the relevant term, where this typically has the effect of shifting its extension; and they must be conservative in the sense of being an analysis of the

[24] Haslanger uses the term 'ameliorative' and there is a question about whether and in what way an ameliorative analysis is to be understood as revisionary. I take the distinction between linguistic meanings and concepts to help clarify this issue too, but do not have the space to discuss it here. I use the term 'revisionary' because of its general applicability.

original subject matter. If they do not satisfy the former constraint they are merely descriptive projects; and if they do not satisfy the latter constraint they have not provided an analysis of the relevant subject matter but have simply changed the topic. This creates a puzzle, since it is unclear how both constraints can be satisfied at once. The problem is that a change in extension brought about by a change in meaning appears to result in a change in topic; thus satisfaction of the first constraint appears to preclude satisfaction of the second.[25]

The puzzle is driven by the (false) assumption that a change in extension corresponds to a change in subject matter, or topic. The assumption fits naturally within the kind of view that identifies linguistic meanings and concepts—the kind of view I have been arguing against. On such a view, there is a single representational element that determines both extension and subject matter, and hence it is possible (and natural) to identify (or conflate) the two. But the solution to the puzzle requires their separation.[26] One of the theoretical benefits of the dual-aspect theory of representation I am advocating is that it not only accommodates the separation of extension from subject matter, but it actually implies that the two are distinct and, moreover, that they can easily diverge. According to the account of linguistic meaning provided in section 3, the linguistic meaning of a term determines an extension descriptively. According to the account of concepts provided in section 4, the concept expressed by a term determines a subject matter non-descriptively. This means, effectively, that the extension of a term is determined by theory, whilst the concept expressed by a term is not. As a result, concepts can provide an anchor to a stable subject matter, or topic, about which there may be different proposed analyses.

Consider this in the context of Haslanger's revisionary analysis of 'woman'. The account of linguistic meaning I have offered implies that the extension of the term 'woman' will change if the traditional meaning of the term is replaced by the proposed revisionary analysis. Not all women in the traditional sense are women in Haslanger's sense. But the account of concepts I have offered does not imply that the subject matter will change as a result. This is because it is consistent with a change in the extension of the term 'woman', underwritten by a change in linguistic practice, that the concept expressed by the term nonetheless secures the same subject matter: women. Indeed, I take this to be the most plausible understanding of what is going on in this and many other cases of conceptual engineering and conceptual ethics. My view applies equally well, for example, to debates surrounding the nature of race, gender, marriage, consent, property, rape, and personhood, each of which has a history of revisionary analyses grounded in substantive disagreement.[27]

[25] The concern is reminiscent of Strawson's objection to Carnap's views on conceptual explication on the grounds that conceptual explication implies a change in extension which amounts to a change in topic. See Carnap (1947) and Strawson (1963). My proposal is also, then, a response to Strawson on behalf of Carnap.

[26] This is recognized by Cappelen, who provides an alternative way to distinguish topic and extension. See Cappelen (2018). A discussion of the relative merits of the two approaches will have to await discussion on another occasion.

[27] For revisionary analyses of race, see for example Appiah (1992) and Haslanger (2012). See Clark and Chalmers (1998) for a revisionary analysis of belief.

In this sense, a revisionary analysis is not intended to describe our current linguistic practice but is intended as the analysis of the relevant subject matter that we *ought* to accept. This is because a proposed revisionary analysis is, in well-intentioned cases, an attempt, by those who take themselves to have more clearly conceptualized the subject matter, to initiate positive theoretical or social change.[28] In effect, the analysis contributes a revised characterization of the subject matter to the dialectic, thereby altering the status of the relevant term from one with a (relatively) stable meaning to one with a contested meaning, grounded in substantive disagreement about a subject matter. If the revisionary analysis is correct and accepted, the effect is to bring the extension of the linguistic meaning of a term in line with the extension of the concept it expresses (i.e., in line with the relevant subject matter); it moves linguistic practice closer to the truth. The analysis has to be both true and accepted to achieve this aim, since an unaccepted truth would not change linguistic practice, and an accepted falsehood would not bring linguistic practice closer to the truth. Conceptual engineering is, at its most interesting, the result of an attempt to uncover facts that we are partially aware of but have not yet fully grasped, whether mathematical, logical, philosophical, natural, social, or moral.[29] This does justice to the kind of normativity that underlies conceptual engineering projects by connecting it to questions about the way we ought to think and talk, as well as to questions about how we ought to act.[30]

Burgess and Plunkett have raised a general concern for externalist accounts of conceptual engineering, responding to which will help to clarify my view. They write:

> The textbook externalist thinks that our social and natural environments serve as heavy anchors, so to speak, for the interpretation of our individual thought and talk. The internalist, by contrast, grants us a greater degree of conceptual autonomy. One salient upshot of this disagreement is that effecting conceptual change looks comparatively easy from an internalist perspective. We can revise, eliminate, or replace our concepts without worrying what the experts are up to, or what happens to be coming out of our taps. (2013a: 1096)

The objection assumes that conceptual engineering involves revising, eliminating, or replacing our concepts. But conceptual engineering as I understand it does not involve revising our concepts. Indeed, there is an incoherence to the suggestion that it does, both for the externalist and the internalist. So long as we assume that concepts have their intensions and extensions essentially, it will not be possible for them to be revised. Nor does conceptual engineering involve eliminating or replacing our concepts. This would, I take it, amount to changing the topic of inquiry, which would not do justice to the interest and importance of the relevant projects in conceptual engineering. My account, of course, has two representational elements:

[28] I say 'in well-intentioned cases' because it is clearly possible for someone to propose, with evil intentions and for personal gain, a revised analysis which knowingly subverts the truth. The idea is explored through Orwell's use of the language 'Newspeak' in his (1949) novel *Nineteen Eighty-Four.* I restrict my focus to the kinds of examples I mention in the chapter, which, I take it, are all well-intentioned.

[29] For a mathematical example, see Frege's revisionary analysis of number in his (1884).

[30] The importance of the normative aspect of conceptual engineering is emphasized in Burgess and Plunkett (2013a,b).

concepts and linguistic meanings. And although conceptual engineering does not involve revising, eliminating, or replacing the former, it does involve revising, eliminating, and replacing the latter. In this sense, Cappelen is right when he says that conceptual engineering is about 'fixing language'; it is, as he says, about better ways of talking about a topic.[31] But there are two points to note here. First, talking about a topic requires a representational relation between us and the topic, and one which does not vary with the variation in our ways of talking. This is provided, I have argued, by concepts that are individuated in part by relations to objective properties in the world beyond us. Externally individuated concepts, then, provide the stable background against which conceptual engineering in the form of linguistic, and hence theoretical, change can take place. This means that a representational anchor to the world is an advantage of the account I have offered, and one which internalist theories cannot provide.[32] Second, conceptual engineering as I understand it is a form of theorizing. This presents no specific problem for the externalist account I have suggested, which incorporates a notion of linguistic meaning that floats free from the heavy anchors of the social and natural environments accepted by the 'textbook externalists'. A revisionary analysis, which can be proposed by any member of a linguistic community, will inevitably, in virtue of its revisionary status, go against the mainstream 'expert' opinion. The difficulty lies in persuading the mainstream of the merits of the revisionary analysis, but this is a difficulty that must be faced no matter which theory of representation is true.

7. Truth and Normativity

Having distinguished the linguistic meaning of a term from the concept expressed by that term, and hence the semantic content of a sentence from the thought expressed by that sentence, we need to separate the elements that are appropriate to language and the elements that are appropriate to thought.

One particularly important such property is truth.[33] I take truth, ultimately, to be a property of thought. The truth-value of a sentence, then, depends on the truth-value of the thought it expresses rather than on its (descriptive) semantic content. The two representational elements are governed by distinct norms: conceptual norms (norms of thought), and linguistic norms (norms of language). Conceptual norms concern truth. Linguistic norms, as stated in section 3 above, concern use in accordance with linguistic meaning at a time. For example, suppose that rape was standardly defined as possible only outside of marriage in 1800, but was standardly recognized (correctly) to be possible within marriage in 2000.[34] The sentence 'Rape can occur in marriage', then, expressed the same, true thought in 1800 as it did in 2000. But an utterance of the sentence in 1800 would have violated the linguistic norms of the time, whereas an utterance of the sentence in 2000 would have conformed to the

[31] Cappelen (2018).

[32] A general argument against internalism on these grounds is given in Sawyer (2007).

[33] For these purposes I do not take any specific stand on the nature of truth.

[34] For a history of substantive disagreement concerning rape, see Hasaday (2000).

linguistic norms of the time. This captures one aspect of the revisionary nature of conceptual engineering. A proposal that violates linguistic norms is bound to be regarded as revisionary from the perspective of the theoretical and linguistic practice at the time, even if the proposal correctly characterizes the subject matter.

A second such property is that of analyticity. I take analyticity to be a property of language. As such, a statement will be analytic not *per se*, but only relative to a linguistic practice at a time. For example, the sentence 'Rape can only occur outside of marriage' was plausibly analytic in 1800 but not in 2000. I agree with Haslanger, and for similar reasons to the ones she provides, that the sentence 'Bachelors are unmarried men', held up as the archetypical analytic statement in the philosophical classroom, is no longer analytic.[35] Quine was right when he said that no statement is immune to revision.[36] However, the dual-aspect theory of representation I have been suggesting allows us to distinguish analytic statements from conceptual truths in a way in which theories that identify linguistic meanings and concepts cannot. The status of a statement as analytic may be relative to a linguistic practice at a time, but conceptual truths are not; conceptual truths are eternal truths. The statements we accept as analytic, I suggest, are the statements that we take to express conceptual truths, and, of course, at the heart of a revisionary analysis is the claim that we erred in what we took the conceptual truths to be.[37]

The distinction between linguistic meanings and concepts also has implications for the practice both of reporting what is said from across a theoretical divide and of reporting what is thought from across a theoretical divide. The case of reporting what is thought is relatively straightforward given my account. For example, Tilly, in 2000, can utter the sentence 'In 1800 Abe believed that rape is not possible within marriage', thereby attributing to Abe in 1800 the belief that rape is not possible within marriage. The practice of ascribing propositional attitudes to others is secured if the ascriber and the ascribee have access to the same thought, which, on the view I have proposed, is not jeopardized by a difference in linguistic meaning. The case of reporting what is said is *prima facie* more complicated, precisely because indirect speech reports are reports of what was *said* rather than of what was *thought*. As such, they appear to concern linguistic meanings rather than concepts, where linguistic meanings, I have argued, are precisely what differ across theoretical divides. How, then, can Tilly in 2000 report what Abe in 1800 said when he uttered the sentence 'Rape cannot occur in marriage'?[38] Here we return to Frege's initial claim that sentences express thoughts. Given this, we can say that Tilly's report of what Abe said is true in virtue of the fact that it reports Abe as having uttered a sentence that expressed the thought that rape cannot occur in marriage. And this, we are assuming, is true.

[35] See the discussion in Haslanger (2006: section VII).

[36] See Quine (1951).

[37] The theoretical benefits of the distinction between analytic statements and conceptual truths will have to be explored in more detail on a future occasion.

[38] Saul (2006: 141) raises a version of this objection against a contextualist interpretation of Haslanger (2000).

8. Conclusion

In this chapter, I have urged the merits of a dual-aspect theory of representation that distinguishes the linguistic meaning of a term from the concept expressed by that term. Linguistic meaning, I suggest, is determined by patterns of actual use and tracks a community's understanding of a subject matter. Concepts, in contrast, are determined by real relations to objective properties and secure a subject matter about which there might be different understandings and different proposed analyses. I have argued that the distinction brings theoretical gains in a number of related areas. There is more to be said, but if I am right about the theoretical gains, then the view is worth taking seriously.

References

Appiah, K. A. 1992. *In My Father's House: Africa in the Philosophy of Culture.* Oxford: Oxford University Press.

Burge, Tyler. 1973. Reference and Proper Names. *Journal of Philosophy* 70 (14):425–39.

Burge, Tyler. 1986. Intellectual Norms and Foundations of Mind. *Journal of Philosophy* 83 (12):697–720.

Burge, Tyler. 1989. Wherein is Language Social? In Alexander George (ed.), *Reflections on Chomsky* (pp. 175–92). Oxford: Blackwell.

Burgess, Alexis, and Plunkett, David. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, Alexis, and Plunkett, David. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–11.

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering.* Oxford: Oxford University Press.

Carnap, Rudolf. 1947. *Meaning and Necessity. Meaning and Necessity: A Study in Semantics and Modal Logic.* Chicago: University of Chicago Press.

Chalmers, David. 2003. The Nature of Narrow Content. *Philosophical Issues* 13:46–66.

Chalmers, David. 2011. Verbal Disputes. *Philosophical Review* 120 (4):515–66.

Clark, Andy, and Chalmers, David. 1998. The Extended Mind. *Analysis* 58 (1):7–19.

Crane, Tim. 1991. All the Difference in the World. *Philosophical Quarterly* 41 (162):1–25.

Fodor, Jerry. 1980. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioral and Brain Sciences* 3:63–73.

Fodor, Jerry. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: MIT Press.

Frege, Gottlob. 1884. *Die Grundlagen der Arithmetik: Eine logisch-mathematische Untersuching uber den Begriff der Zahl.* Breslau: Verlage Wilhelm Koebner. Trans. 1950. J. L. Austin as *The Foundations of Arithmetic: A Logico-mathematical Enquiry into the Concept of Number.* Oxford: Basil Blackwell.

Frege, Gottlob. 1892. On Sense and Reference. Trans. 1948. In *Philosophical Review* 57 (3):209–30.

Frege, Gottlob. 1918. The Thought: A Logical Inquiry. Trans. 1956. In *Mind* 65 (1):289–311.

Gallie, W. B. 1956. Essentially Contested Concepts. *Proceedings of the Aristotelian Society* 56:167–98.

Hasaday, Jill. 2000. Contest and Consent: A Legal History of Marital Rape. *California Law Review* 88 (5):1373–505.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Noûs* 34 (1):31–55. Reprinted in Haslanger (2012).

Haslanger, Sally. 2006. What Good Are Our Intuitions? *Aristotelian Society* Supplementary Volume 80 (1):89–118.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique.* Oxford: Oxford University Press.

Kaplan, David. 1989. Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives. In J. Almog and J. Perry (eds.), *Themes from Kaplan*. Oxford: Oxford University Press.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.

Orwell, George. 1949. *Nineteen Eighty-Four.* London: Secker & Warburg.

Plunkett, David, and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Putnam, Hilary. 1973. Meaning and Reference. *Journal of Philosophy* 70:699–711.

Putnam, Hilary. 1975. The Meaning of "Meaning". *Minnesota Studies in the Philosophy of Science* 7:131–93.

Quine, W. V. O. 1951. Two Dogmas of Empiricism. *Philosophical Review* 60:20–43.

Sainsbury, Mark. 2014. Fishy Business. *Analysis* 74:3–5.

Saul, Jennifer. 2006. Gender and Race. *Aristotelian Society* Supplementary Volume 80 (1):119–43.

Sawyer, Sarah. 2007. There Is No Viable Notion of Narrow Content. In B. McLaughlin and J. Cohen (eds.), *Contemporary Debates in the Philosophy of Mind* (pp. 20–34). Oxford: Blackwell.

Sawyer, Sarah. 2010. The Modified Predicate Theory of Proper Names. In S. Sawyer (ed.), *New Waves in Philosophy of Language* (pp. 206–25). New York: Palgrave MacMillan Press.

Sawyer, Sarah. 2018. The Importance of Concepts. *Proceedings of the Aristotelian Society* 118 (2):127–47.

Segal, Gabriel. 2000. *A Slim Book about Narrow Content*. Cambridge, MA: MIT Press.

Stalnaker, Robert. 1990. Narrow Content. In C. A. Anderson and J. Owens (eds.), *Propositional Attitudes: The Role of Content in Logic, Language and Mind*. Stanford: CSLI Press.

Strawson, Peter. 1963. Carnap's Views on Conceptual Systems versus Natural Languages in Analytic Philosophy. In P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap* (pp. 503–18). London: Open Court.

Sundell, Timothy. 2011. Disagreements about Taste. *Philosophical Studies* 155 (2):267–88.

Sundell, Timothy. 2012. Disagreement, Error, and an Alternative to Reference Magnetism. *Australasian Journal of Philosophy* 90 (4):743–59.

19

Philosophy as the Study of Defective Concepts

Kevin Scharp

For the past decade or so, I have tried to make sense of the liar paradox, which is a 2,300 year old problem associated with truth. It goes like this. Think of a sentence that says that that very sentence is not true. Is it true or not true? No matter what answer you give, it is easy to derive a contradiction in just a few steps. The view that I developed is that the liar and the other paradoxes to which truth gives rise are symptoms of an underlying defect in the concept itself. It's not that the reasoning in the paradox involves some trivial mistake or faulty assumption. It is that the concept of truth itself is to blame for these paradoxes that we have found ourselves in for a very long time now. We might say that when we reason to the contradiction in the paradox, we are using all our concepts according to rules that are built into those concepts.

The second half of project is to replace the concept of truth for certain purposes. Truth is a defective concept, and there are certain jobs and it's not very good for. We should replace it with a team of concepts that together can do its job without giving us any of the paradoxes or problems that plague the concept of truth. The job that I really focus on is explaining the meanings or contents of natural language sentences by way of natural language semantics. A very popular form attributes truth conditions to sentences of natural languages. The paradoxes that truth generates mean that it can't do that job very well at all. Anytime you try to use truth to give a semantics for a natural language like English, or really any expressively rich language at all, you end up contradicting yourself. You end up saying things that are inconsistent. And it's the paradoxes that force this upon us.

The replacement concepts, which I call ascending truth and descending truth, can do this job perfectly, and the resulting theory agrees with traditional semantics as a special case everywhere the latter provides coherent results. So it is a lot like advances in science, where the successor theory does everything that the earlier theory did, and solves some extra problems as well.[1]

I call the method followed in this project of replacing truth *conceptual engineering*. I take conceptual engineering to be actively changing some aspect of our concepts—eliminating bad ones, deciding which ones we should use, and which word should

[1] Scharp 2013.

Kevin Scharp, *Philosophy as the Study of Defective Concepts* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Kevin Scharp.
DOI: 10.1093/oso/9780198801856.003.0019

express them. Although there are plenty of instances of conceptual engineering in the history of philosophy, it hasn't really been a focus of attention; I borrowed the term from a comment by Simon Blackburn in his little introductory book called *Think*.[2] The idea of conceptual engineering is really taking an active role with respect to our conceptual scheme and changing it when one finds defects in those concepts. In earlier work, I argued at length that the concept of truth does have this kind of defect and is ripe for replacement.

I'm not going to develop that project; instead what I'm going to do is introduce something that I've come to believe as a result of engaging in it. I've come to think that conceptual engineering can and should play a much larger role philosophical theorizing. Indeed, I've come to think that most, if not all, commonly discussed philosophical concepts are inconsistent. Some in the way the truth is inconsistent, others in more subtle ways. Some based on the way they interact with one another, and others just by themselves. As such, I have come to think that philosophy is for the most part the study of what have turned out to be inconsistent concepts.

One way to make sense of this idea of an inconsistent concept is to say that concepts have constitutive principles—principles that constitute that concept in the sense that they tell you which concept it is, and often those constitutive principles are inconsistent with one another or with obvious facts about the world. The concepts that I think are inconsistent include truth, knowledge, nature, meaning, virtue, explanation, essence, causation, validity, rationality, freedom, necessity, person, beauty, belief, goodness, space, time, and justice. So when I say I think that philosophy is for the most part the study of inconsistent concepts, that's really what I mean. I think those are all inconsistent concepts; those are all defective concepts.

1. The Radical Therapeutic Program

I want to stay a little bit about the role that I think conceptual engineering should play in philosophical methodology, and I in order to do that I want to paint a picture of what I think philosophy is like—an account of the nature of philosophy. One way to do this is to appeal to some folks from history of philosophy, namely, Socrates, Nietzsche, and Wittgenstein.

From Socrates, and by 'Socrates' here I mean the early Platonic Socrates, the one from the early dialogues of Plato, we get the idea that the unexamined life is not worth living.[3] I take it he means the life without critical thinking. Subjecting ones' beliefs to critical scrutiny is the key aspect of critical thinking. So if your life is without critical thinking, then it's not worth living. Critical thinking is an essential aspect of living the good life—being in the right way.

Nietzsche. The idea from him is that in the absence of any divine or objective standards for human life we ought to craft our own.[4] That is, you want to take an active attitude toward your own life—for the creation of the structure of one's own life.

And from Wittgenstein, I take the idea that philosophical problems are manifestations of being trapped by our language, and philosophy should take the form of

[2] Blackburn (1999). [3] Plato (1961). [4] Nietzsche (1886).

therapy that ultimately dissolves the philosophical problems.[5] The key claim here is that the aim of philosophy is to show fly the way out of the fly bottle. That's a nice metaphor that Wittgenstein used, and I'm going to take that metaphor pretty seriously.

Conceptual engineering is taking a Socratic, that is, critical, and Nietzschean, that is, active, attitude toward one's own conceptual scheme. Right now, many of us think that we should already take this attitude toward our beliefs. For example, we should subject our beliefs to a battery of objections, see how well we can reply to those objections, and if a belief doesn't fare well in this process, then that's a good indicator that it should be changed. And by doing this—by subjecting one's beliefs to critical scrutiny—one can craft and sculpt and mold a system of beliefs for oneself, You can do that rather than just doing what most other people do, which is borrowing a set of beliefs from whoever raised you and living throughout your whole life without really thinking about them very much. It turns out that the beliefs you borrow from your ancestors might work pretty well today, but there are also going to be some places where they don't work very well.

I think we should take this same attitude toward our concepts. The central idea of conceptual engineering is that we ought to take the same critical attitude toward our concepts. Likewise, if a concept doesn't fare well under critical scrutiny, the active attitude kicks in and one crafts new concepts to do work wants without giving rise to the problems inherent in the old ones. By doing this, one can sculpt and craft a conceptual repertoire of one's own, rather than just living one's life with the concepts borrowed from one's ancestors. Conceptual engineering is doing to concepts what most of us already think we should be doing to our beliefs. A nice quote from Alexis Burgess and David Plunkett: "Our conceptual repertoire determines not only what we only what we can think and say, but also, as a result, what we can do and who we can be." Through conceptual engineering, we can take some control over what we can think, say, do, and be.

Conceptual engineering can been seen as in the service of an overarching therapeutic program in the spirit of Wittgenstein. However, Wittgenstein's infamous conservativism, that philosophy should leave everything as is, has no part in this project at all. Our beliefs are not fine as they are. Our concepts are not fine as they are. But we can make them better.

The radical therapeutic program does share with Wittgenstein's methodology the goal of showing the fly the way out of the fly bottle. And conceptual engineering can help. Consider the thesis that philosophy is the study of what have turned out to be inconsistent concepts. Put this idea into a Wittgensteinian program, and you get the following picture. Philosophers are arguing about how best to make sense of concepts that are actually inconsistent. We're trying to figure out how to analyze concepts that are internally defective. The arguments one finds in philosophy journals and books rely on privileging certain constitutive principles here and others there, but ultimately the debates rarely make discernible progress because the concepts being analyzed and the concepts used to conduct the debate are defective. That is one reason

[5] Wittgenstein (1953).

philosophers end up dealing with so many paradoxes and conceptual puzzles. And it can sometimes seem like that is our whole job.

That's the fly bottle. That's where you are and where I am. Trapped by our most important and cherished concepts into accepting absurdities and reasoning our way to contradiction.

How do we escape? For the past 400 years, the domain of philosophy has been shrinking. That is a sociological fact. Physics, geology, chemistry, economics, biology, anthropology, sociology, meteorology, psychology, linguistics, computer science, cognitive science. Each of those subject matters was a part of philosophy a mere 400 years ago. And we are talking about a discipline that has 2,600 year history; 400 years in 2600 years is nothing. As the scientific revolution ground on, more and more sciences were born. This process is essentially philosophy *outsourcing* its subject matter as something new. As sciences. The process is rather complicated, but an important part of it is getting straight on right concepts to use. That subject matter then gets brought under scientific methodology.

Showing the fly the way out of the fly bottle is taking an active role in this outsourcing process—pushing faster and faster to identify the conceptual defects (the Socratic idea), craft new concepts to avoid the old defects (the Nietzschean idea) with an eye toward repairing the philosophical subject matter for outsourcing as a science.[6] The ultimate goal of the process is the potential end of philosophy. Escape for the fly. The end of philosophy is only potential because it is likely that new technologies and ways of life will give us new inconsistent concepts that are philosophically significant, and these will need to get sorted out as well. So it is not obvious that our stock of defective concepts is ever going to effectively decrease. It really depends on how much conceptual engineering occurs. Speeding it up is up to us. The speed with which we get new defective concepts is mostly *not* up to us. People just make them up as they are needed or wanted. Nevertheless, we can envision a world where we have succeeded in making philosophy evaporate. Sometime after that, it might show up again with new philosophically significant defective concepts. And after that, philosophy might break out during especially rapid technological or social growth. Somewhat like acne.

That's the idea I want to call *the radical therapeutic program*. It calls for actions that might ultimately do away with philosophy.

The scientific element in this radical therapeutic program, which I call *metrological naturalism*, is separable from the conceptual engineering element. However, the two go together well. Metrological nationalism is more successful with consistent concepts, and in order to do conceptual engineering, we need to know what kind of replacement concepts to aim for. One might even say that metrological naturalism without conceptual engineering is empty, and conceptual engineering without metrological nationalism is blind.

Contrast this radical therapeutic program that I've just outlined with the most prominent philosophical methodology of our time, philosophical analysis in general,

[6] Contrast this position with Capellan's recent work in which he argues that we have little or no control over the concepts we use.

and the Canberra Plan in particular, which owes much to David Lewis. According to Lewis' methodology, one begins by assembling the platitudes for a philosophical term, and then one tries to figure out which real, relatively fundamental thing the platitudes might describe. If the platitudes are inconsistent, one tries to find something that satisfies a weighted majority of them—try to figure out what comes the closest to satisfying them and that is what the philosophical term designates. That's it.

The dominant, Lewisian methodology is static, and so has nothing to do with change and improvement. Lewis writes,

> One comes to philosophy already endowed with a stock of opinions. It is not business of philosophy either to undermine or to justify these pre-existing opinions to any great extent, but only to try to discover ways expanding them into an orderly system.[7]

I think it is hard to be more wrong than that.

2. Conceptual Engineering

There are several conceptual engineering projects already in philosophy, and I want to lay out two of them. One kind of project is labeled *amelioration* by Sally Haslanger. She argues that we need to change various key terms, especially terms associated with gender and race, for social justice reasons.[8] For example, 'woman,' the word, seems to currently express the concept of an adult human female. However, Haslanger argues that 'woman' is used primarily to subordinate people based on their stereotypical female characteristics. She suggests that the word 'woman' should instead be used to express something like the concept of a person subordinated based on stereotypical female characteristics. She wants to change the word 'woman,' and by doing that, the goal is to fight that subordination by making it explicit in the concept expressed by 'woman.' Rather than having it be something implicit, she suggests making the subordination explicit so that is something that is right there in front of one's face. Then, if the word 'woman' is used in this new way, a supporter of social justice can have as a goal the elimination of women—the elimination of people who are subordinated by appeals to their stereotypical female characteristics. Haslanger's amelioration project is obviously a conceptual engineering project, and there are clear similarities between her project of mine.

There are conceptual engineering projects in contemporary metaphysics as well. For example, Theodore Sider's introduction of the concept of structure as a generalization of David Lewis' notion of naturalness.[9] Another conceptual engineering project in this area concerns the idea that some or all metaphysical disputes are pointless. Metaphysics has come under attack lately, and one of the major criticisms is that metaphysical disputes are merely verbal, where the participants are just talking past one another rather than having a substantive disagreement. For example, how many things there are in a universe with two simple objects (i.e., objects with no proper parts). Are there two things in that universe or are there three things (the two simple things plus the one complex thing made up of the two simple things as parts)?

[7] Lewis (1973: 88).

[8] Haslanger (2000).

[9] Sider (2011).

One view is that two people engaged in that dispute are just talking past one another—the dispute is merely verbal, and so not worth having. It is a merely verbal dispute because some key term used in it—maybe the word 'thing' or maybe the quantifier 'there is'—means different things for the people involved. Those who advocate this position are called metaphysical deflationists, and they employ some sophisticated tools with which to formulate and defend this criticism. Probably the most well-known tool is quantifier variance, which is the idea that there are multiple equally good interpretations of what people mean by the existential quantifier, 'there is,' involved in formulating these metaphysical questions. Those engaged in the ontological disputes—the metaphysicians—are simply talking past one another according to this criticism.

The metaphysicians have strong objections to this criticism, but they also have proposed a new kind of strategy for conducting metaphysical disputes just in case the metaphysical deflationists turn out to be right. The strategy is called *Plan B* by Sider who is one of its primary advocates. Plan B is to give up using natural languages like English for doing ontology, for doing metaphysics, and instead stipulate a fundamental meaning for existential quantifiers in a new language often called *Ontologese*. According to Plan B, metaphysical disputes can then be conducted using this new language and its new existential quantifier, which is stipulated to be fundamental. That is obviously a conceptual engineering project. This project is less focused on improving our conceptual scheme in light of discovering one of our concepts is defective, and it is more focused on avoiding criticism by talking in a new way. Still, it is a conceptual engineering project.

3. Metrological Naturalism

There are two aspects to the radical therapeutic program outlined above: conceptual engineering and metrological naturalism. Let us turn to metrological naturalism. *Metrological* naturalism is a kind of *methodological* naturalism, which is a somewhat unpopular philosophical methodology these days. Methodological naturalism is captured by the following quote from Wilfrid Sellars. "In the dimension of describing and explaining the world, science the measure of all things."[10] Moreover, scientific methods are the most reliable route to true beliefs despite scientific results being fallible. Finally, philosophy should be continuous with the sciences in two senses: (i) sciences do not require justification or grounding from philosophy, and (ii) philosophy and the sciences should pursue similar goals and employ similar methods. Philosophers have typically for a very long time now taken it as our job to assess whether scientists are spending their time productively, and we often assumed that science is in need of some kind of justification from first principles; that is, philosophers have historically taken sciences to be ungrounded unless they have been provided a grounding from philosophy. I suppose this isn't surprising given the philosophical sources of all the sciences, but that is no reason to retain it, and methodological naturalism rejects it. Moreover, methodological naturalism

[10] Sellars (1956).

enjoins philosophers to look to the sciences for methods and goals. Beyond that vague advice, there isn't much agreement about how to be a methodological naturalist.

Here is a problem for methodological naturalism: it seems like many of the things that the naturalist says are not scientific. For example, no science concludes that science is the best road to truth about the physical world. That is not a scientific result. Another problem is that there is no consensus about what demarcates science from non-science. The demarcation problem is notorious as a standard example of a problem in philosophy that has received tons of attention but resists any solution. And there is no consensus on the nature of scientific methodology either. Obviously there is observation, hypothesis, prediction, and experiment, but beyond those kinds of banalities, it is hard to say exactly what the scientific method consists in, and it is not exactly obvious that someone who is working at the Large Hadron Collider, the particle accelerator, is following the same kind of methodology as a sociologist who studies rural food distribution networks or as a biologist who is investigating fetal development in marine mammals. Also, many philosophical topics are abstract and so resist scientific methods, which tend to emphasize causal interactions. And finally, scientific methods aim for descriptive results, but many philosophy philosophical topics are normative.

The version of naturalism that I want to advocate is called *metrological* naturalism because the Greek word for measure is '*μέτρον*' (*metron*). My version is a *measurement-theoretic* methodological naturalism, which we can therefore call *metro*logical naturalism.

The big idea from scientific revolution is that we can use mathematics to describe, explain, predict, and control the world around us. That is the idea that comes out in Galileo's very nice formulation, "mathematics is the language of nature."[11] It is since the early 1600s that we have had this idea, but it was not really until the 1800s, in fact that late 1800s, that theorists turned their attention to providing a *scientific* understanding of how mathematics is applied in this way in the sciences. The result of these investigations is called *measurement theory*: the study of how mathematics applies to the natural world in the sciences. You can think of measurement theory as an all-purpose foundation for scientific theorizing in much the same way that set theory can be thought of as an all-purpose foundation for mathematics. Metrological naturalism has as a methodological principle that philosophers should use measurement theory as a guide or model in philosophical theorizing.

According to this methodology, we should be using resources of the sciences in philosophical theorizing. There are many ways of doing this, but let me just briefly present three of them. First, cast your philosophical theories of X, where X is whatever philosophical concept you want to be thinking about, as measurement systems for X. There is a large literature on how to construct measurement systems, and we know pretty well how to do this for things like length and weight. Trying to figure out how to construct a measurement system for something like truth or justice is a lot more complicated, but this isn't just an analogy.

[11] Galilei (1623) as translated in Popkin (1966: 65).

For example, Donald Davidson, over the course of a famous career, crafted a detailed measurement system for belief, desire, and meaning.[12] He called it the unified theory. In the case of length, we start with people's judgments about which of two objects *extends beyond* the other when the two are lined up, and thinking of two objects *end to end* as a single thing having a length just like any regular object. We then translate these judgments into a different language—a language with terms like 'longer than' and 'concatenation.' Call it the relational language. If object A extends beyond object B, then A is longer than B. If A is end to end with B then there is an object C that is identical to the concatenation of A and B. One benefit of this translation is that 'longer than' and 'concatenation' are very clearly behaved with explicit laws defining them, whereas the judgments we started with are pretty messy. From here, the next step is to prove a certain kind of result called a representation theorem, which says that we can translate from the language with 'longer than' and 'concatenation' into a mathematical language, which contains 'greater than' and 'plus' and numerals as well. Proving the representation theorem shows that you can use the numbers talked about by the mathematical language to keep track of the objects talked about by the relational language. Each object gets a number, and the number assigned to A is greater than the number assigned to B if and only if A is longer than B. And the number assigned to A plus the number assigned to B equals the number assigned to the concatenation of A and B. The representation theorem says that you can assign numbers to the objects in a way that makes all these principles, and a bunch of other ones, true. Another kind of result, called a uniqueness theorem, says how many different ways there are of assigning numbers like this. Overall, the measurement system begins with basic judgments about one object extending beyond another and putting two objects end to end. And the measurement system delivers something remarkable! A number for every direction of every object so that we can use these numbers to do everything from carpentry to identifying the distances to distant galaxy clusters. And we also get a plethora of length scales from stadia to megaparsecs. All of that comes from making some basic assumptions about 'longer than,' 'concatenation,' and how to translate them into basic English (extends beyond and end to end) and into mathematics (greater than and plus).

In Davidson's work, the same kind of measurement system is laid out for belief, desire, and meaning. He makes some assumptions about the how people think about sentences that they think are true, and some assumptions about people's preferences. Together these allow him to show how to assign truth conditions to the sentences of the language in question and beliefs and desires to the person in question. In his demonstration, Davidson utilizes a fictional character called the radical interpreter, who goes through a certain procedure to figure out meanings, beliefs, and desires, from holding-true and preferences. This procedure is showing a representation theorem for the measurement system. The same idea is worked out in considerably more detail in Robert Matthews' book *The Measure of Mind*.

[12] Davidson (1990).

So the first bit of advice from metrological naturalism is to cast one's philosophical theories as measurement systems in much the same way that Davidson and Matthews do. Second, focus on semantic theories of philosophical locutions rather than trying to analyze philosophical concepts. Philosophers have been, since the early twentieth century, trying to analyze concepts. Even today the majority of talks at most major conferences are dedicated to philosophical analysis projects, where the philosopher is attempting to put forward a philosophical analysis of some concept like obligation or beauty or reason.

Instead of doing philosophical analysis or conceptual analysis we should be focused on semantics for philosophical terms, because arriving at an adequate semantic theory for a philosophical term can often cut through many of the unnecessary and confusing assumptions associated with that term. One example here would be the semantics for reasons project I have jointly undertaken with Brian Weaver.[13] Getting straight on the semantics for reasons locutions helps tremendously in assessing traditional philosophical debates about reasons. For example, there is a debate between factualists, who say that reasons are facts, and mentalists, who say that reasons are mental states. Understanding the proper semantics for reasons locutions exposes that debate as not very substantive.

Third, utilize the tools of measurement theory for answering philosophical questions. For example, invariance and symmetry play major roles in measurement theory, and they can be used to make sense of the philosophically significant distinction between objective and subjective phenomena. Consider the claim, 'In Scotland, it is twice as hot in May as it is in March'; one might feel compelled to assert this sentence because the average temperature in May is around 10°C and the average temperature in March is around 5°C. However, if we transform these values from the Celsius system to the Farenheit system, we get: 50°F in May and 41°F in March. But of course 50 isn't twice 41. The lesson is that one *can* of course multiply temperature measurements, but multiplication with temperature measurements isn't objective. When one says that it is twice as hot in Scotland in May as it is in March, one is talking about one's way of representing the world (using the Celsius scale) rather than talking about the world itself. Here the key to identifying objective features of the world is invariance. Something is an *objective* feature of temperature if it is invariant across different scales, whereas something is a *subjective* feature of how we measure temperature if it fails to be invariant. In contrast to the case of temperature, multiplication is an objective feature of length because it is invariant across length scales; for example, 20 feet is twice 10 feet, and 6.096 meters is twice 3.048 meters.

It is worth highlighting some features of metrological naturalism:

(i) It says nothing about the methods one uses to arrive at or justify philosophical theories. So metrological naturalism is not opposed to *apriori* methods (e.g., intuitions, deductions) in philosophy.

(ii) It is not opposed to *apriori* or theoretical philosophical claims. In fact, it is plausible to think that certain aspects of a measurement system for some concept will be constitutive of that concept.

[13] Weaver and Scharp (forthcoming).

(iii) It does not offer analyses of concepts or necessary and sufficient conditions for concept application.
(iv) It does not offer reductive explanations in any sense.
(v) It is not a metaphysical thesis about what exists or does not exist.
(vi) It does not need a leading science, a criterion for what demarcates science from non-science, or an account of scientific methodology.
(vii) It is applicable to abstract topics (e.g., one can do measurement theory for mathematics and logic).
(viii) It is applicable to normative topics (e.g., formal axiology).

There is much more to be said for metrological naturalism both as an independent methodology and as a companion to conceptual engineering, but the above will have to do for this occasion.

4. Elements of Conceptual Engineering

Conceptual engineering is actively changing some aspect of our concepts—eliminating bad ones, adding new ones, deciding which ones we should use for which purposes, and choosing which words should express which concepts. There are plenty of instances of conceptual engineering in the history of philosophy, and we have considered a couple above.

Another term that is used often in this area of research is 'conceptual ethics,' but the two are distinct. *Conceptual ethics* is the study of evaluative and normative issues associated with our concepts and the words that express them. *Evaluative* issues are those pertaining to how good something is; for example, oxygen is a better concept than phlogiston, and the luminiferous ether is not a good concept. *Normative* issues are those pertaining to obligations and permissions—to what we ought to do and what we may do; for example, we ought to use the word 'woman' to express the concept of someone oppressed on the basis of stereotypical female characteristics, and we ought not use the concept of mass from Newtonian mechanics when calibrating the atomic clocks on GPS satellites.

Conceptual ethics is clearly involved in conceptual engineering because the latter often relies on evaluative and normative judgments about our concepts. However, not all conceptual ethics is conceptual engineering. For example, judging that our concept of mental illness is just fine for our purposes is doing conceptual ethics, but it isn't doing conceptual engineering. And the converse holds as well. For example, establishing a relative consistency proof for an axiomatic theory of ascending truth and descending truth, which have been suggested as replacements for our defective concept of truth, is a part of conceptual engineering, but it isn't a part of conceptual ethics. Hence, I do not see the two terms as competitors for describing a single area of philosophy.

Among conceptual engineering projects, two kinds deserve to be singled out as significant and distinct: conceptual revision vs. conceptual replacement. *Conceptual revision* is changing a concept so as to improve it in some way, but the concept persists through whatever changes happen to it. It is the same concept before and after the revision to it. I think some people conceive of Carnap's method of explication as a kind of conceptual revision by adding a degree of clarity to an otherwise fuzzy concept. Sally Haslanger is often read in this way, but it is isn't obvious that it is the most accurate interpretation.

On the other hand, *conceptual replacement* doesn't cause any changes to any concepts at all; rather, these projects introduce new concepts to our conceptual scheme and prescribe a particular role for them to play; this role might already be filled by one of our existing concepts, so a replacement project might suggest that one of our existing concepts isn't cut out for one of the jobs we think it can do. Conceptual replacement is the kind of conceptual engineering project I take up with respect to the concept of truth in my book *Replacing Truth*. There I argue that truth is an inconsistent concept, and I offer two replacement concepts that, together, will do some of the work we have been using truth to do. Note that replacement does not entail elimination—we still retain the defective concept of truth, because in the vast majority of situations, we can use it without running into any trouble whatsoever. However, we do eliminate one or more roles for truth to play once we have the replacement concepts.

To illustrate the distinction between conceptual revision and conceptual replacement, we can think about three distinct readings of Haslanger's conceptual engineering project for the concept expressed by 'woman.' The variance is due to the fact that Haslanger both presents her project as actively choosing to do something to the concept expressed by the word 'woman' and appeals to a particular thesis in philosophy of language. She even suggests that 'woman' right now expresses one concept—the concept of being an adult human female—but that we should, for the purposes of social justice, choose to make it express a different concept—the concept of an adult human who is oppressed on the basis of stereotypical female characteristics. She also sometimes suggests that instead of picking a new concept, we are changing the existing concept expressed by the word 'woman.' Moreover, Haslanger also appeals to a controversial, but highly discussed thesis in the philosophy of language and mind called *semantic externalism*, which entails that the concept one uses to think or the concept that is expressed by a certain word is, to some extent, determined by the physical or social environment of the person thinking or the person using that word. As a result, it could be that 'woman' already expresses the concept of an adult human who is oppressed on the basis of stereotypical female characteristics. It might already express that concept because that is the role the word 'woman' plays in our social structures, whether we know it or not. So, Haslanger's project involves an element of conceptual engineering—choosing for our word 'woman' to express a certain concept—and it sure seems like it doesn't express that concept right now. And her project involves a commitment to semantic externalism as well, which to some extent, takes the control over which concepts our words express away from us. The three distinct readings of what Haslanger is up to are based on privileging these different aspects of her project.

1. A *conceptual revision* project. Prior to Haslanger's work, the English word 'woman' expressed the concept of an adult human female, but Haslanger suggests we ought to change this concept so that its content is *an adult human oppressed on the basis of stereotypical female characteristics*. That is,

Haslanger enjoins us to change our concept of woman so that it has a slightly different content. Although Haslanger sometimes talks as if these are distinct concepts, what she means is that they are distinct contents or readings of the same concept, before and after a change in that concept.

2. A *conceptual replacement* project. Prior to Haslanger's work, the English word 'woman' expressed the concept of an adult human female, but Haslanger suggests that we ought to change which concept is expressed by this word 'woman' so that it expresses the concept of an adult human who is oppressed on the basis of stereotypical female characteristics. That is, Haslanger enjoins us to change which concept is expressed by the word 'woman' so that it expresses a different concept. Although Haslanger sometimes talks as if these are the same concept, what she means is that they are distinct but similar concepts.
3. A *belief replacement* project. Prior to Haslanger's work, the English word 'woman' expressed the concept of an adult human who is oppressed on the basis of stereotypical female characteristics. However, no one really believed that it expressed this concept. Haslanger suggests we ought to change our beliefs about which concept is expressed by the word 'woman,' so that we stop believing it expresses the concept of an adult human female and start believing that it expresses the concept of an adult human who is oppressed on the basis of stereotypical female characteristics. That is, Haslanger enjoins us to change what we believe about the concept expressed by the word 'woman' so that our beliefs are true instead of false. Although Haslanger sometimes talks as if we should change the concept expressed by the word 'woman,' what she means is that we should have true beliefs about the concept expressed by the word 'woman.'

The third reading is the one that results from emphasizing semantic externalism over conceptual engineering. The first two projects result from emphasizing conceptual engineering over semantic externalism, and they differ on whether Haslanger offers a conceptual revision project in reading one or a conceptual replacement project in reading two. For what it is worth, I prefer the project described in reading number two, but it doesn't leave much room for semantic externalism to play a role in the account. Moreover, the standard term for the study of how rational agents change their belief systems is belief *revision*, not belief *replacement*. I've chosen the latter to be consistent with my distinction between conceptual revision and conceptual replacement. Belief revision, according to my usage, would be somehow changing a single belief so that it was still the same belief but was somehow different in content. This sort of thing is impossible according to most analytic philosophers because we tend to individuate beliefs by their contents.

5. Tools for Conceptual Engineering

There are many approaches one might take in conceptual engineering. Here I canvass three. Each of these has something to be said for it, and they need not be competitors. Nevertheless, they do have their own problems and pitfalls, so it helps to be clear about which of these tools one is using. Each of these tools is essentially a standard of *evaluation*—a way of assessing concepts that is relevant to potential conceptual engineering projects. As such, they all belong to conceptual ethics as well, and to the evaluative branch of conceptual ethics in particular.

The *metaphysical approach* is to evaluate our concepts for naturalness or fundamentality—how well they carve nature at its joints. One can see this approach in Sider's recent book, *Writing the Book of the World*. There he advocates fundamentality as the primary epistemic virtue associated with concepts. Usually naturalness or fundamentality are taken to come in degrees, so we can speak of one thing being more fundamental than another and of something being relatively fundamental. Presumably, Sider would accept relative fundamentality as a basic evaluation of concepts—other things being equal, more fundamental concepts are to be preferred.

The *pragmatic approach* is to evaluate our concepts for how well they work. How well does a concept do what we use it to do (or what we ought to use it to do)? This seems to be the genus for ameliorative projects like those advocated by Haslanger, and I often cast my truth project in these terms. Overall, talk of jobs or purposes for concepts is no more heavyweight than talk of things we do with concepts. For example, we formulate truth conditional semantic theories for gradable adjectives (like 'tall'), and many of the most powerful of these semantic theories have the word 'true' in them in way that doesn't seem to be eliminable. From this we can conclude that one job of the concept of truth is serving an explanatory role in these semantic theories.

The *constitutivity approach* is to identify the constitutive principles for a concept and then evaluate those principles using our toolkit for evaluating beliefs. Are they true? Are they justified? Are they helpful? On my view, constitutivity is a pragmatic issue. A constitutive principle for a concept is a principle that is used to guide interpretation—if my interlocutor rejects a principle that I take to be constitutive for a concept that figures in our conversation, then that is a *pro tanto* reason to think that we do not mean the same thing by the word in question. Constitutivity is a descendent of analyticity, but there is no reason to think that constitutive principles should be true or vacuous or uninformative or *apriori*. Moreover, one can reject a principle constitutive of a certain concept without thereby losing possession of that concept. By focusing on constitutivity, we can characterize individual concepts as inconsistent if their constitutive principles are inconsistent with established facts (e.g., truth), and we can characterize mutually inconsistent groups of concepts.

The constitutive approach also allows us to model concept change using the elaborate and influential tools of formal epistemology on belief change (e.g., AGM (Alchourrón, Gärdenfors, and Makinson) theory; Alchourrón et al. 1985). These are nice tools and I think they can help us identify conceptual clashes and evaluate proposed conceptual engineering projects. Some examples: (i) one can think of getting rid of a pejorative concept as contraction (belief elimination), (ii) we can define the entrenchment of a concept by the average entrenchment of its constitutive principles, (iii) belief replacement is the adding of a new belief to the set, and that is akin to what happens in conceptual replacement, which is the particular kind of conceptual engineering that I have been pursuing, and (iv) consolidation is an operation on an inconsistent belief base, and this can be thought of as identifying potential replacements in a conceptual engineering.

6. Constraints on Conceptual Engineering

I have laid out the radical therapeutic vision of what philosophy is all about. I have presented conceptual engineering as one aspect of a proper philosophical methodology. We should take an active role in altering and improving our conceptual scheme. I have also advocated a scientific element in this philosophical methodology that I have called metrological naturalism.

One of the big things that I think needs to be explored is the extent to which there are important constraints on conceptual engineering. I can imagine a debate over the legitimacy of a certain kind of conceptual engineering project. Consider the following argument.

> The pro-choice position (abortion is morally permissible in normal circumstances) and the anti-infanticide position (killing an infant is not morally permissible in normal circumstances) are the right ones to have with respect to abortion and infanticide. Furthermore, when I reflect on the nature of time, I find myself committed to the idea that I do not have temporal parts—I am wholly present (metaphysically) at every moment. If so, then I am an endurantist, and presumably I am an endurantist about a zygote/infant, which is the thing that is wholly present throughout the change from having no rights (as a zygote) to having rights (as an infant). So far, so good, but now I start thinking about metaethics and theories of justice, and I arrive at the plausible view that rights are properties had by objects and that our judicial locutions denote these properties; so I'm a realist about rights. Now I think about it a bit more and I find it difficult to believe that having rights is not an intrinsic property of an entity if one is a realist about rights. So now I have a problem making sense of how a single thing could be intrinsically killable at one time (as a zygote) and then not intrinsically killable later on (as an infant). Now I need to think hard about which concepts of time, persistence, and rights I should use, given my commitments to prochoice/anti-infanticide positions.

Is this a good reason to replace my concept of time, concept of rights, concept of persistence or any of the related concepts appealed to in the inset reasoning above? To be clear, I am not attributing this project to anyone in particular. I am just trying to think through the kinds of conditions that one might want on an acceptable conceptual engineering project.

One way to think about it is that I surely have a reason to change my concepts in this way, but it is *the wrong kind of reason*. Reasons for belief of the right kind are those relevant to the truth of the belief, whereas reasons for having a belief, which are irrelevant to the truth of the belief, are reasons of the wrong kind. There is a larger philosophical literature on reasons of the wrong kind. Pascal's wager is a good example; Pascal's reasons for belief in God are entirely focused on the consequences of having that belief, rather than on whether the belief is true. These same considerations seem to suggest that to the reason for changing our concept of time or our concept of rights is *not* a reason to think the new concept is a good one. Instead, the reason given is maintaining certain moral and social commitments (e.g., pro-choice and anti-infanticide). This is a reason to change the concepts I use, but it is not a reason to think that the new ones are valuable or right for the job. As such I am somewhat uneasy about the conceptual engineering project that might be inspired by the inset argument.

On the other hand, conceptual engineering is essentially oriented to action, not belief. Making changes to our conceptual scheme or our language is an activity. And it isn't clear whether there are reasons of the wrong kind for *actions*—notice all the examples given involve beliefs instead of actions. For example, there is no difference in kind between volunteering at a homeless shelter because I want to help the homeless and volunteering at a homeless shelter due to receiving some incentive. Hence, it might be that there are no reasons of the wrong kind to promote moral rightness. So promoting moral rightness is on par with any other reason when it comes to adopting a certain concept of time or a certain concept of persistence, or a certain concept of rights. This result would have tremendous consequences—given the complexity and abundance of connections between concepts, this policy for conceptual engineering would effectively moralize and politicize our entire conceptual scheme. Even logic or mathematics could conceivably be affected.

My attitude on this issue hasn't been entirely stable, but I tend to err on the side of letting a thousand flowers bloom rather than figuring out from the armchair which conceptual engineering projects are kosher. Still, I can envision balking at certain proposals if they were to, for example, promote conceptual confusion or inconsistency for political gain. I think conceptual engineering should always make our conceptual scheme better for us and the concepts we use better for what we use them for. I like to think of conceptual engineering as a wide category, but it certainly has limits.

7. Non-Scientific Exports

I'll consider three objections, one here and the others in the next two sections.

It is clearly false to think that as the subject matter of philosophy shrinks, this subject matter is exported only in the form of sciences. So the entire part of the radical therapeutic program dedicated to helping us escape our predicament of a conceptual scheme dominated by inconsistency (the fly bottle) is baseless.

It is surely right that as the subject matter of philosophy has constricted, science has not absorbed all of it. For example, astrology was a huge part of western philosophy and a primary driver of innovation in astronomy from antiquity until last couple of centuries. Which sciences ended up with this subject matter? None. Throughout the history of western philosophy, there have been changes that eliminate some subject matter from philosophy as not fit for philosophical thinking. Astrology is one example, and sophistry is another expulsion, but it occurred very early in the tradition by those under the influence of Socrates and Plato. In our own modern period, it has become unacceptable within western philosophy to make appeals to God or God's works to explain philosophical puzzles, except at the current time in philosophy of religion and parts of metaphysics. That is a huge change that has happened over the past few centuries. Think of how important appeals to God are in the solution to scepticism according to the orthodox reading of Descartes' *Meditations*.[14] Now imagine trying to publish a paper in a top journal today arguing

[14] Descartes (1641).

that God's works are the best solution to a problem like the liar paradox. Unthinkable! So philosophy often dumps parts of its own subject matter and these dumps need not be exports to the sciences. They need not even be illegitimate—international relations comes to mind as a topic that started as part of philosophy in the early 1800s, but is now its own respectable discipline alongside philosophy in the university, despite the fact that international relations isn't a science.

Nevertheless, my point is unaffected by this complexity. Philosophy changes in all sorts of ways, and one of the most significant and impactful changes it has undergone is the colossal outsourcing of its material to the sciences. Here is a nice quote from Alexis Burgess and Brett Sherman about what has happened just in philosophy of language and just in the last few decades.

> It's not easy being a philosopher of language these days. Work is hard to come by, and we don't just mean jobs. The subject matter itself seems to be getting smaller and smaller. What were once proprietary issues in the field (like the semantics of names, descriptions, quantifiers, etc.) are now quite rightly seen as scientifically tractable research programs in linguistics and psychology.... [T]he marked progress of linguistic semantics obviously owes volumes to the foundamtional work of philosophical luminaries like Frege, Tarski, Davidson, Montague, and Lewis, who helped erect a basic framework for articulating and evaluating claims about verbal meaning. As these foundations have solidified, however, questions once assumed amenable to *apriori* reflection have been exposed as properly empirical quarries. Handmaiden to the science of meaning might be a perfectly respectable job title. But some of us who self-identify as philosophers of language will naturally want to seek out new work.[15]

The process by which philosophy is giving way to the sciences on dozens of fronts is absolutely massive, and it is one of the most significant things that has ever happened in western civilization.

The radical therapeutic program has two parts—it identifies our problem, which is that most or all of our core concepts are inconsistent. And it offers a solution: use conceptual engineering to change our conceptual scheme so that we have concepts that work for us rather than concepts that tangle our thinking, confuse our beliefs, and interfere with our plans. To accomplish this, conceptual engineering aims for certain things in the new concepts, and that is where it relies on metrological naturalism. The solution—the way out of the fly bottle—is to promote the already massive exodus from philosophy that is science. The establishment of science doesn't have to be the only exodus from philosophy for it to be the most significant and for it to be our role model.

8. Defective Concepts in Conceptual Engineering

Another objection: How can I be sure that we won't find awful defects in the concepts employed by conceptual engineering projects themselves?[16]

My reply is that I can't be sure that these concepts aren't defective as well. In fact, I think they probably are defective. High on the list of probably defective concepts is

[15] Burgess and Sherman (2014: 1).

[16] Neil Tennant offered this question at the lecture.

the concept of a concept itself. It has well-known problems and many theorists engaged in conceptual engineering projects even go so far as to be concept eliminativists—they think that talk of concepts has no place in a proper conceptual engineering project. Instead, these theorists contend, we can make do with less controversial tools like extensions, which are just sets of individuals denoted by a predicate, and intensions, which are assignments of extensions to various possible situations. Herman Cappelen is a major proponent of the "no concepts" version of conceptual engineering, and there are others as well.[17]

However, there are two things to say to those in the "no concept" wing of the conceptual engineering movement. First, concept eliminativism is itself a conceptual engineering project, and I have yet to see anyone carry that project out in a careful and detailed way. So far, we have some proposals for how to do conceptual engineering without appealing to concepts, but we have very little in the way of reasons to think that this is a good idea. Moreover, there are bound to be more defective concepts utilized by the conceptual engineer than just this one, so if concept eliminativism is appropriate, then presumably other kinds of eliminativism with respect to the tools of conceptual engineering are appropriate as well, and it isn't clear that there will be enough left for the conceptual engineer to use for her projects.

However, the fact is that concept eliminativism is unjustified even if the concept concept turns out to be defective and in need of replacement. The reason is that defective concepts can still be useful, and even those who know they are defective can still employ them without thereby being irrational. For example, I think the concept of truth is seriously inconsistent, but I am not a truth eliminativist. The analogy I like to use is the concept of mass in classical mechanics. Mass is inconsistent concept but is still extraordinarily useful use it for all kinds of things from building houses to landing robots on comets. Think about how insane it would be to use general relativity to, say, design a sturdy bridge. It would be extremely unwieldy and it would take one far longer, and one would end up with the same bridge that would have been designed using Newtonian mechanics. Therefore, although it is likely that the concepts involved in my own methodology are themselves defective, that does not mean they are not useful for this purpose. When one aims to replace some concept, one tries to figure out whether the defect in that concept actually inhibits its utility—whether its defect actually gets in the way of certain applications. If the defect in a concept does undermine its utility for some purpose, then that is a decisive consideration in support of replacing that concept for that purpose. That is exactly the case with our concept of truth; its defects prevent it from effectively performing the role we ask of it in certain applications of natural language semantics. That is, when one tries to provide a truth-conditional semantics for a fragment of natural language that contains liar sentences, then one ends up with an inconsistent semantic theory. If I were to be shown conclusively that one of the concepts involved in my own methodology had a defect *that was impeding its utility in my methodology*, then that would be a problem for me, and I would focus attention on how to effectively replace that concept for my purposes. Therefore, there is a considerable gap between

[17] Cappelen (2018).

the suggestion that some of the concepts I rely on are defective, and a substantive objection to conceptual engineering as I understand it and practice it. I admit the former, but the latter I have yet to see.

9. Is Philosophy about Concepts?

Here is another objection to what I have said so far: Philosophy isn't the study of concepts at all, so it cannot be the study of what have turned out to be inconsistent concepts.[18] Philosophers do on occasion study concepts, but only as one item among many in other things in the world. For example, there is a difference between the concept of truth and truth itself. Truth is, presumably, a property that things like sentences or theories or propositions can have, whereas the concept of truth is something like a mental representation or a constituent of thought or some other kind of thing that people grasp or possess or understand. Philosophy isn't the study of *the concept of* truth or *the concept of* knowledge or any of the other concepts. Instead, philosophy is the study of certain *phenomena*, like truth, knowledge, freedom, justice, and the rest.

That is all well and good as a start, but as philosophers we must to do better—we need to think a bit deeper about the issue. If our philosophical concepts are as defective as I have suggested, then there is no reason to expect there to be a property of truth or a property of knowledge or any of the rest. At least, not if one thinks of the property of truth as anything like what our concept of truth leads us to think it would be like, and if the property of knowledge is anything like what our concept of knowledge leads us to think it would be like. If the principles for these concepts are inconsistent, then no property can satisfy them. If they are seriously inconsistent, then no property can even come close to satisfying them.

For example, a philosopher might decide to study whether truth is a substantive property that can explain things or a deflationary property that doesn't explain anything. This is a huge area of contemporary philosophy covering the last half-century and involving hundreds of theorists and thousands of publications. One might think that such an inquiry has absolutely nothing to do with our concepts—it's about truth, not the concept of truth. But what, exactly, is the property of truth taken to be? Which property is it? It sure isn't the property had by a sentence 'grass is green' just in case it turns out that grass is green, and the property had by the sentence 'snow is white' just in case it turns out that snow is white, and in general the property had by the sentence <p> just in case it turns out that p. Why isn't it this property? *Because there is no such property.* To suppose there is such a property is inconsistent, as shown in the reasoning of the liar paradox. And there are dozens of other paradoxes associated with truth as well. In fact, truth is such a defective concept that no property satisfies even small subsets of the principles we think of as constitutive of truth. So there is no property of being true, not if that property is anything like what the concept of truth leads us to think it would be like. Philosophy cannot be about the

[18] Williamson (2007).

property of being true because there is no property of being true for philosophers to investigate.[19]

From the point of view of the radical therapeutic program, there might not be anything like what our philosophical concepts lead us to expect in the world. There might not be properties in the world that correspond to our philosophical concepts. In fact, there probably aren't. Perhaps there is *some* philosophical concept that is consistent enough for there to be something in the world that comes close to satisfying its constitutive principles, but that isn't the case for the vast majority of philosophical concepts. Hence, it makes the most sense to think of philosophy as the study of certain concepts—there isn't much else for it to be about. Even philosophers who think of themselves as studying genuine phenomena in the world are usually just exploring one aspect or another of an inconsistent concept. For example, internalists about knowledge and externalists about knowledge aren't investigating some property—the property of knowing something. Instead, each side takes for granted some of the constitutive principles for the concept of knowledge and uses them to argue against those on the other side in the debate, who take for granted *other* constitutive principles for the concept of knowledge. The debate seems interminable and deadlocked because each side is right—each side has latched onto some aspect of our concept, but each side is wrong as well, in that they reject some other aspect of our concept. The fact that internalists and externalists about knowledge—or pretty much any of the sides in any philosophical debate—can refute each other only shows that all the constitutive principles for knowledge, taken together, are inconsistent. That is, it only shows how defective the concept of knowledge is. Another way of putting the point, in terms of subjects or properties instead of in terms of concepts, would be that the subjects or properties that philosophers might think of themselves as investigating are delineated according to inconsistent principles. So there are no such things. The very idea that there is something like truth or knowledge or freedom or justice or virtue for us to investigate at all is inconsistent. Of course, we have the concept of truth and the concept of knowledge and all the rest, and philosophy is primarily the study of these concepts.

So if there is no such thing as truth or knowledge or freedom or virtue, then what is there? We don't know. And we won't know until we have done far more conceptual engineering.

10. Conclusion

Philosophy, or western philosophy at least, has been focused throughout its history on certain topics or certain concepts—truth, knowledge, value, virtue, freedom, justice, etc. The radical therapeutic program presented here is based on the idea most or all of these concepts are inconsistent. Or, alternatively, most or all of these subject matters are delineated in an inconsistent way. Our philosophical concepts, which are the heart and soul of our conceptual scheme, are organized and distinguished by principles that are themselves inconsistent with one another.

[19] See Scharp (forthcoming) for more details on this example.

The result is that just about any time we think or talk about philosophical topics and we try to follow these principles, we end up contradicting ourselves. That is our predicament. The solution sketched here relies on conceptual engineering—charting out the defects in and among our concepts and proposing new concepts that will do the work we demand without causing the problems we currently encounter.

Imagine, for a moment, what it would be like to have a consistent conceptual scheme. No paradoxes. No puzzles. Just clarity. We can do it. You can help.

Acknowledgements

Inaugural address delivered 25 February 2015 at The Ohio State University in Columbus, Ohio. The text has been changed only slightly, but has a few additions. Thanks to everyone present and to those in audiences at the University of Oslo and the University of St Andrews as well.

References

Alchourrón, Carlos, Peter Gärdenfors, and David Makinson. 1985. On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic* 50 (2):510–30.

Blackburn, Simon. 1999. *Think*. Oxford: Oxford University Press.

Burgess, Alesis, and Sherman, Brett. (eds.) 2014. A Plea for Metasemantics. *Metasemantics*. Oxford: Oxford University Press.

Burgess, Alexis, and Plunkett, David. 2013. Conceptual Ethics I. *Philosophy Compass* 8:1091–101.

Cappelen, Herman. 2018. *Fixing Language: Conceptual Engineering and the Limits of Revision*. Oxford: Oxford University Press.

Davidson, Donald. 1990. The Structure and Content of Truth. *The Journal of Philosophy* 87:279–328.

Descartes, Rene. 1641. *Meditations on First Philosophy*. Trans. John Cottingham. Cambridge: Cambridge University Press, 1996.

Galilei, Galileo. 1623. *The Assayer*. Rome: Giacomo Mascardi.

Haslanger, Sally. 2000. Gender and Race: What are They? What do we want them to be? *Nous* 34:31–55.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

Nietzsche, Friedrich. 1886. *Beyond Good and Evil*. Trans. R. J. Hollingdale. Harmondsworth: Penguin Books, 1973.

Popkin, Richard. 1966. *The Philosophy of the Sixteenth and Seventeenth Centuries*. New York: Free Press.

Plato. 1961. Apology. In *Plato: Collected Dialogues*. Trans. F. Cornford. Princeton: Princeton University Press.

Scharp, Kevin. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Scharp, Kevin. forthcoming. Conceptual Engineering for Truth: Aletheic Properties and New Aletheic Concepts. *Synthese*.

Sellars, Wilfrid. 1956. Empiricism and the Philosophy of Mind. In Herbert Feigl and Michael Scriven (eds.), *Minnesota Studies in the Philosophy of Science*, Volume I: *The Foundations of*

Science and the Concepts of Psychology and Psychoanalysis. Minnesota: University of Minnesota Press.

Sider, Theodore. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Weaver, Bryan, and Scharp, Kevin. forthcoming. *Semantics for Reasons*. Oxford: Oxford University Press.

Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Oxford: Blackwell.

20

Linguistic Intervention and Transformative Communicative Disruptions

Rachel Katharine Sterken

Introduction

Sometimes having certain word-meaning pairs in circulation in a population of speakers at a particular time, in a particular social-historical milieu, can be bad. Such word-meaning pairs might cause injustice or disadvantage, stifle discourse, deliberation and inquiry, or stall social progress. It's not hard to think of examples—take any slur. The population would be better off without such word-meaning pairs.

Likewise, sometimes *not* having certain word-meaning pairs in circulation in a population of speakers at a particular time, in a particular social-historical milieu, can be bad. *Not* having these word-meaning pairs can cause injustice or disadvantage, stifle discourse, deliberation and inquiry, or stall social progress. Two prominent examples are discussed by Miranda Fricker (2007) in relation to the notion of hermeneutical injustice: 'sexual assault' and 'postpartum depression'. The population is better off with such word-meaning pairs.[1]

Still further, sometimes certain word-meaning pairs in circulation in a population of speakers at a particular time, in a particular social-historical milieu, could be better. Consider, for example, the recent revision of the meaning of 'marriage' to include same-sex couples: Here the word was kept but the meaning improved. In this way, sometimes changes in meaning for existing words can bring about various favorable effects or prevent various unfavorable ones.

I take it, then, that it matters what word-meaning pairs are in circulation for a given linguistic population; that which language we speak can have a significant

[1] Note that Fricker wouldn't quite frame it this way: she would speak of people being better off with such *hermeneutical resources*, but the point is essentially the same. On the topic of terminology: throughout I speak of word-meaning pairs where others might use talk of concepts or intensions and extensions. There isn't any particularly deep reason for my doing so, but it helps me formulate some of my claims more neatly.

Rachel Katharine Sterken, *Linguistic Intervention and Transformative Communicative Disruptions* In: *Conceptual Engineering and Conceptual Ethics*. Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020).
© Rachel Katharine Sterken.
DOI: 10.1093/oso/9780198801856.003.0020

impact on whether or not the world is as it should be[2] (e.g., if our language lacks the word-meaning pair of our 'postpartum depression', then the world, at least as far as it concerns the treatment of new mothers, isn't as it should be because their sufferings might go unrecognized); and hence, that normative claims about word-meaning pairs are important to reflect on.

It is clear, then, that speakers sometimes have good reasons to and should have a strong interest in *eliminating* existing word-meaning pairs from circulation, *introducing* new word-meaning pairs into circulation, or, indeed, introducing and eliminating word-meaning pairs in tandem—that is, what has been variously called *changing*, *shifting*, *engineering*, *replacing*, *revising*, *improving*, *innovating*, or *ameliorating*.[3] Whether or not they should always act on those reasons or interests, and what they should do to act on those reasons or interests, are different and more difficult questions to answer.

In this chapter, I attempt to characterize one potentially controversial, though sometimes justified, means by which to act on those reasons and interests. I provide a descriptive account of a kind of linguistic strategy speakers of a language can, and sometimes do, engage in to bring about changes to the word-meaning pairs in circulation (sections 1, 2, 5, and 6). I outline what it is I take to be controversial about this strategy—that is, the kind of moral and linguistic challenges there are to justifying the strategy (sections 3 and 4). Finally, I argue that there is a way around these challenges by way of a linguistic phenomenon I call *transformative communicative disruption* (section 5).

1. Meaning Change, Linguistic Intervention, and Linguistic Transgression

Linguistic Interventions

Before we can understand whether or not speakers should act on their reasons or interests in changing the word-meaning pairs in circulation, it would first be useful to understand from a semantic, pragmatic, and metasemantic perspective what it is speakers are doing when they intentionally and strategically try to eliminate, introduce, or change the word-meaning pairs in circulation, and what kinds of effects these doings can have. Call communicative activities on the part of a speaker that (intentionally and strategically) attempt to change the word-meaning pairs in circulation, *linguistic interventions*.

I take it that having a proper semantic, pragmatic, and metasemantic account of linguistic interventions is of independent theoretical interest, because regardless of whether speakers should act on their interests in changing word-meaning circulation,

[2] I don't mean to commit myself, in this chapter, to a particular normative theory or meta-ethical stance. I take it that the contents of this chapter can be suitably rephrased without loss of the central observations and claims therein, if one does want to commit to some theory or stance.

[3] These notions are inspired by the work of various authors: Haslanger (2012); Burgess and Plunkett (2013a,b); Plunkett and Sundell (2013, forthcoming); Scharp (2013); Richard (2014); Eklund (2015, 2017); Cappelen (2018). It should be clear, however, that my understanding is narrower and should not be confused with the different views of these authors.

it is clear that speakers often *do* engage in communicative activities that aim to introduce and/or eliminate word-meaning pairs from circulation. Several recent prominent examples include: 'marriage', 'rape', 'sexual assault', 'organic', 'terrorist', 'migrant', 'fake news', as well as pronoun introduction and use. Such cases may plausibly be construed as largely driven by grass-roots, bottom-up linguistic intervention, but there are numerous examples of top-down attempts at institutional intervention as well: where institutions attempt to change the word-meaning pairs in circulation via legislation, authority, or influence (think, to use an example George Lakoff made famous, of the Republican Party's mostly successful attempt to replace 'tax cut' with 'tax relief'). I will be principally concerned with processes that are primarily bottom-up, though of course both are important and interesting forms of linguistic intervention.

I take it that the processes of meaning change in cases of linguistic intervention are different from standard processes of meaning change whereby input to that process is primarily constituted by normal usage, permissible pragmatic operations on existing meanings, changes in the world, or speakers' conceptions of the world. Processes of meaning change that are driven by linguistic intervention, by contrast, have as a crucial part of their input strategic, intentional, or project-like plans on the part of speakers to change which word-meaning pairs are in circulation. Such speakers have a *metalinguistic standpoint*—a set of beliefs about what word-meaning pairs should be in circulation amongst their linguistic community, and their linguistic activities are in part guided and influenced by that metalinguistic standpoint. They imagine that if our language were like this, then our language would be better or the world would be better off. Their intentions and metalinguistic standpoint can affect their linguistic activity, and hence, the semantic and pragmatic properties of their speech.

Linguistic interventions are similar to, but also importantly different from, other communicative exchanges discussed in the literature, most notably metalinguistic negotiations. Consider the following example of a metalinguistic negotiation from David Plunkett and Tim Sundell (2013: 14—15). We're making chili for dinner. You see me chopping up several more scotch bonnets to add to the already pepper-filled chili. You say that you don't like things too spicy, and I ask what counts as spicy. You taste the chili, turn red, and say '*That* is spicy'. I taste it and say: 'You wuss, that's not spicy'.

What's going on here, arguably, is a disagreement about how to use the word 'spicy': that is, a metalinguistic negotiation, one about words. In answering my question, you are conveying that that chili *should* fall under the extension of 'spicy'. In responding, I'm denying that it should. What we're doing is negotiating about how precisely to use the word 'spicy' in this and future culinary endeavors. In what follows let's use metalinguistic negotiation to refer to metalinguistic disagreements which have this normative component (by contrast, we'll say a *metalinguistic use* need not have this normative component. If we're in the zoo and I see a sign saying 'Pachyderms to the left', I might ask you what pachyderms are and you might say, pointing to a big elephant, '*that* is a pachyderm'. You thereby say something, at least in part, about the word 'pachyderm', namely that that elephant falls under its extension, and so your use is metalinguistic, but it's not a normative claim about how we should use 'pachyderm' for our conversational purposes. (See Plunkett and Sundell 2013:14 and Ludlow 2014a: 13—14 for more on metalinguistic use.)

As I understand Plunkett and Sundell (2013) and Ludlow (2014a,b) metalinguistic negotiations are limited in their scope—the aim of the negotiation is to settle what a given word should mean *in the context of a given communicative exchange*. Speakers in metalinguistic negotiations needn't have diachronic intentions to change the meaning for the linguistic community as a whole, in (all foreseeable) future contexts. Linguistic interventions, by contrast, have this much more ambitious goal. The linguistic properties of metalinguistic negotiations and linguistic interventions differ, then, in that negotiations are attempts for the target word *w* to mean *A* in context *c* (or some suitably limited set of contexts *C*), whereas interventions are attempts for the target word *w* to mean *A sans phrase*. The semantic, pragmatic, and metasemantic properties of the respective utterances differ and the two differ in how they are the input to processes of linguistic change. Metalinguistic negotiations are input in the process of meaning change by altering use-facts, resolving underdetermination or changing speakers' conceptions. Linguistic inventions, by contrast, are attempts to introduce a new meaning or eliminate an old one—to anchor a new word-meaning pair or derail an old one.

Amelioration and Facilitating Meaning Change by Use

As a hypothetical example of linguistic intervention, consider Sally Haslanger's ameliorative analysis of the concept 'woman':[4]

S is a 'woman' iff

i. S is regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction;
ii. that S has these features marks S within the dominant ideology of S's society as someone who ought to occupy certain kinds of social position that are in fact subordinate (and so motivates and justifies S's occupying such a position!); and
iii. the fact that S satisfies i and ii plays a role in S's systematic subordination, that is, along some dimension, S's social position is oppressive, and S's satisfying i and ii plays a role in that dimension of subordination. (2000: 42)

Call the meaning of 'woman' that corresponds to this ameliorative analysis *A*. For my purposes, nothing hangs on the specifics of the analysis—what does matter is that the analysis offers a significantly different meaning than that currently in circulation. One way to put this is to say that the conventional meaning of the term 'woman' is significantly different, perhaps incommensurable with, the proposed ameliorated meaning for 'woman'.

Now, imagine a speaker with the metalinguistic standpoint that the word 'woman' should mean *A*, and who engages in linguistic activity, at least some of the time, whereby she uses the word 'woman' in a way so as to facilitate meaning change. The ameliorator might do various things to facilitate meaning change. She might, for

[4] I add a proviso here that Haslanger would not necessarily endorse any claims that I make by way of using this example.

example: (i) assert or propose that 'woman' should mean *A* in hopes of changing people's conceptions or use; (ii) explicitly mark her speech or make her intentions manifest—that she means *A* by 'woman'; or (iii) metalinguistically negotiate, and thereby potentially change some metasemantic determinants of the meaning of woman (e.g., the use facts); or (iv) she might make a word-meaning pair taboo by attempting to invoke worldly consequences for its use (see Anderson and Lepore 2013). I won't focus on these sorts of linguistic activities.

Rather, I want to focus on linguistic activities on the part of the ameliorator whereby she uses 'woman' as if the word means *A*, and/or interprets others' uses of 'woman' as if the word means *A*.[5] In particular, I will focus on cases whereby the ameliorator treats 'woman' as meaning A even though she doesn't believe that 'women' means *A*. I want to focus especially on cases where the ameliorator's interlocutor is not (fully) aware that she is speaking/interpreting in this way. This might be so because the ameliorator's intentions were not manifest to her audience, or because her interlocutor isn't sufficiently aware of her project and metalinguistic standpoint. Linguistic activities of this sort I understand as an interesting and important form of linguistic intervention.

One might judge such linguistic activity with suspicion. Why use 'woman' as if it means something it doesn't, when you're aware your audience has little chance of understanding what you are saying? Why purposely misunderstand what someone is saying to you? Despite the apparent unreasonableness of this sort of linguistic engagement, I think there is a way to make sense of it, and I think it is more common than one might initially have thought. In the case of introduction and change, it is an attempt to anchor a new word-meaning pair—to homophonically baptize—by way of metalinguistic use—while, simultaneously, attempting to render defective the interpretive common ground of the original word-meaning pair (cf. Barker 2002; Krifka 2013; Richard MS). In the case of elimination, it is an attempt to break the communicative chain (Kripke 1980) or dominance facts (Evans 1973) that connect speakers to the problematic anchoring event, again while simultaneously rendering defective the interpretive common ground.

Linguistic Disruption

One important and distinctive thing about this kind of linguistic activity is that it is *disruptive*. Linguistic interventions are disruptive in at least two senses. First, linguistic interventions of this sort are a disruptive form of communication. The ameliorator's linguistic activity attempts to disrupt the interpretive common ground so as to affect metalinguistic reflection and reconstruction on the part of her interlocutor—it attempts to disrupt the interpretive resources of the interlocutor so that she engages in imaginative and counterfactual thought about language and its potential role in the world. Second, linguistic interventions disrupt the normal functioning of the language system. Rachel Ann McKinney (2015), amongst many other theorists thinks of our language system as serving certain functions and that a

[5] See Thomasson 2016 for a view on which much of metaphysicians' talk should be understood in this way: the nihilist about composition, for example, uses *table* as if it has no meaning (or at least an empty extension).

well-functioning language system has value insofar as it serves those functions—as McKinney (2015) nicely puts it: a well-functioning language system allows us "to learn from each other, to inquire and deliberate together, to pool information, to coordinate action, express care and concern, to reflect on and solve common problems, and so on" (p. 54). What such linguistic interventions do or aim to do is to disrupt the functions of the language system, so as to effect change.

In what follows, I will make a case for the claim that this kind of linguistic activity is in fact as normal and perhaps even as pervasive as other forms of linguistic activity that are disruptive—for example, lying. I will also argue that it is reasonable, from a theoretical perspective, to construe such activity as part of the normal maintenance of our language system, and perhaps even the (social) world more broadly. Such an observation is significant, as it ultimately makes acts of linguistic intervention easier to justify.

On the other hand, despite any social, practical, and intellectual benefits such changes might potentially achieve, such changes are by no means always easy to bring about, if they can be brought about at all. (In sections 3 and 4, I outline complications which ultimately make acts of linguistic intervention harder to justify.) Some authors have already noted that meaning shifts seem hard to achieve (Burgess and Plunkett 2013; Cappelen 2018). For example, if one endorses semantic externalism, then speakers can't always simply change what their words mean, neither by individual nor by collective will. Many of the metasemantic facts that fix the meanings of our expressions are simply outside of our control. To give an incomplete but illustrative list, none of the following can be said to be in full control of any given speaker or even community of speakers: naturalness and magnetism (Lewis 1983, 1984; a recent exhaustive discussion is Dorr and Hawthorne 2013); patterns of past usage or future usage (Williamson 1994; Jackman 1999); linguistic conventions (Lewis 1969); features of the event where the meaning was introduced (Putnam 1975; Burge 1979); causal chains or dominant sources (Evans 1973; Kripke 1980); speaker intentions (Kaplan 1989; Stokke 2010; King 2014), some even argue modal facts (see Cappelen and Dever 2018: 92ff) and normative facts (Haslanger 2012; see Cappelen 2018: 79 for discussion) about usage. Endorsing semantic internalism won't do us any better either (Cappelen 2018: 91; *pace* Burgess and Plunkett 2013: 1096). This makes acts of linguistic intervention difficult to justify because it simply seems like there is no way for us to fruitfully control the relevant metasemantic facts.

Linguistic Transgression and the Linguistic Reformer's Dilemma

I won't focus on the above-mentioned set of difficulties; instead I will focus on another worry. The worry centers on the fact that some acts of linguistic intervention involve what we might call *linguistic transgression*. The issue can be usefully illustrated by considering what is known as the Reformer's Dilemma. Suppose for the sake of illustration that semantic conventionalism is true—that is, that what our words mean is tightly constrained by the linguistic conventions of the relevant community of speakers. The *Linguistic Reformer's Dilemma* is as follows: Suppose speaker *S* is a linguistic reformer and thinks that word *w* should mean *A* where *A* is not the meaning determined by the linguistic conventions of *S*'s community. If *S* uses *w* to mean *A*—that is, speaks to and interprets others as if *w* means *A*—then *S* has done something wrong *qua* member of her linguistic community (supposing ordinary, non-figurative use). Either *S* can speak and interpret others correctly (according to conventional meaning) or *S* can reform the language, but *S* can't do both. Thus, an act of attempted reform of this sort will involve a fault on the part of speaker—a linguistic transgression.

I think there is a way out for the linguistic reformer—the reformer can overcome the challenge linguistic transgression poses for her realizing her aim of linguistic reform. Later, I argue that an important strategy for linguistic reformers is to engage in linguistic transgressions because these elicit *transformative communicative disruptions*. In such cases, the interventionist's transgression is justified (it is outweighed by the potential benefits to be achieved—either representational or worldly), and she is engaging in activity whereby her interlocutor can reflect on the meaning of the given word, acquire the new meaning, and recognize the new meaning as an improvement. Sometimes being a good member of a linguistic community will involve disrupting and transgressing: The reasons for having the linguistic system in the first place can give us reason to flout individual conventions and norms of that system.[6]

2. On the Pervasiveness of Linguistic Interventions

Generally, word-meaning pairs only come into circulation when two things have happened (at least on some prominent theories):

(I1) something like an *anchoring* or *baptism* event (Kripke 1980) has occurred whereby a speaker performs a dubbing and the word becomes connected with a referent, and;

(I2) some process of spread has occurred whereby the word-meaning pair is brought into circulation amongst the community of speakers.

Kripke and subsequent literature discuss introductions by means of explicit metalinguistic discourse involving *deixis* and description (e.g., his 'Feynman' and his 'Jack the Ripper' cases). But there are other options: One might argue that speakers can anchor a new word-meaning pair by way of metalinguistic use, and in particular, a metalinguistic use with an ameliorator's scope and ambitions.[7] In such a case, the word-meaning pair is introduced by way of linguistic intervention of the sort outlined above.

If a word-meaning pair is to be genuinely replaced by a new word-meaning pair, more than (I1) and (I2) needs to have taken place—(I3) needs to have taken place as well:

(I3) a process of elimination has occurred whereby the original word-meaning pair is eliminated from circulation amongst the community of speakers.

[6] Let me ward off some obvious objections: It is worth mentioning that I am supposing as preconditions to engagement in this kind of linguistic activity that the reformer has good reason to believe that: (i) her project has a fair chance at success; and (ii) her speech does not pose a detrimental threat to the functioning of the language system.

[7] For some work on this see the discussion in Armstrong (2016) on lexical innovation.

A word-meaning pair might disappear from circulation by simply becoming obsolete, but it can also disappear with the help of linguistic intervention. An example of such an intervention is discussed below (see section 2). In these cases, the interventionist is attempting to eliminate the word-meaning pair from circulation by breaking the communicative chain that connects the linguistic community to the anchoring event of the original word-meaning pair.

To see that linguistic interventions are as pervasive and normal as other forms of linguistic activity that disrupt the normal functioning of the language system, like lying, consider the following collection of phenomena, which are now part of or could easily be considered part of descriptive, empirical projects in linguistics and philosophy of language. One example we already considered above—that of ambitious metalinguistic negotiations without strict limitations of contextual scope. I consider five more examples in turn: (i) neologisms, protologisms, and semantic introductions; (ii) the reappropriation of slurs and insults; (iii) transgressive uses of definitional or normative generics; (iv) semantic elimination and interpretive uncharity; and (v) blocking and flouting (semantic) presuppositions.

(i) Neologisms, Protologisms, and Semantic Introductions

Word-meaning pairs where (I1) has taken place, but where the process in (I2) is incomplete are called *neologisms* or *protologisms*. An example from the feminist movement in the 1970s is 'womyn's herstory'. Examples abound in the age of social media: 'hangry', 'tweet cred', and '#X', for any *X*. As an example of semantic introduction consider the slang word, 'cool'. 'Cool' had a meaning before its slang use, and arguably its slang use is related in some way to its original meaning. In this case, a speaker introduced a new meaning homophonically by introducing an ambiguity or polysemy, and it spread and acquired a meaning. Other examples include: the introduction of 'administrative assistant' to replace 'secretary', the introduction of 'firefighter' to replace 'fireman', the introduction of 'server' to replace 'waitress'.

(ii) Reappropriation of Slurs or Insults

Consider the reappropriated slur, 'bitch'. This can be seen a further example of linguistic intervention. In its original use the term had an oppressive, insulting meaning, but once it was reclaimed the term was used in an acceptable, non-oppressive, and often even a positive manner. Initially, the term was used within a local community of users that included the reappropriators, but later it was also understood as such by the larger linguistic community.

The reappropriators of the term 'bitch', during the period of trying to reappropriate the term, proposed and tried to get others to accept a revised meaning for the existing slur. In doing so, we can suppose, reappropriators used the word 'bitch'. In particular, reappropriators might have used the term when it had its conventional meaning, though they used it as if it had its reappropriated meaning.

(iii) Transgressive Uses of Normative or Definitional Generics

Consider the following examples of normative or definitional generics (Cohen 2001; Haslanger 2007; Krifka 2013; Leslie 2015):

(2) a. Girls are tough.
b. Crop tops are cute.

One prominent view of definitional or normative generics treats them as metalinguistic claims involving a metalinguistic use. For example, (2a), makes a statement about the meaning of the term *girls* and how it is or should be used:

(3) The term *girls* is/should be used so that it applies to things that are tough.

But (3) doesn't merely say something about the linguistic expression *girls*, it can also advocate for a particular meaning the speaker endorses, regardless of whether "toughness" is part of the conventional meaning. In addition, in so advocating, the speaker uses the term *girls* to express her endorsement of a definition or descriptive generalization involving the existing linguistic expression.

(iv) Interpretive Uncharity and Semantic Elimination

Consider the following communicative scenario involving an ameliorator of the term 'woman':

> Suppose that Ben is walking with a friend in a park. He says, of his small, female child, 'I hope she grows up to be a strong and powerful woman'. The ameliorator follows through with her commitment to her project, and understands Ben to hope that his small child will grow up to be strong, powerful, and subordinated on the basis of the reproductive organs she is perceived to have. The ameliorator thus responds, 'That's perverse. Why would you want your child to be unjustly subordinated?' Further, if Ben does not take his assertion back, the ameliorator then reports what he said to others, saying that he wants his own child to be subordinated on the basis of the reproductive organs she is perceived to have.[8]

That might seem a bit artificial. But real life examples aren't hard to find. Thus consider:

> In the autism community, there's a debate about the correct terminology to use to discuss autistic people. There's a push for using 'autistic person' as opposed to 'person with autism' because, it's argued, the former phrase better reflects the centrality to the person's identity of autism. Autism, the thought goes, is not something incidental to a person and language should reflect this.[9] Now imagine Ben's friend is someone who agrees with this perspective, and so thinks 'person with autism' is not the right way to refer to autistic people, because to be a person with autism is to be someone for whom autism is a merely incidental feature of their identity, and there are no such people. And imagine Ben has just learned his daughter is autistic, but Ben doesn't know about the linguistic debate. He says 'Children with autism grow up to be adults with autism, so I've already got used to the thought that my daughter will face some challenges as an adult.' Sarah replies, uncharitably, 'What? You think autism is just an accidental feature of people? That's weird. I would have thought you would have known better.'

[8] Thanks to Jack Spencer for this example.

[9] See, for example: http://autisticadvocacy.org/about-asan/identity-first-language/. Thanks to Matthew McKeever for this example.

This kind of communicative deviance is perhaps more involved, and harder to justify, but it can also be effective at eliminating word-meaning pairs from circulation. If the ameliorator engages in this sort of linguistic activity, as a hearer, her interlocutors will tend to stop using the word 'woman' with its non-ameliorated meaning, or stop using 'person with autism' completely.

(v) Blocking or Flouting (Semantic) Presuppositions

When babies started sporting onesies with 'I love my mommies!' printed on them,[10] these uses of 'my mommies' were meant to provoke the idea that 'my mommies' can mean 'my parents'. It's not far off to think that this intervention was intended to induce a permanent shift in the (semantic) presuppositions associated with 'my mommies'. One might imagine that attempts to use 'my mommies' in the ameliorated sense before the intervention resulted in presupposition failure or were met with the reaction that the speaker was confused, whereas after the intervention 'my mommies' could be used unproblematically to refer to two (or more) women as a parental unit. In this example, the ameliorator in producing the onesie was likely motivated, in part, by the belief that 'my mommies' should have an unproblematic lexical meaning in such circumstances. One way of understanding the example, then, is to think of the onesie intervention as a strategic and intentional attempt to change the (semantic) presuppositions of 'my mommies'.

The important point is this: All of the above linguistic activity seem like things speakers are regularly and extensively engaged in, despite its transgressive and disruptive nature. Like lying, linguistic interventions are a normal and regular form of linguistic engagement, with complex and interesting social and moral implications.

3. Challenges for the Ameliorator I—Transition Periods and the Inevitability of Miscommunication

In Evans's *The Causal Theory of Names*, he describes a case of reference shift which is illustrative:

> A youth *A* leaves a small village in the Scottish highlands to seek his fortune having acquired the nickname 'Turnip'. . . . Fifty or so years later a man *B* comes to the village and lives as a hermit over the hill. The three or four villagers surviving from the time of the youth's departure believe falsely that this is the long-departed villager returned. Consequently, they use the name 'Turnip' among themselves and it gets into wider circulation among the younger villagers who have no idea how it originated. I am assuming that the older villagers, if the facts were pointed out, would say 'It isn't Turnip after all' rather than 'It appears after all that Turnip did not come from this village'. In that case I should say that they use the name to refer to *A*, and in fact, denoting him, say false things about him (even by uttering 'Here is Turnip coming to get his coffee again').

[10] Thanks to Joshua Armstrong and Samia Hesni for this example.

> But they may die off, leaving a homogeneous community using the name to refer to the man over the hill. I should say the way is clear to its becoming his name. The story is not much affected if the older villagers pass on some information whose source is A by saying such things as 'Turnip was quite a one for the girls, for the younger villagers' clusters would still be dominantly of the man over the hill. But it is an important feature of my account that the information that the older villagers gave the younger villagers could be so rich, coherent, and important to them that A could be the dominant source of their information, so that they too would acknowledge 'That man over the hill isn't Turnip after all'. (1973: 23)

In Evans's case, there is a period where reference shift has not yet occurred and the villagers end up saying false things about the initial referent. There is also a period where the older villagers die off and the young ones establish a new "dominant source" for the name 'Turnip'. Before the dominant source is established, however, the young villagers say meaningless things since there is no established referent for the term during that period. It should be clear that both periods, during transition from one meaning of a term *w* to a new one, are extremely important for anyone engaged in a project of linguistic reform.

Call a *transition period* the period during a project of linguistic revision before meaning change is successful. During the transition period, there will be many uses of *w* by the interventionist (and her local speech community) where meaning change has not yet occurred. These uses are important to the success of her project, so it is important to understand what the semantic and communicative properties of these uses of *w*. As Evans observes, such uses will sometimes be false or nonsensical, and hence semantically or communicatively *deviant* in some way. This presents a challenge to the linguistic interventionist: attempts to introduce and use *w* with the desired new meaning will result in linguistic transgressions in the form of false and meaningless speech.

Jennifer Saul (2006) makes similar observations in raising concerns for Haslanger's ameliorative analyses of gender, claiming that ameliorative projects have the strange consequence that an ameliorator's speech inevitably leads to misunderstandings and confusion:

> Imagine that Amanda takes a feminist philosophy class and is convinced by Haslanger's views. She decides to use the terms 'woman' and 'man' in the way that Haslanger suggests in order to explain to her friend Beau what she has learned. Amanda utters (1):
>
> (1) All women are subordinated by men.
> Beau does not use 'woman' and 'man' in the way that Amanda uses these terms. He uses them, let's say, as sex terms. A first question is what Amanda has said. Since the speaker and audience have different meanings in mind . . . it is genuinely unclear what the right answer is. Possibly, the right answer is that Amanda has failed to say anything. This seems strange. Perhaps more plausibly, Amanda has said one thing and Beau has understood her as saying another. . . . These difficulties . . . point to the seriousness of the confusion that is possible with a contextualist version of Haslanger's view. In so doing, they offer some reason for resisting it. (2006: 141–2)

This case again shows that the linguistic interventionist, in attempting to introduce and use *w* with an improved meaning, will linguistically transgress by causing misunderstandings and confusion.

The extent of the misunderstanding and confusion that the linguistic interventionist can cause is even worse than it seems from these two examples. The false or meaningless beliefs, misunderstandings, and confusion can be spread throughout the community of speakers by way of (i) misattribution in speech and thought reports and (ii) failed testimonial chains. I consider examples of each in turn.

Consider again Saul's ameliorator Amanda from the above quoted example. Direct assertions won't be the only utterances involving 'woman' that Amanda will make. Imagine Donald Trump uttering 'she is not socially subordinated' (pointing at his daughter Ivanka). Amanda can report what Donald Trump said using 'woman', by saying:

(2) Donald Trump said that Ivanka is not a woman.

Supposing that 'woman' is used as a sex term in Amanda's linguistic community, in uttering (2), Amanda says to her hearer that Donald Trump said that Ivanka is not female. Thus, saying what others say and think while using their words becomes difficult for the linguistic interventionist.

Another type of example is what are called *failed testimonial chains*: Amanda says 'women are F' and another speaker, Eliot, is told that Amanda says that women are F. Eliot trusts Amanda, and passes this along to others, but Eliot is unaware of the ameliorated meaning. Here we have miscommunication spreading false or defective beliefs via failed testimony.

This presents a challenge to the linguistic interventionist as it seems that any attempt at intervention will lead to various kinds of defective communication and linguistic transgression. Hence, there is a precise sense in which the ameliorator's speech disrupts her linguistic engagements and undermines the functioning of the language system.

4. Challenges for the Ameliorator II—From Miscommunication to Lying, Misleading, and Bullshitting

So far this might all look like an argument to the effect that meaning transitions will lead to false or defective beliefs and miscommunication. But does the inevitability of miscommunication entail the inevitably of more serious forms of deviant speech and transgression? On the one hand, interventionists are merely attempting to introduce new word-meaning pairs into circulation. On the other hand, they may know their speech will lead to disruption, miscommunication, and confusion.

Let us distinguish some phenomena:

Uncooperative Speech: Speaker and audience do not have common conversational goals and do not share the relevant mutual attitudes.

Lying: A speaker lies iff:

(i) She says that q;
(ii) She believes q to be false;
(iii) She intends the hearer to believe that q is true. (Mahon 2015)

Misleading: A speaker misleads iff:

(i) She communicates that q;
(ii) She believes q to be false;
(iii) She intends the hearer to believe that q is true. (Mahon 2015)

Bull-shitting: A speaker bull-shits when they say something without caring whether what they say is true or false. (Frankfurt 1986)

These analyses are not the state-of-the-art in the literature on insincere speech. However, they have the benefit of being straightforward and relatively good reference points for connecting to the state-of-the-art should one be inclined towards a particular analysis (e.g., Fallis 2009; Stokke 2013). Let us see then if our examples of interventionist speech are in a position to satisfy these definitions.

It seems obvious that the kind of interventionist speech of interest counts as uncooperative: she clearly has different conversational goals from her interlocutor. Let's take a more difficult case. The least obvious case is whether the interventionist counts as lying. Consider whether the ameliorator Amanda counts as lying given the criteria (i) to (iii) in our definition of lying stated above. To fix examples, consider Amanda who utters to her friend Beau 'We need to get rid of women'. In Amanda's mouth, she is saying that we need to get rid of subordination on the basis of perceived female biological features. However, even if Amanda wants the meaning of 'woman' to be *A*, it often cannot be because her hearers will not recognize her intention or recover her intended meaning. Beau will understand her as saying that we need to get rid of females (or some pragmatically modulated version of this).

In this example, Amanda's speech says that we need to get rid of females. Moreover, she knows that this is what the sentence 'We need to get rid of women' says, and she knows that this is false. Thus, Amanda's speech satisfies (i) and (ii) in the definition of lying above. What about criteria (iii)? Does Amanda intend for her hearers to believe that we need to get rid of females (is true)? This last criterion is tricky: Amanda certainly intends that her audience believe that it's true that we need to get rid of subordination. But she didn't succeed in communicating this content and knew at the time of speaking that she wouldn't succeed. Is Amanda's belief that her audience would pick up the literal content and believe it enough to count as intending her audience to believe it?

Even if my reader isn't convinced of this, I will argue that this intention is actually crucial for the success of the interventionist project. In the next section, I will outline why this is so.

Before that, however, it is worth noting that some of the interventionist's speech more easily qualifies as misleading or bullshitting. In the case of misleading, the ameliorator need only communicate something false and in the case of bullshitting, she need only show indifference towards the truth of what she says or communicates.

5. Effective Linguistic Interventions—Transformative Communicative Disruptions

Ameliorators, given the right motivations, can use deviant communication—like miscommunication, uncooperative or insincere speech—to accomplish their projects of linguistic change. The disruption of standard communicative patterns can help them accomplish their goals. The disruptions are a good thing as they can have the effect of making the hearer stop and reflect on her usage, and this reflection can be *transformative*. In other words, the deviant communicative activity of the ameliorator can engage the hearer in the sort of metalinguistic reflection needed to acquire the new meaning and understand the ameliorator's utterances as she intends. In addition, it can engage the hearer in reflection about the representational and worldly consequences of her speech, and how a change in word-meaning pair may help bring about representational and worldly benefits.

Return to the example involving Amanda's utterance of 'We need to get rid of women'. Consider someone not initiated in or not knowing about Amanda's project who is told that her goal is to get rid of women. On the supposed conventional reading, this is horrific. That reaction of horror, which is a result of the miscommunication and maybe even an intended one, is constructive. It makes the audience stop and think, and that thinking will trigger further communicative efforts on the part of the hearer. It will lead her, one might hope, to start reflecting on the meaning of her words, and that process itself is part of the goal of the project of linguistic intervention.

Here is a further example of how a false or defective belief can trigger reflection of this sort: Imagine a reappropriator, Mia, who says 'Samantha is a bitch' and miscommunicates that Samantha is bossy, etc. Her audience will be shocked, as Mia is normally such a nice person and knows what it's like to be called a bitch. This cognitive dissonance, Mia's kindness on the one hand and her use of the term 'bitch' on the other, can make them reflect on their usage, and their grasp of the meaning of 'bitch' could undergo a transformation.

Similarly, in the cases of the uncharitable interpreter and the onesie-interventionist that flouted the (semantic) presuppositions of 'my mommies'. Each of their communicative transgressions results in the kind of reflection that can be transformative.

So far, these are transformations that occur in individual communicative exchanges, involving individual speakers. Such exchanges might only have minimal effect on achieving meaning transformation. However, individual transformations can eventually spread across the linguistic population and lead to full meaning transformation. Once this has occurred, the project of linguistic intervention is successful.

Disruptiveness is not sufficient for meaning transformation. In meaning transformation, coming to understand the proposed amelioration is transformative. The deviant communicative event, together with a number of other events, triggers a full-on meaning change that is transformative. Meaning changes are transformative when they enable an interpreter to think and communicate things she could not have thought or said without having that meaning—having that meaning gives the interpreter new abilities to imagine, recognize, create cognitive models, and communicate using that meaning. The change offers a new way of understanding the world, and results in substantially revised normative commitments and core preferences.

Think about each of the cases we have considered so far and how the new meanings result in discontinuous, revised understandings of the world and of the kinds of normative commitments and preferences one has. Amanda understands woman as unduly subordinated and now wants to take action. The reappropriator recognizes that she shouldn't ever be called a bitch (in the old sense) and that taking control of the term disempowers those that would. The onesie intervention allowed for a new understanding of who counts as parents, and people's normative commitments and preferences surrounding parenthood change.

So far, I have discussed meaning transformation, and events that trigger meaning transformations. These are not yet transformative experiences in L. A. Paul's sense (2014). But I think there are some close analogies. The first is that authors in the literature on amelioration and conceptual improvement sometimes think of conceptual change analogously to how Paul thinks of transformative experiences. Paul (2014: 17) describes a transformative experience as "an experience that is both epistemically and personally transformative" where an experience is *epistemically transformative* when it "teaches you something you could not have learned without having that kind of experience. Having that experience gives you new abilities to imagine, recognize, and cognitively model possible future experiences of that kind" and where an experience is *personally transformative* when it "changes you in some deep and personally fundamental way, for example, by changing your core personal preferences or by changing the way you understand your desires, your defining intrinsic properties, or your self-perspective". The authors in the literature on amelioration and conceptual improvement also observe something that might be classed as epistemically and personally transformative. Burgess and Plunkett, for example, nicely summarize something that might be considered a transformation of our concepts:

> Arguably, our conceptual repertoire determines not only what beliefs we can have but also what hypotheses we can entertain, what desires we can form, what plans we can make on the basis of such mental states, and accordingly constrains what we can hope to accomplish in the world. Representation enables action, from the most sophisticated scientific research, to the most mundane household task. It influences our options within social/political institutions and even helps determine which institutions are so much as thinkable. Our social roles, in turn, help determine what kinds of people we can be, what sorts of lives we can lead. Conceptual choices and changes may be intrinsically interesting, but the clearest reason to care about them is just that their non-conceptual consequences are pervasive and profound. (2013: 1096–7)

But the meaning transformation discussed here might seem very different after all from transformative experiences. Meaning transformation is a drawn-out event, consisting of innumerable small interactions between speakers and audience members. However, I think this is also true of some of Paul's core cases—like becoming a doctor (Paul 2014: chapter 3): It's not a single event, or if it is one event, then it's one that is stretched out over time.[11] Moreover, it is unclear to what extent meaning

[11] That said, it's arguable that some such events are single and non-stretched out: seeing the mommie onesie could immediately make something about same-sex partnerships click for someone who had previously, for example, not paid attention to the debates and news stories about the matter, causing her to fundamentally rethink the nature of parenthood.

transformations are first-personally transformative for all speakers. These issues I leave as open questions.

6. Justifying Linguistic Interventions—the Long Game and Diachronic Communicative Intentions

Some pressing concerns regarding the communicative strategy discussed have been raised. The disruptions are risky communicative behavior. They won't always work. The ameliorator intentionally risks prolonged misleading or worse. It's not always going to be the case that her audience can look back on the exchange and see that it was the improved meaning that was intended. Interventionists even risk detrimentally undermining the functioning of the language system, or they may end up silencing themselves—in a manner much like what McGowan (2013) calls *sincerity silencing.*

But it is important to note that even though Amanda, the reappropriator Mia, the uncharitable interpreter, and the onesie-interventionist used disruptions to contribute to their projects, they didn't have ill intentions in doing so. In this section, I outline how we might understand the communicative intentions of the ameliorator in a way that makes them unproblematic.

In order to do so, let's summarize the communicative situation I am characterizing: *A* utters *S* to *B*, *A* knows that in their public language *S* means that *q*. *A* doesn't believe that *q* or is indifferent towards *q*, but *A* wants *S* to mean that *p* and wants this speech act to be part of the revisionary process. *A* intends for *B* to, at first, get the false or defective belief that *q*—then, somehow, that belief, in addition to a number of other events, will accomplish the revision. *B* can then think back on this communicative exchange and realize that it was *p* that was the ultimate communicative intention. *B* can also gain the now true, intended belief that *p*. In this way, *A* and *B* were part of a transformative miscommunication.

In characterizing the situation in this way, we see that the ameliorator is interested in the communicative "long game", not just what's going on in her own communicative context, or with the community of speakers that speak their shared language at the time of her utterance.

She intends for her speech to eventually be understood as she wants it to be—that is, with its ameliorated meaning. The ameliorator has *diachronic communicative intentions*—communicative intentions that are not relevant to her context, but project into future communicative contexts and future linguistic communities.

Diachronic communicative intentions might be more common than one might think, even independently of the issues of amelioration and conceptual engineering discussed here. There are circumstances when speaking in this way seems justified and appropriate. For instance, parents and teachers often tell white lies to their children or students in order to aid their understanding of some difficult subject matter. They do so, knowing that the child (student) will eventually understand the full complexity of the subject, and that some white lies acted as a stepping stone to that understanding.

Conclusion

Changing language, while necessary, is difficult. This chapter has considered one important reason for its difficulty: it can often lead to miscommunication and confusion. I presented several ways we can use language which might lead to this sort of miscommunication and confusion. But then I argued that these supposed problems can actually be beneficial. It's *good* that changing language leads to miscommunication and confusion, because that can cause speakers to reflect on their language, and that will lead them to focus on its flaws and ways to improve them. I called this process 'transformative communicative disruption'. The sort of reflection transformative communicative disruptions can bring about is the sort of thing anyone interested in changing language for the better should care about fostering, and so we should embrace transformative communicative disruptions.

References

Anderson, Luvell, and Lepore, Ernie. 2013. What Did You Call Me? Slurs as Prohibited Words. Analytic Philosophy 54 (3):350—63.

Armstrong, Josh. 2016. The Problem of Lexical Innovation. *Linguistics and Philosophy* 39 (2):87—118.

Barker, Chris. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy* 25 (1):1–36.

Burge, Tyler. 1979. Individualism and the Mental. *Midwest Studies in Philosophy* 4 (1):73–122

Burgess, Alexis, and Plunkett, David. 2013. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Cappelen, Herman, and Dever, Josh. 2018. *Puzzles about Reference*. Oxford: Oxford University Press.

Cohen, A. 2001. On the Generic Use of Indefinite Singulars. *Journal of Semantics* 18 (3):183–209.

Dorr, Cian, and Hawthorne, John. 2013. Naturalness. In Karen Bennett and Dean Zimmerman (eds.), *Oxford Studies in Metaphysics*, Volume 8. Oxford: Oxford University Press.

Eklund, Matti. 2015. Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In Chris Daly (ed.), *The Palgrave Handbook of Philosophical Methods*. London: Palgrave Macmillan.

Eklund, Matti. 2017. *Choosing Normative Concepts*. Oxford: Oxford University Press.

Evans, Gareth. 1973. The Causal Theory of Names. *Aristotelian Society* Supplementary Volume 47 (1):187–208

Fallis, Don. 2009. What Is Lying? *Journal of Philosophy* 106 (1):29–56.

Frankfurt, Harry G. 1986. *On Bullshit*. Princeton: Princeton University Press.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Noûs* 34 (1):31–55.

Haslanger, Sally. 2007. 'But Mom, Crop-tops Are Cute!' Social Knowledge, Social Structure and Ideology Critique. *Philosophical Issues: The Metaphysics of Epistemology* 17:70–91.

Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press.

Jackman, Henry. 1999. We Live Forwards but Understand Backwards: Linguistic Practices and Future Behavior. *Pacific Philosophical Quarterly* 80 (2):157–77.

Kaplan, David. 1989. Afterthoughts. In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes From Kaplan* (pp. 565–614). Oxford: Oxford University Press.

King, Jeffrey C. 2014. Speaker Intentions in Context. *Noûs* 48 (2):219–37.

Krifka, Manfred. 2013. Definitional Generics. In Mari Alda, Claire Beyssade, and Fabio del Prete (eds.), *Genericity*. Oxford: Oxford University Press.

Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Leslie, S. J. 2015. 'Hillary Clinton Is the Only Man in the Obama Administration': Dual Character Concepts, Generics, and Gender. *Analytic Philosophy* 56 (2):111–41.

Lewis, David. 1969. *Convention: A Philosophical Study*. Oxford: Wiley-Blackwell.

Lewis, David. 1983. New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61 (4):343–77.

Lewis, David. 1984. Putnam's Paradox. *Australasian Journal of Philosophy* 62 (3):221–36.

Ludlow, Peter. 2014a. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford: Oxford University Press.

Ludlow, Peter 2014b. Norms of Word Meaning Litigation. *ProtoSociology* 31:88–112.

Mahon, James Edwin. 2015. The Definition of Lying and Deception. *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University Press.

McKinney, Rachel Ann. 2016. Extracted Speech. *Social Theory and Practice* 42 (2):258–84.

Paul, L. A. 2014. *Transformative Experience*. Oxford: Oxford University Press.

Plunkett, David, and Sundell, Timothy. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1–37.

Plunkett, David, and Sundell, Timothy. forthcoming. [title to be determined].

Putnam, Hilary. 1975. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science* 7:131–93.

Richard, Mark. 2014. Analysis, Concepts, and Intuitions. *Analytic Philosophy* 55 (4):394–406.

Richard, Mark. MS. *Meaning as Species*.

Saul, Jennifer. 2006. Gender and Race. *Aristotelian Society* Supplementary Volume 80 (1):119–43.

Scharp, Kevin. 2013a. *Replacing Truth*. Oxford: Oxford University Press.

Stokke, A. 2010. Intention-sensitive Semantics. *Synthese* 175 (3):383–404.

Stokke, A. 2013. Lying and Asserting. *Journal of Philosophy* 110 (1):33–60.

Thomasson, Amie L. 2016. Metaphysical Disputes and Metalinguistic Negotiation. *Analytic Philosophy* 57 (4):1–28.

Williamson, Timothy. 1994. *Vagueness*. London: Routledge.

21

A Pragmatic Method for Normative Conceptual Work

Amie L. Thomasson

I have argued elsewhere (2016), that we should see metaphysics as fundamentally involved in conceptual work, where this conceptual work includes not only descriptive conceptual work like that undertaken by such figures as Gilbert Ryle (1949), Peter Strawson (1963, 1992), or Frank Jackson (1998), but also—and often more interestingly—*normative* conceptual work.[1]

The claim that metaphysics centrally involves both descriptive and normative conceptual work can itself be read as either a descriptive or normative claim. Taken in the descriptive sense, the claim is that a great deal of historical and contemporary work in metaphysics can be aptly *interpreted* as engaging in (descriptive and/or normative) conceptual work. Often this must be a matter of 'deep' interpretation, since many metaphysicians think of themselves as discovering worldly facts when they do metaphysics. Nonetheless, many classic debates in metaphysics can be understood as—at least *implicitly*—involved in conceptual analysis or in negotiating what terms or concepts we should use. As I have argued elsewhere (2016, 2017a), seeing many past debates as implicitly involved in negotiating for what terms or concepts we should use often makes better sense of what competing metaphysicians are *doing* in their debates, and gives us new tools with which to assess and make progress on those debates. But I don't aim to make the case here that a great deal of historical and contemporary work in metaphysics can be so interpreted. While the job has been started elsewhere,[2] continuing work on it would require detailed case studies that must be left for other occasions.

[1] This echoes suggestions by others—classically, Carnap; more recently, Simon Blackburn, who writes that "just as the engineer studies the structure of material things, so the philosopher studies the structure of thought" (1999: 2). More recently still, similar views have been defended by David Plunkett (2015) and Matti Eklund, who suggests more broadly that "Philosophy should . . . be thought of as *conceptual engineering*" (2015: 36). Eklund argues that philosophy could be seen as conceptual analysis, but that it would be better to take it as conceptual engineering in the sense of "a study of what concept best plays the theoretical role of our concept of [e.g.] truth and what features this concept has . . . " (2015: 376).

[2] A. J. Ayer makes the case that Locke, Berkeley, and Hume can be seen as engaged in (descriptive) conceptual analysis (1946: chapter 2). In my view, not only ordinary language philosophy but also phenomenology can easily be understood as engaged in descriptive conceptual work (see my 2007)—including under this heading transcendental conceptual work that examines what concepts we *must* have

Amie L. Thomasson, *A Pragmatic Method for Normative Conceptual Work* In: *Conceptual Engineering and Conceptual Ethics.* Edited by: Alexis Burgess, Herman Cappelen, and David Plunkett, Oxford University Press (2020). © Amie L. Thomasson.
DOI: 10.1093/oso/9780198801856.003.0021

Taken in the normative sense, the claim is that we *should* come to understand the legitimate work of metaphysics as (descriptive and/or normative) conceptual work. If the case *can* be made that much historical work in metaphysics involves conceptual work, so much the better for the proposal that *we should* come to understand the proper work of metaphysics in that way. For that would lend plausibility to the idea that, even if this doesn't match the explicit self-conception of many contemporary metaphysicians, it maintains enough continuity with what both historical and contemporary metaphysicians have been *doing* to deserve the title. As I have argued elsewhere (2016) understanding the *legitimate* work of metaphysics in this way brings other advantages, too, such as avoiding epistemic mysteries and apparent rivalries with science.

But whether we take the claim that metaphysics centrally involves normative conceptual work as an *interpretation* of what is implicitly at stake, or as a *proposal* for how to understand what metaphysics legitimately *can* do (or both) an important question immediately arises. If metaphysics centrally involves normative conceptual work, how ought we to be doing it? What methods and standards should we employ?

While the idea that work in metaphysics may involve normative conceptual work has begun to gain some traction, far less has been said about *how* that work is to be done. Herman Cappelen aptly notes that we should be interested not just in how conceptual revision has been done, but "We should also be interested in how [conceptual revision] *should* be done—what *should* be considered successful" (2018: 119). Answering the methodological question of how it should be done is important. As Alexis Burgess and David Plunkett put it:

> if certain views about metalinguistic disputes are on the right track, we actually already engage in normative argument about representational choices much more often than one might realize. We would therefore do better to pursue these issues wittingly, overtly, and with greater care. (2013b: 1091–2)

And again:

> If we already practice conceptual ethics, let's do it well. (2013b: 1097)

But what sort of care should we take—what is it to do normative conceptual work well? Presumably there must be constraints, better and worse ways of engaging in this sort of conceptual choice. We do often find *critiques* of certain concepts plausible and powerful—whether critiques like Foucault's (2006) that show the uses of certain

to be able to reason or experience the world at all. Some more recent work on what it is to be a person, a work of art, or artifact can also be seen as engaged in descriptive conceptual work. Much revisionary work in recent metaphysics can be seen as engaged in normative conceptual work. Mark Richard (forthcoming) argues that many philosophical debates, for example, concerning knowledge, free action, truth... can be seen as engaged in something like conceptual engineering. David Plunkett makes the case that many philosophical disputes can be interpreted as not canonical disputes but rather cases of metalinguistic negotiations (2015: 861). Some even make it explicit that they are advocating for adopting a new conceptual scheme rather than analyzing the old conceptual scheme or discovering some worldly 'essence'. See, for example, Haslanger (2012) on race and gender; Millikan (1984) on function; Bell (1914/1977); and Davies (2004) on art. This also lies just beneath the surface in the debate between Price (2013) and Brandom (2013) about how to use the term "representation(alism)".

concepts to be driven inappropriately by power relations, or by feminist philosophers of science in showing how certain concept choices in the sciences have been driven by sexist assumptions, or critiques of past or current concepts of race (Appiah 1992). To the extent that concepts are open to critique, we must presuppose that there are at least some standards for evaluating concepts: standards that are not lived up to in these cases.

Comparatively little (that I have been able to find) has been said about how this sort of normative conceptual work should proceed. Burgess and Plunkett (2013b) formulate the question, list a range of goods (clarity, consistency, naturalness, social justice . . .) and goals (cooking, playing soccer, doing mathematics) that might be relevant in doing what they call 'conceptual ethics'. They also note that goals may play a central role in determining what goods matter, and what norms we should follow in conceptual ethics—"it's natural to think that different *goals* give rise to different norms" (p. 1105). But—as is appropriate in their agenda-setting paper—they leave most of these questions open, and don't propose anything like a unified approach to addressing problems in conceptual ethics. Even Peter Ludlow, who argues that normative negotiations of meanings are pervasive in everyday speech, says little about how these *should* be conducted, writing:

> Application conditions for terms and phrases like 'murder' 'life' 'family values', and 'good character' must be fleshed out and precisified, and it would seem to be a mistake to just blindly follow our neighbors or the powerful for their precisifications. Here is a place where we want to insist on deliberation and good reasons for a choice. (2006: xx)

What might such good reasons consist in? Here we get little guidance. Matti Eklund, similarly, having argued that philosophy is best seen as conceptual engineering, admits, "Obviously some big questions in the vicinity remain entirely unresolved, such as what the proper methodology for conceptual engineering is" (2015: 382).[3]

1. The Metaphysical Approach and the Pragmatic Approach

In this chapter I aim to sketch a method for approaching normative conceptual work, with an ulterior motive in view. My own interest in the idea that metaphysics can be understood as (to be) engaged in conceptual work—including normative conceptual work—comes from seeing this position as a way of adding strength and depth to the deflationary metametaphysical position I have argued for elsewhere (2015). Call a metametaphysical approach 'deflationary' if it relies on nothing more mysterious than (normative and descriptive) conceptual work—perhaps combined with straightforward

[3] Some readers may have noticed that some of the relevant works mentioned above speak of concepts, while others speak of language. There is as of yet no consensus on whether one should speak (primarily) of engineering concepts or language. I am inclined to think that it is language that is primary—for (parts of) language we can clearly see as historical and changeable and as having functions. Nonetheless, speaking of concepts is a useful way of abstracting from phonetic and orthographic features of words that are typically irrelevant to the engineering at issue in philosophical work. But I will not address these issues here, and will continue to speak of both below.

empirical work—in answering those metaphysical questions that are well formed and fit to answer. Call a metametaphysical approach 'heavyweight' if it presents metaphysics as involved in discovering deep facts about the world that are not knowable by employing straightforward empirical and/or conceptual means—but are rather (in the words of Theodore Sider) 'epistemically metaphysical' (2011: 187). By acknowledging the role of *normative* (as well as descriptive) conceptual work, I have argued (2016), we can retain the epistemic clarity of a deflationary metametaphysical approach and yet also respect the depth, worldly relevance, and difficulty of many metaphysical disputes.

But accepting that a great deal of the interesting work in metaphysics has involved and should involve normative conceptual work does not *commit* one to a kind of metaontological deflationism, as Plunkett makes clear. Having argued for the related point that many philosophical disputes (including disputes about ground, supervenience, and essence) turn out to be cases of metalinguistic negotiation (2015: 859–60), he goes on to emphasize that:

> Everything I have said is consistent with thinking that there are important issues [in metaphysics, concerning ground, supervenience, real definition, essence, etc.] here to investigate. To see this, consider the following. Suppose one argued... that an important part of communication among biologists involves metalinguistic negotiation. (The different meanings of 'species' is a good place to start with such a proposal, as is the different meanings of 'intelligence'). Would that mean that there aren't facts about animals and their behavior to investigate, and that all biological argument is just about normative issues about word and concept choices? Clearly not. (2015: 860)

For, as he (aptly) emphasizes, our views in conceptual ethics are generally tied up with commitments on a range of first-order normative and descriptive claims about the world.

Once we embrace the idea that much metaphysical work involves normative conceptual work, the question of whether we should adopt a deflationary or heavyweight metametaphysical approach comes down in large part to the question of what methods we can and should use for undertaking normative conceptual work.

One might take either of two broadly different orientations to answering this methodological question—one of which is tied to a heavyweight approach, and the other of which is consistent with a deflationary approach. Either approach could no doubt be fleshed out in many ways, and there may still be other approaches to consider.[4] Call the first of these approaches the 'metaphysical' approach. Someone who takes the metaphysical approach to conceptual ethics, in broad strokes, is someone who accepts that many debates in metaphysics may be seen as negotiating which terms, concepts, or conceptual scheme we ought to use, but who thinks that it

[4] One might also take a transcendental or Kantian approach, for example, to the question of what *basic* concepts we must have—say, in order to have cognition of objects at all. (Thanks to Lanier Anderson for bringing this option to my consideration, and to Jessica Leach for further discussion.) One might be able to consider a transcendental approach as part of the pragmatic approach—but a part on which the choice of certain concepts is determined by certain requirements that are non-optional (and so not 'merely pragmatic', in the everyday sense). This possibility suggests that, in taking a pragmatic rather than metaphysical approach, one need not be committed to the view that all conceptual choices are *merely conventional*, or built on contingent practical needs.

is the metaphysical facts of the world that provide the primary standard, for example, suggesting that we should aim to choose concepts that correspond to things that exist, carve the world at its joints, or correctly reflect the essences of the things described.[5] Such views remain meta-metaphysically heavyweight insofar as, even while acknowledging that some of the central work of metaphysics involves normative conceptual work, they hold that this ought to be driven by the metaphysical facts about the world—where these are not discoverable by straightforward empirical work or conceptual analysis. I have argued against such metaphysical approaches to conceptual choice elsewhere (2017b), on grounds that they leave us with epistemological mysteries that are hard to resolve.

In any case, it is clear that the metaontological deflationist cannot say that our normative conceptual choices ought to be determined by the metaphysical facts of the world. Nor is the metaphysical approach open (on any topic) to those who are deflationary *about that topic*. For example, Sally Haslanger notes that on her approach "the world by itself can't tell us what gender is, or what race is" (2012: 224). Similarly, metaethical deflationists cannot treat our conceptual choices in ethics as requiring guidance from heavyweight ethical facts.[6] Thus—given the difficulties faced by the metaphysical approach, as well as the need many will have for a non-metaphysical approach—there is good reason to investigate the prospects for a non-metaphysical approach to conceptual choice.

Those with heavyweight metametaphysical inclinations are prone to see a threat here: if the metaphysical facts of the world cannot determine which of our concepts or terms we ought to use, one might worry that we will be unable to account for intuitions that the world has *structure*, and that accordingly some conceptual choices (e.g., lithium rather than *lithium on earth*; fish and mammal rather than sea creature and land creature; green rather than grue) are *objectively better* than others (Sider 2011: 18–19). But, as Theodore Sider puts it, "It is really, really hard to believe that the fact that electrons go together, in a way that electrons-or-cows do not, is merely a reflection of something about us" (2011: 18). Not only might we be left unable to account for intuitions that the world has structure; a broader worry arises. That worry is that if we do not take a metaphysical approach to conceptual choice, our conceptual choices will be merely arbitrary, not constrained by the world. As David Plunkett describes the worry (without endorsing it), this view might be thought to suggest "that normative issues about what concepts we should use can be settled by voluntary choices that we ourselves make" (2015: 860–1). Finally, the deflationist might seem unable to make sense of the plausible idea discussed above, that there are standards governing conceptual choice, which justify us in critiquing certain concepts.

In short, then, abandoning a metaphysical approach to conceptual choice leads to three interrelated worries: (1) That we will be unable to account for intuitions of structure; (2) That we will have to treat conceptual choices as merely arbitrary, not

[5] Peter van Inwagen (2016: 17) suggests something like the view that the metaphysician's conceptual choices ought to be driven by facts about what exists, while Theodore Sider (2011) develops and defends the view that (at least when doing fundamental metaphysics) our conceptual choices should be held to the standard of carving the world at its joints or mapping 'structure'.

[6] Thanks to David Plunkett for emphasizing this point.

worldly; and (3) That we will be unable to critique conceptual choices. If the deflationist is really saddled with these problems, it threatens to leave the deflationist's view of metaphysics (as engaged centrally in normative conceptual work) inherently uninteresting, or even suspect—with each practitioner simply attempting to wield her power in imposing her own perhaps idiosyncratic conceptual choices, incapable of worldly validation.

But these fears are misplaced. There is a far better option available for the deflationist to take: namely, adopting a pragmatic (rather than metaphysical) methodology for normative conceptual work.

We can think of normative conceptual work as involving two projects. One is conceptual *engineering*: that is, holding in place some function or functions the concept is to serve, we may aim to redesign old concepts to serve that function better, or to engineer new concepts that can serve a function that was done imperfectly or not at all by our prior conceptual scheme. But we can also engage in work in conceptual *ethics* proper. Burgess and Plunkett take conceptual ethics broadly, as involving 'normative and evaluative issues about representation' (2013a: 1091), including deeper normative questions regarding what functions our concepts *should* serve, which functions we should pursue and abandon, and more generally "which concepts [we should] use to think and talk about the world" (Burgess and Plunkett 2013a: 1091). This work in conceptual ethics can also be undertaken at two levels. First, keeping fixed some goals we wish to fulfill or see as desirable, we can ask what functions our concepts should serve, to help fulfill these goals. This will issue in hypothetical imperatives: if you have these goals or purposes, then you should employ this range of concepts (using them in this way). Second, there are deeper questions one can raise about which goals we should adopt, and so about which concepts should we employ, all things considered (not just given some presupposed range of common goals and interests). I will leave these questions to the side here. For one thing, addressing such questions would bring us into deep metanormative issues that are beyond the scope of this chapter. For another thing, we (fortunately) do not need to appeal to such deep normative claims to do the work needed here. For I will argue that, *presupposing various widely shared goals that are generally presupposed in debates about what concepts to use*, we can fully account for ordinary intuitions about structure, as well as for the non-arbitrariness, worldliness, and openness to critique of our conceptual choices. In that way (even without commitment to any deep normative claims), we can account perfectly well for the metaphysician's driving intuitions that that some conceptual choices are better than others, and give a picture on which (given those shared interests) conceptual choice is not at all arbitrary, and is open to critique.

The key to developing a plausible pragmatic approach, as I will argue below, is to take the functions of our (ranges of) concepts as playing a central role. As I shall argue below, when we aim to engage in normative conceptual work, we must pay close attention to the purpose or function that is (to be) served by the relevant term, concept, or conceptual scheme.[7] To engineer a concept well, we must attend to its

[7] In asking these functional questions, we need not take a stand on whether the concepts and terms we are investigating are natural products of evolution, explicit artifactual creations, or something in between

function. To do first-level conceptual ethics well, and determine what concepts we should adopt, to meet goals that we have, we must attend to function.

By appealing to the idea of function, I will argue, the pragmatic approach can respect our ordinary intuitions about structure, give constraints for conceptual choice that ensure that conceptual choice should not be seen as merely arbitrary or subjective, and must be responsive to worldly constraints (though it is responsiveness to empirical facts, not special metaphysical facts, that is needed), and leave room for critique. As a result, I will argue, those inclined to think that the work of metaphysics centrally involves normative conceptual work may combine that with a plausible methodology consistent with a deflationary metametaphysics. The prospects for a pragmatic approach to conceptual choice are far better than its critics have suspected.

2. A Defense of Function

But although it seems natural to think that our concepts (or perhaps better: ranges of concepts) serve certain functions, and that these are relevant to the projects of conceptual engineering and conceptual ethics,[8] Herman Cappelen rejects this function-driven view, denying that concepts or words have functions: "I don't think concepts have purposes and certainly not words (or extensions or intensions)" (2018: 180). Cappelen argues that the appeal to the function or purpose of a concept doesn't "do any work"—in particular, that it fails to provide an adequate answer to Strawson's challenge against conceptual engineering. Thus, before going on to develop the function-driven pragmatic approach below, we should confront these doubts.

Carnap aimed to replace certain everyday concepts with "exact and fruitful" concepts for use in the formal or empirical sciences. Strawson's challenge against Carnap is that any attempt to solve traditional philosophical problems involving concepts used in non-scientific discourse by engaging in conceptually engineering exact scientific concepts would not be "to solve the typical philosophical problem, but to change the subject" (1963: 506). For "the clarification of philosophically puzzling concepts is not achieved by the introduction of related scientific concepts" (p. 506). For, Strawson insists, shifting to the new scientific concepts may simply fail to address the old philosophical perplexities (pp. 504–5).

Cappelen interprets Strawson as presenting a general challenge to the tenability of conceptual engineering, generalizing it as follows:

> Generalized challenge: Change of extension and intension . . . is a change of topic, so revisionary projects [in conceptual engineering] are bound to fail. Even if the revisions succeed, they do not provide us with a better way to talk about what we were talking about; they simply change the topic. (2018: 100)

(as with the nests of birds or cities of ants). In any of these cases, we can profitably ask questions about the function of the heart, the toaster, or the nest.

[8] This idea has also been suggested by others working on conceptual engineering, including Brigandt (2010) and Haslanger (2012).

For if we change the conditions that need to be satisfied in order for something to fall into the extension of our concept as we engineer it, the objection goes, we can't claim to have *improved* the concept—we will have changed the concept and thereby changed the topic in such a way that we can't even be answering the questions posed using the original concept. Mark Richard (forthcoming) argues similarly that one can't improve on a concept by changing its intension and extension, since concepts possess these essentially, ensuring that any such changes leave us with new concepts rather than improvements of the old concepts. The challenge, then, is to find a way to understand sameness of topic in such a way that one can allow that conceptual engineering enables us to improve our concepts without changing the subject.[9]

Cappelen takes the question "does conceptual engineering always involve topic revising... or can conceptual engineering in some cases preserve topic" as "a central question for anyone interested in conceptual engineering and its foundations" (2018: 97). One of his recurrent objections against the view that words or concepts have function is that, he claims, appeal to a function won't help answer this question.

The first thing to say in response to the generalized challenge is that we shouldn't take the challenge (as generalized by Cappelen) with high seriousness and feel pressed to search for a univocal answer—to whether we "really" have the same concept or topic as before. The generalized challenge presses us to say when terms are the same, when concepts are the same, when a topic of conversation is the same. But it is entirely coherent with the conceptual engineer's point of view to think that the key terms here: 'term', 'concept', and 'topic', and the like, are themselves underspecified in ordinary language and up for metalinguistic negotiation and re-engineering. In general, as I have argued elsewhere (2017a) (following Plunkett and Sundell 2013), debates about what is/is not essential to Ts are often debates (disguised in the object language) about how we ought to use the relevant terms ('T'). The deepest, though not most direct, response to the generalized challenge is to urge that we not presume that there is an objectively correct 'discovery' of what does/does not count as sameness of topic, concept, or term. What we count as sameness of concept or term may aptly be engineered or negotiated differently depending on the purposes we have. Sometimes (say, in doing etymology) we may wish to track historical continuity; sometimes we might need track sameness of extension, inferential role, or even phonetics or lexical effects... In other cases, including many of those centrally at issue for conceptual engineering, we may want to track sameness of function. In short, the best response seems to me to be a bootstrapping response that begins by asking what function we want the relevant terms (including 'same word', 'same concept', or 'same topic') to serve and presses for a view that will do that job well.

The purpose that is at issue in Strawson's challenge is to give us a way of understanding *concept* that can preserve the sense in which people are 'talking about the same subject' over time—not just 'changing the subject'. To do this, it seems we might do better to look to function and historical continuity in individuating concepts than to

[9] Cappelen himself does this by an appeal to same-saying (2018).

rely on precise intensions and extensions. For appealing to function provides a promising way of giving a sense in which we remain on topic across changes in intension and extension—a sense in which we aim to solve the same problem, or to pursue the same goals.[10] Consider, for example, recent revision of our concept of marriage to include same-sex couples: how can we consider it an improvement to our old concept of marriage rather than a simple change of subject? One way to do so is by appealing to the continuous function (or functions) the concept of marriage was to serve. Why it is useful to have a concept like *marriage*: what legitimate role(s) might it play (perhaps along with other social concepts) in our overall conceptual system, and what would we be missing if we lacked such a concept? If we suppose that one legitimate and desired function of a concept of marriage is to mark a range of close relationships that we would help protect by affording a special legal and social status (tied up with some 3,000 relevant legal obligations and entitlements in the U.S.), then one can see that function as served—and served even better—by extending the criteria to include same-sex relationships that otherwise are similar in character to those previously included in the extension. In that way, we can see the change as a conceptual improvement, rather than a mere change of topic. So thinking of concepts or words in functional terms provides ways of legitimating the feeling that we haven't simply 'changed the subject' when we engage in conceptual engineering.

Nonetheless, Cappelen opposes appealing to function to do this work. He initially considers and rejects two versions of a functional view:[11] Haslanger's appeal to the 'central functions of a concept' (2000: 35), and Brigandt's appeal to a concept's epistemic goals (the kinds of inferences and explanations) a concept was intentionally introduced to serve (2010).[12] Haslanger argues that shifts in the meaning of a term are semantically justified "if central functions of the term remain the same, e.g., if it helps to organize or explain a core set of phenomena that the ordinary terms are used to identify or describe" (2000: 35). Against this, Cappelen argues that the only non-controversial approach to identifying the relevant 'core set of phenomena' is disquotationally: for example, that the concept of salmon is to identify or describe salmon (2018: 183). But this clearly won't help with answering the generalized challenge, of saying how we could still be talking about the same things, after we have revised a concept and changed its extension. Otherwise, Cappelen suggests, we simply need more guidance about how to identify the relevant phenomena a concept was to identify or describe—and Cappelen adds a suspicion that "there simply *isn't* a good way to identify 'the phenomenon' except disquotationally and the disquotational identification is unresponsive to the challenge of articulating the limits of revision" (2018: 184).[13] The other approach to identifying conceptual function that

[10] Plunkett and Sundell (2013), and Warren (2015), provide a good account of this for the moral case.

[11] Cappelen also considers contextualist approaches that hold that the function of a concept varies from concept to concept, but since this is clearly a non-starter for solving Strawson's challenge of accounting for sameness of topic, I will leave it to the side here.

[12] Cappelen attributes this view to Brigandt, though noting that Brigandt makes only the limited claim that some scientific concepts have such epistemic goals—not that all concepts, or even all scientific concepts do.

[13] To be fair, Haslanger does point to ways of identifying relevant phenomena in the case of race and gender, including the need to "identify and explain persistent inequalities between females and males, and

Cappelen considers is from Brigandt, who appeals to the idea that conceptual change in science can be understood as rational by appealing to its epistemic goals: the kinds of inferences and explanations the concept is intended to support. This of course, as Cappelen notes (2018: 185), is too narrow to apply generally as a view of the function of a concept. One might attempt to develop this view more broadly by identifying the function of a concept with whatever function it is intended to serve. But this will clearly be problematic as well, since few concepts are intentionally designed at all: any that are innate or that gradually evolve in a community rather than being explicitly designed and introduced for a purpose will not be intentionally designed to serve any function.

This needn't be a worry, however, for those inclined to appeal to function in laying out a method for conceptual ethics. We need only attend to the recent philosophical work on function to see that there are more plausible options available, which don't identify function with intended function, and aren't left with a mere disquotational story about the function of words or concepts. There have been two large trends in understanding the notion of function in post-Darwinian biology: in terms of evolutionary/selection history (a historical story about what the ancestors of such things did that accounts for their reproduction and survival) ('proper function'), and in terms of a thing's current capacities/dispositions, with particular attention paid to the role such things play in the overall system in which it is embedded (what Beth Preston calls 'system function') (1998: 221). In neither case do we need to identify function with what anyone intends the function to be.

The notion of proper function has been most famously developed by Ruth Millikan, who is not concerned merely with biological functions, but rather explicitly aims to identify functions of 'language devices' that are "not found either by averaging over idiolects or by examining speaker intentions" (1984: 4). On Millikan's view (roughly), a member of a 'reproductively established family' has as its proper function whatever its ancestors did that contributed to the reproductive success of the family, which contributes to explaining the existence of that member (1984: 28). Millikan explicitly applies the view to cultural products, including language, as much as to biological entities such as hearts and lungs. Meaningful linguistic devices, on her view, are also members of "first-order reproductively established families" (p. 29), and Millikan argues that "language devices must have direct proper functions at some level or levels. It must be because they correlate with functions that they proliferate" (p. 31). This is clearly a view on which the (direct) proper function of a concept or term need not be identified with anyone's intentions or beliefs about what the function is, nor with actual dispositions of speakers to use the term in certain ways, or an average over the relevant ways in which, or purposes for which, it is used. It is thus a view that avoids Cappelen's objections.

The other dominant approach to function is to adopt a 'system function' view like that defended by Robert Cummins. On this model, the function of an item is whatever it *does* within the system as a whole—whatever its current capacities

between people of different 'colors'", and to be "sensitive to both the similarities and differences" among people considered to be male/female or of different 'colors' (2000: 36). So, it seems further discussion should focus on these suggestions, and whether and if so how more general guidance can be given.

contribute to the capacities of the whole, so that we can give a compositional analysis "of the capacities of containing systems in terms of their component parts" (Preston 1998: 225). As Preston argues, these two notions of function—as proper function and as system function—must not be conflated, but we also need not consider them rivals. Instead, we may need to recognize them as "equally important for a viable general theory of function"—whether we are concerned with the functions of biological entities or of artifacts (p. 226). Typically new proper functions begin life as *system* functions: it is because these entities can do something for the system that they (and their later copies) tend to be kept around; that is why identifying what a range of vocabulary *can* do (its system function) can be a useful tool in determining its *proper* function.

Cappelen, however, also rejects the idea that we can usefully identify function in these ways.[14] He argues, first, that we can't identify the functions of vocabulary, considered as *words*, by asking what it "enables us to do that we couldn't do (or couldn't do as effectively or efficiently) without it". For we could clearly do the same work by exchanging one symbol ('1') with another ('2') (2018: 187). But this misses the overall point. The point is to ask (e.g.) what *nominative number terms* (or ethical terms, or a truth predicate) do for us that we couldn't do, or couldn't do as effectively or efficiently without them. The question clearly isn't a matter of what this typographical shape type can do for us. Nor are functional analyses that identify what hearts or forks do for us, which might account for their being reproduced, undermined by imagining that the same work might be done by an artificial heart or a differently shaped piece of silverware. The relevant counterfactual asks us to evaluate what we couldn't do as effectively without hearts or forks, *holding other aspects of the background context in place*—not while we vary the background by providing a substitute to do the job. In any case, we can clearly avoid the above worry by rephrasing, asking: what a range of vocabulary "enables us to do that we couldn't do (or couldn't do as effectively or efficiently without it, *or an apt translation*)".[15]

While Cappelen expresses some willingness to accept that we can identify functions "by looking for what makes [terminologies] useful for us (and hence perpetuated in our culture)" (2018: 187), such functions, he suggests will be nothing more than disquotationally specified functions, such as "the reason 'salmon' is useful for us is that it can be used to talk about salmons (or denote salmons)" (p. 187). For "beyond these disquotationally specified functions, there's variability" in how the terms are used in different speech acts (p. 187).

But this response misses two important points. The first is a functional pluralist point, which (following Huw Price (2011)) I have emphasized elsewhere: that we shouldn't assume that all terms have as their function (or even among their central functions) to track or denote entities of a certain kind.[16] As I discuss in section 3 below, many of our most philosophically interesting terms (such as mathematical,

[14] What follows in this section responds to the published material in Cappelen (2018), which was criticizing material found earlier in an earlier draft of this chapter.

[15] Another option is to speak in terms of concepts rather than words here. I aim to remain neutral here regarding which provides the best approach.

[16] The tracking or denoting function is what Price (2011) calls an 'e-representational function'.

moral, and modal terms) may plausibly serve very different functions from this kind of denoting or tracking function that can be given disquotationally.

The second important point is that variability in how something is used does not entail that nothing informative can be said about its function. The parallel argument would never be accepted for holding that biological or artefactual entities can't be said to have functions, in a way that can be given substantively and informatively. What is the function of a dog's mouth? There are a great many things that can be, and have been, done with them—but that does not show that the mouths of dogs can't be said to have a biological proper function, or functions. The case is even clearer for artifacts: there are a great many things that can be, and have been, done with forks, or screwdrivers (they are exapted in all sorts of ways), but that does not show that forks or screwdrivers do not have proper functions (Preston 1998). For a proper function is not identified just by looking to anything that *can be* or *has been* done with the item in question. If we think of a language, and the terms in it, as human creations—as artifacts (abstract artifacts, in the sense I have elsewhere articulated (1999)), it is natural to think that linguistic items, too, may have proper functions, identifiable separately from the diverse uses to which they are put. I have elsewhere[17] distinguished the 'practical significance' or proper function of a range of vocabulary (the function it serves in our overall linguistic apparatus, which explains why it is useful to have vocabulary like that in place), from that of its use in different speech acts (what it is used to do on particular occasions). Even where uses vary, a more stable proper function or functions may be identified.

In any case, I want to leave open what view, precisely, of function should we adopt with respect to the functions of our terms (or ranges of vocabulary) or concepts. That is a major topic for discussion in its own right. A great deal of work remains to be done in determining how we should understand the notion of linguistic function most relevant to normative conceptual work, and how we can best discover the relevant functions of our concepts, terms, or ranges of vocabulary.

The central point for now is that we can legitimately maintain that our terms and concepts have functions, without thereby having to think of all functions as intentionally endowed, and without having to limit them to functions that can be understood disquotationally. I have tried to at least point towards some familiar ways of understanding 'function' that might help us do the job here. With that defense in place, we can return now to our story—and try to utilize this appeal to function as a way of sketching a pragmatic method for undertaking normative conceptual work.

3. A Pragmatic Method for Normative Conceptual Work

Here I aim to sketch the beginnings of a pragmatic approach to normative conceptual work—and to show that such approaches have every prospect of accounting for core

[17] In my *Norms and Necessity* (forthcoming), following Michael Williams (2010), who uses slightly different terminology.

intuitions about structure, and of avoiding accusations that such approaches must leave normative conceptual work arbitrary, subjective, or insusceptible to critique. The key, as I have already suggested, involves beginning from an appeal to function. As Strawson puts it, "The kinds of concept we employ are not independent of the kinds of purpose for which we employ them" (1963: 506).

While Carnap was interested in conceptual engineering, primarily in the sense of devising new, technical languages, most work in metaphysics (traditional and contemporary) does not involve devising new terms, but rather working with, and making normative choices regarding, common terms of our long-familiar vocabulary. Thus, if we think of metaphysics as engaged in conceptual negotiation regarding terms such as 'freedom', 'person', 'art', 'good', 'responsible', 'number', 'property', 'species', and the like, then we must acknowledge that these are terms that are *already* part of our shared vocabulary and conceptual scheme—not terms we do or can engineer on a blank slate. So how should we begin?

3.1. Reverse Engineering

Revisionary work has been increasingly popular in metaphysics, sometimes including recommendations that we do away with certain ranges of vocabulary—often in order to 'avoid' 'problematic ontological commitments'. But before removing a piece of a car engine, lines in a software program, or an organ from the body, it is always a good idea to begin with reverse engineering: working out what the part does for the engine, program, or organism as a whole. Thus, in conceptual engineering as it is (to be) practiced in philosophy, we must often begin not with simple constructive conceptual engineering, but rather with 'reverse engineering'.[18]

In some cases, we may get important clues about the functions a range of vocabulary has served by engaging in conceptual genealogy—looking back to when and why the term was introduced, how it has been used, and what functions it served in its original and later historical contexts.[19] As David Plunkett argues, "conceptual history can help us when we engage in conceptual ethics" (2016: 59). Understanding conceptual history can not only help us determine what functions those concepts have served, but also (Plunkett argues) might help in (re-)evaluating our purported justifications for using those concepts, help us in engaging in conceptual analysis, and thereby aid us in determining which concepts will be the most helpful for a given sort of inquiry.

But whether one aims to discover the function of a term or of other artifacts, such historical information is not always available, and gives only defeasible clues about how the relevant item *currently* functions. Even in those cases, however, one may engage in reverse engineering. But how is reverse conceptual engineering to be done? Consider how one approaches reverse-engineering a piece of software—say, a discovered piece of malware. One may have no access to historical information about how and why it came into being. Instead, one proceeds first by investigating what it

[18] I am indebted to David Sanford for suggesting this point.

[19] For an excellent overview of the role of conceptual genealogy in philosophy, see Dutilh Novaes (2015). For a general defense of the idea that conceptual history may play a useful role in undertaking work in conceptual ethics, see Plunkett (2016).

does and can do, and thereby gains clues to determine what (system) function(s) it serves and how it serves them. So similarly, in engaging in conceptual engineering one may aim to engage in reverse engineering the concepts in question—aiming to determine what they do or can do—why it is useful to have such (ranges of) concepts at our disposal, what we can do better with such a range of concepts than if we lacked it. For that identification of system function might provide an explanation of why ranges of vocabulary to express these concepts emerged and persisted—providing evidence of their proper function.

It is not hard to find philosophical analyses of functions that fit in these molds. Consider, for example, Stephen Yablo's (2005) analysis of what we can do using mathematical discourse that we couldn't otherwise. On his analysis, introducing noun terms for numbers enables us to simplify our statements of laws in certain effective ways—so that we can state in finite form laws that otherwise would take an infinite series of infinitely long sentences. Or consider the expressivist's analysis of the point of moral discourse, as enabling us to express and coordinate our attitudes in ways that put pressure on certain forms of agreement that thereby enable us to better live together. Paul Horwich's (1999) view of the role that the truth predicate serves as a device of generalization can also be understood in this light. In each case, these analyses purport to identify something that this range of concepts *does* or (better) *enables us to do*, that we couldn't do (or couldn't do as effectively or efficiently) without it (or an apt translation). Analyses like these can serve as clues to proper function analyses: to why it would have been useful to have concepts like this, why terms that express them might have been perpetuated in our culture. Nonetheless, it is important to emphasize that saying that a concept has a function is not to say that it is overall beneficial, aids the general utility, or anything of the sort. Some concepts, or ranges of concepts, may serve a function not *for us*, but rather for *some*: perhaps for those in power, who have the power to keep them in use. Of course, this does not make terms different from other artifacts—say, weapons, poisons, fences, elitist school systems, gendered clubs—which may serve functions for some, without being beneficial overall.

Engaging in conceptual genealogy and/or reverse engineering may yield various important results. On the one hand, we may find a useful function a range of vocabulary has served for us. For example, if Yablo is right about the useful functions of using nominative terms for numbers, and we have purposes that would be well served by being able to state scientific laws of these sorts in finite form, then we should hesitate before we suggest doing away with the vocabulary and 'making do' with some replacement nominalist language on grounds of alleged 'ontological' concerns. (Similar points could be made about mentalistic vocabulary, property talk, truth talk, etc.) On the other hand, we may find that the terms have served an insidious function that we don't think should be served—I will come back to this when we return to discussing critique below.

3.2. Identifying the Function to Be Served

A second crucial step is to undertake work that is more explicitly in conceptual ethics: determining what functions (if any) these concepts *should* serve, and are *to* *serve* going forward, given the goals and purposes we have. In some cases, this may involve no change (supposing the original concept served a function or functions that were useful, and that we still desire to be served). In other cases, it may involve proposing changes. In the most radical cases, it may involve proposing that we drop or replace a range of concepts or terms entirely, if we find that they have no hope of serving their function (say, in the case of terms for failed scientific posits such as phlogiston or Vulcan), or if we find that they have served an insidious function that we want to abandon.[20] For example, one might argue for rejecting pejorative terms from our vocabulary on grounds of their serving functions of insult and exclusion, which (given our current and public purposes) we think should be dropped. In other cases, the second step may involve a proposed shift in function, while retaining the old terms. Sally Haslanger's work on race and gender concepts makes this move explicitly. While the old concepts may have served to give artificial pseudo-scientific legitimacy to discriminatory practices, Haslanger argues that concepts in this vicinity *should* be retained to serve a different function: serving as "effective tools in the fight against injustice" (2012: 226). For without race concepts, it is difficult to address questions about the consequences being African American, say, has on one's job opportunities, educational prospects, treatment by the police, and so on—and thus hard to identify, and aim to fight against, racism.

3.3. Engineering to Serve the Function

Once we identify what function the relevant concepts are *to serve*, we can do far better at engaging in the third step, of constructive conceptual engineering. It is no good engineering a boat, or deciding which boats should be kept, modified, or replaced without a clear idea of whether the boat is to function in providing a fast and nimble escape from police boats, in transporting masses of heavy cargo across the wide oceans, or in undertaking exploration in the icy arctic. Some features would require repair for any boat—failure to float for example. So similarly some features of a concept would require repair regardless of purpose—such as its propensity to ensnare us in contradictions.[21] But no detailed evaluation can be made without an assessment of the functions that are to be served. Once a purpose (or multiple purposes) is/are identified, we can go on to use that in engaging in conceptual engineering—determining what sorts of rules or constraints would best (or better) enable it to fulfill its function(s), going forward.

[20] This also enables us to properly criticize other cases in which one might engage in conceptual negotiation with merely personal goals in mind—as with the case Burgess and Plunkett raise of Karl, a politician who (for purely personal gain of getting elected) employs traditional race and gender concepts rather than ones that would better serve social justice (2013b: 1105–6).

[21] However, Alexis Burgess' (2013) interesting arguments that we may, for example, have reasons to retain our concept of truth 'as is', even if its meaning is given by principles that classically entail a contradiction, should give us pause before assuming that even inconsistency always is sufficient to justify revising or replacing a concept.

4. A Defense of the Pragmatic Approach

Bringing functions into account enables us to develop a pragmatic approach to conceptual ethics (taken broadly as including practical work in conceptual engineering) that clearly avoids the problems often thought to plague deflationary methods.

4.1. Preserving Core Intuitions about Structure

First, it enables us to give due respect to the idea that the world is structured into natural kinds to which our concepts should be responsive, and to the idea that some concepts seem 'objectively' better than others. For it is plausible that some words serve something like a joint-tracking function—and are best left serving that function. The prime candidates for these 'joint-tracking' terms are the terms David Lewis and D. M. Armstrong began with in developing the joint-carving idea: predicates that aim to pick out 'natural' properties and relations. These are those that will figure in laws (Armstrong) and in our natural-scientific theories. Since these terms have the function of serving in explanatory and predictive scientific theories, which in turn aim to predict and explain, there are worldly constraints on what concepts we ought to adopt.

Given the relevant function and the constraints that come with it, it is easy to see why we do better to have in our chemical theory the current chemical concept of 'lithium' rather than a concept that would apply to lithium on earth, but not to the same chemical kind if found on Mars (cf. Sider 2011: 7). Geographic constraints in themselves are not helpful to chemical explanations and predictions, so the limited 'lithium' concept would not be as useful in a chemical theory. Similarly, it is easy to see, on these grounds, why the concepts of 'fish' and 'mammal' would be more useful than 'sea creature' and 'land creature' to serve the function of figuring in explanations and predictions in biology: more predictions of behavior, internal construction, disease susceptibility, reproduction, and so on will be facilitated by use of the former concepts than the latter concepts.[22]

I have respect for natural kind structure. I have a child with a nut allergy. It is a matter of life and death ("death in seven minutes", her allergist tells us) whether something is biologically a tree nut or is something *called* a 'nut'. It is a matter of life and death because it enables us to *predict* whether ingesting something will cause a life-threatening allergic reaction. It is not just a subjective matter whether 'tree nut' is a better concept than one that includes all and only things *called* 'nut' (including hazelnuts, peanuts, coconuts, nutmeg, and doughnuts (only the first of which is biologically a tree nut), and excluding cashews, pistachios, and almonds). That one concept but not the other is usefully and efficiently *predictive* in this way, which has life-or-death consequences, is all I need to be fully convinced that one set of concepts is objectively better. Moreover, the choice of concept of course has other useful consequences beyond predicting allergic reactions—consequences for its use in biological theory, farming, government regulations, etc. The important thing to

[22] On the other hand, as Sundell points out, where the concept of *fish* is employed with a different role in old seafaring contexts (aboard whaling ships, say), it is far less clear that the current biological concepts are better (2011: 14–15).

note here is that there *is* a way to justify the claim that one concept is better than another, for worldly reasons, but that our choice of concepts is vindicated *empirically*, given our shared purposes—it does not require additional *metaphysical* vindication.

The pragmatic approach to conceptual ethics can and does take into account all of these perfectly objective, worldly, empirically-driven reasons for choosing one set of concepts over another, *where these concepts are designed to figure in our explanatory and predictive theories*. And the advantages are not just in the simplicity of stating our biological theories, but also in the living of our lives, communication with others, safety, formulation of laws of state (as well as expression of laws of nature), etc. We can thus take into account our ordinary intuitions that some concepts are 'objectively better' than others, given a widely shared and generally unquestioned set of purposes: in this case, the purpose of designing theories that enable us to better predict and explain. In these cases, clearly our normative conceptual choices must also be world-responsive—in an empirical way. For when concepts (such as natural kind concepts) are designed to be useful in our empirical explanatory and predictive theories, we are thereby committed to being deferential to the world—letting experimental evidence help determine which *do* best serve in our predictions and explanations. Those concepts that function well in these explanatory and predictive roles will tend to be those we think of as 'carving at the joints' in the ordinary sense of marking those similarities and differences that are most relevant to our overall body of predictions and explanations, and so can preserve the everyday sense in which we think of the world as 'structured'.

Sider insists that "Joint-carving thought does not have merely instrumental value" (2011: 61) but rather is a constitutive aim of inquiry. But the instrumental value of employing concepts in our scientific theories that are particularly useful in predicting and explaining is sufficient to account for the ordinary intuitions about structure used to motivate the theory: that we ought to employ concepts like the purely chemical concept of water, the biological concept of mammal, and even the color concept of green (rather than grue). But it does so without positing extra metaphysical facts about 'structure', without invoking epistemic difficulties about how we could know such facts, or about why such metaphysical facts should be theory-guiding.[23] Once we can account for the worldliness and objectivity of criteria for conceptual choice in these cases, it's worth asking how very powerful are any remaining intuitions about 'real structure' that aren't accommodated in this way, and whether we need any further metaphysical vindication of our intuitions that some concepts are objectively better than others.

4.2. *Respecting Non-Arbitrariness*

However, we should not assume that all terms serve the same function: of tracking features of reality that enable us to better explain and predict.[24] Even in cases where

[23] On the last point, see Dasgupta (2018).

[24] Sundell accepts that 'metaphysical naturalness' serves as a norm governing the use of our scientific terms, but he too insists that "across a wide range of activities, speakers regulate their usage according to norms that are largely orthogonal to metaphysical naturalness" (2011: 10).

this is not the function, however, we can preserve the idea that our conceptual choices are not merely arbitrary, and often must be responsive to worldly constraints.

As our examples above show, many of the functions attributed to philosophically interesting concepts, including mathematical concepts, moral concepts, modal concepts, the concept of truth, and so on, plausibly serve functions very different from those of natural kind terms. Where functions vary, the criteria for evaluating, retaining, rejecting or rejigging extant concepts will vary accordingly—in Timothy Sundell's terms, there will be different 'measures of appropriateness' (2011: 15). It won't *always* be an apt criticism of a concept to say that it doesn't 'track the joints of reality' or serve in our best scientific theories, any more than it is always an apt criticism of a boat to say that it couldn't carry more than 1,000 tons of cargo. For social and institutional terms like 'married', 'citizen', 'person', or 'voluntary', or philosophically interesting and contested terms like terms for the moral, modal, or mathematical, the proper criteria for evaluation might not be whether the terms or concepts serve well in building explanatory and predictive theories, but whether they properly serve other purposes we have—say, endowing certain close human relationships with legal protections, enabling us to assign legal and moral rights and responsibilities, enabling us to coordinate our plans and attitudes, or enabling us to simplify our expressions of laws.

Even where the function of a concept is not predictive/explanatory, however, the pragmatic approach can nonetheless allow that our choices in engineering the concept are not merely arbitrary or subjective. For such conceptual choices also must be responsive to worldly factors. Consider as an example the concept of death, as examined by Bernard Gert, Charles Culver, and K. Danner Clouser (2006). They argue that the concept of death serves a variety of functions, including to enable us to determine when medical care should cease, funeral preparations should begin, survivors' benefits put into effect, and so on. Yet (they argue), there is no precise joint in nature marked by the concept of death, but rather a continuum of changes that go on in the process. Choosing, precisifying, or engineering the concept of death must be responsive to worldly matters. Certain empirical discoveries might place new pressure on our old vague concept of death. First, the old way of treating cessation of spontaneous breathing and circulation as a criterion for death comes into question with the availability of artificial ventilation, and puts pressure on finding new ways of identifying criteria for death. At the same time, the use of new and increasingly expensive medical technologies, and the critical need for organ transplantation to be done quickly, puts new pressure on determining the time of death more precisely than before, so that expensive treatments can be stopped, and organs harvested with greater chance of success for the recipient. These are empirical factors that put pressure on the old concept, and give reasons for at least precisifying the concept and altering the criteria typically used in applying it, so that it may continue to serve its functions. Where the function of a concept like 'death' involves, at least in part, enabling us to make decisions about when medical care should cease, and that medical care becomes increasingly costly or scarce, we may have a need to precisify the concept of death beyond the vaguer concept that served us well a hundred years ago (Gert et al. 2006: 284). The pragmatic approach to normative conceptual work is certainly worldly in that, to do it well, one must be responsive to worldly constraints and new empirical situations. In conceptual engineering no less than civil engineering, the question of which design (of concept or bridge) will best fulfill the relevant function, given the requirements, does not leave room for a merely 'arbitrary' or power-driven answer, and must be addressed while being sensitive to a variety of worldly factors.[25]

This is not the only way, however, in which our conceptual engineering work is subject to constraints that make our choices non-arbitrary. Civil engineering projects must take into account not only the function to be served, but also the constraints of the site: what the relevant land and geography are like, what the constraints are on surrounding extant structures and geographic features, etc. Similarly, when we engage in conceptual engineering, we must engage in descriptive conceptual work so that we can analyze, assess, and go on to be mindful of the multiple inferential connections our concepts bear to other concepts and practices.[26] Gert et al. again emphasize this point for the concept of death, appealing to the conceptual connections between death and a wide range of other social and personal (not merely medical) concepts as a way of criticizing the conceptual revisionism of certain medical doctors, who aimed to (re-)define death in such a way that they would be permitted to harvest organs sooner, when they would have a greater chance of success with transplantation. Such physicians, they argue, make the mistake of noticing only the connections between 'death' and other *medical* terms and practices, not the wider system of concepts and social practices in which 'death' plays a central role. Gert et al. use this example as part of a generalized argument for conservatism in conceptual change: while new circumstances (such as new medical technologies in keeping patients alive using artificial respiration, and in enabling organ transplantation) may require new precisifications of terms like 'death', Gert et al. argue:

> When a term plays an important part in social and legal practices, as 'death' does, then the greater the change in the meaning of the term, the greater the likelihood that there will be significant social and legal problems. (2006: 284)

Given the dangers of introducing confusion, distrust, and other social and legal problems in changing a common term, they defend a strong principle of conservativism regarding meaning change:

> It is almost impossible to describe a situation in which it is appropriate to redefine a term with widespread ordinary use in order to change any particular medical (or even social or legal) practice, in which that term plays a significant role. (2006: 285)

[25] This is not to say, however, that there will always be a uniquely best answer. While there may be some bridge designs that are far better than others, there nonetheless may be two or more that do the work (of transversing the chasm, safely supporting the intended vehicles, and staying within budget) equally well. So, similarly, in my view, we should allow the possibility that two or more different conceptual choices may (like different axiomatizations of geometry, or choices of different logical constants) serve equally well, without assuming there must be a 'worldly' fact to determine which of these 'carves at the metaphysical joints'.

[26] I suspect that this is related to the point Eklund makes as he argues that one cannot 'selectively' engineer the quantifier (or, presumably, other concepts) (2015: 380).

I think it is an under-appreciated point that conceptual engineering, no less than civil, does not take place in a vacuum, and that it is extremely important to note and be responsive to the inferential connections between the term in question (which we are considering revising or eliminating) and our other terms and broader practices.

Nonetheless, I think this is better taken as a caution than as an argument for a general principle of conservativeness in conceptual engineering. In civil engineering it may be a good—but defeasible—principle in constructing your new bridge or building to interfere as little as possible with surrounding roads and structures. But when problems get bad enough, or there are overriding social or moral purposes at stake, there are times for a more complete clearing. So similarly, though 'marriage' is connected to a wide range of social practices, those who value equality and happiness had good reason to change the legal definition to not precisify but rather expand the applicability of the term to same-sex partnerships, just as those who do not endorse racism had reasons to bulldoze the whole network of race concepts such as 'octaroon', 'quadroon', and 'mulatto' that played an influential social and legal role in former slaveholding and colonial societies. (As Burgess and Plunkett note, one question in conceptual ethics is "whether we ought to be using a given concept *at all*" (2013a: 1095)).

4.3. *Leaving Room for Critique*

This brings us to the third point: respecting the thought that concepts may be subject to critique. The functionally-driven pragmatic approach to conceptual ethics makes it clear why (and under what conditions) conceptual critique may be in place. One way critique can be appropriate, on this model, is in showing that the function of certain concepts cannot be fulfilled. If, for example, the term 'Vulcan' had the function of tracking a certain heavenly body, supposed to explain eccentricities in Mercury's orbit, then later discoveries led us to see that no term could fulfill this function—and it was time for getting this term out of our astronomical theories.

Another appropriate role for critique (closer to Foucault's work) is to show that the ostensible function of a range of terms or concepts comes apart from what it *really* serves to do and has done. For example, if race terms that were ostensibly introduced as natural kind terms—to explain and predict—have failed in that function, and have served instead to lend pseudo-scientific legitimacy to oppressive social practices, we have grounds to show that, whether our purposes were genuinely scientific or anti-racist (or both), they have gone wrong. Similarly if (as Foucault's work (2006) suggests) terms like 'madness' and 'mental illness' have served not so much to diagnose and treat as to give artificial pseudo-medical authority to practices of ostracism and exclusion, we have reason to reevaluate our attachment to and use of these concepts—provided we presuppose shared interests in transparency, human well-being, and/or inclusiveness.

In other cases, we may have reason to engage in critique once we notice that the function certain terms serve is only for a privileged few. If concepts like Hochdeutsch or 'received' English have served to reinforce and to legitimate regional and class biases, which we now seek to undermine, we will have reason to unmask and rethink these concepts.[27] The same goes for the case above of eliminating terms such as ‘octaroon’ and ‘quadroon’, and revising the concept of marriage to meet the goals of building a fairer society.

Uncovering the functions of various ranges of vocabulary can thus pave the way to the sorts of critiques engaged in by Nietzsche and Foucault. Exposing how our terms function and for whom, where those functions can’t be fulfilled or don’t fit our current shared values and goals, may give us entirely non-arbitrary reason to reject or revise the concepts at hand.

5. Conclusion

Where have we come? I have aimed above to sketch a blueprint for how normative conceptual work can be done, on a pragmatic model. No doubt it requires a great deal of expansion and revision. But already here we can note some important features of the proposal.

The first is that there is every prospect of adopting a method that does not require appeal to specifically *metaphysical facts* for guidance. On this model, all that is required is both descriptive and normative conceptual work, and also empirical work. To do normative conceptual work (on our extant concepts) explicitly, we engage in reverse engineering to figure out, empirically, what function(s) the concepts *have served* and *do* serve (where these, of course, might differ), and do descriptive conceptual work in figuring out how they work and what the ‘site constraints’ are: how they are related to other concepts and practices. But we must also do work in conceptual ethics to determine what functions our concepts *are* to serve, going forward, given our shared purposes. Finally, we combine that with empirical work, in doing constructive conceptual engineering: determining whether (given worldly constraints) certain modifications or precisifications would better enable the term to fulfill its function. This gives us a non-mysterious pragmatic approach to conceptual ethics that may not only be defensible against objections, but even be preferable to metaphysical approaches to conceptual ethics. For the latter leave us with familiar epistemic mysteries about how the relevant guiding metaphysical facts may be discovered. Fully evaluating these problems and comparatively evaluating the two approaches requires a more extensive discussion elsewhere.[28] Nonetheless, there is at least *prima facie* reason to think that, by appealing to nothing more than empirical, conceptual, and normative work, the pragmatic approach may

[27] And note that to engage in this kind of critique and make it seem non-arbitrary to do away with such concepts, we need not rely on any form of moral realism—the critique may simply involve pointing out what the terms purport to do (what is the legitimation for having such terms), versus what they actually do, and leave it to readers to decide what, from a practical standpoint, we should do given that uncovering. But nor do we, as readers who share a certain outlook, think that—given the relevant uncovering—it would be simply arbitrary to do away with the critiqued concepts or revise them heavily. *Given the moral views we may be presupposed to share*, such moves will not be arbitrary at all.

[28] See my (2017) for a more thorough evaluation of the metaphysical approach to conceptual choice, and Dasgupta (2018) for arguments against the idea that we can appeal to metaphysical naturalness to guide theory choice.

retain the epistemic high ground over metaphysical approaches to conceptual choice.[29]

Most importantly for present purposes, I hope to have shown that a pragmatic approach to conceptual ethics is a viable option for metaontological deflationists who still hope to make some sense of the difficulty, depth, and value of work that has often gone under the heading 'metaphysics'—as well as being a viable option for those who are (merely) deflationary *about a certain topic.* Like civil engineering, conceptual engineering is not a matter for discovery but for invention. But also like civil engineering, that does not mean that the choices we make are arbitrary, unconstrained, merely subject to our will, or 'subjective'. Which boat, or development of a concept, will work best *given our shared goals, purposes, and situation* may often be an objective matter, once all constraints are in. Of course, this is not to suggest that there will always be a *uniquely best* solution to a problem in civil or conceptual engineering. But that is no embarrassment for the deflationist, who may recognize the value in developing a plurality of concepts to serve a plurality of functions, as well as the possibility that two or more concepts could (like different bridge designs) serve a function equally well.

The crucial point here is that, once we understand the approach better, we can easily see that the problems thought to plague the pragmatic approach are avoidable. If we can properly develop and understand an approach to normative conceptual work in this way, then even metaontological deflationists will be able to account for the intuitions that have motivated many to take a metaphysical approach to conceptual choice. Given the purposes we commonly assume in the background, we can account for the central intuitions that the world is 'structured' and that there are worldly constraints on conceptual choice, allow that our conceptual choices are non-arbitrary and that some are 'objectively better' than others, and leave room open for the critique of problematic concepts. Most importantly, even if we adopt the deflationary metametaphysical approach, and the pragmatic approach to conceptual choice that comes with it, we will have room to account for the difficulty, depth, and importance of work in metaphysics—and to do so without invoking epistemological mystery.

[29] Some might be tempted to worry that any claim to the epistemic high ground is illusory, however. For determining what functions our concepts *should* serve (it might be argued) requires discovering *normative* facts, about what functions our concepts *ought to* have and whether we *ought to* revise them in various ways or reject them, and whether there are sufficient moral or political reasons to justify overriding the usual site constraints. But discovering such deep normative facts (it might be thought) involves epistemic barriers every bit as formidable as discovering metaphysical facts. This epistemic problem is avoidable, however, as long as there is some acceptable and non-mysterious approach to moral epistemology. Certain reductive naturalist approaches, for example, purport to render moral knowledge non-mysterious. Another approach, which goes naturally with the functional pluralism I have advocated above, is to adopt a form of non-descriptivism, for example, seeing our moral statements as expressions of certain kinds of non-cognitive attitudes or plans (as in the work of Blackburn or Gibbard)—thereby eliminating the principled epistemic problem, and leaving us with difficult, but pedestrian, problems of coordinating our plans and attitudes and figuring out what to do. We needn't settle these contested issues here, but only note that as long as some non-mysterious approach to moral epistemology is both tenable and combinable with the deflationary metametaphysical approach, the latter can indeed retain the epistemic high ground, giving it a substantial advantage over the heavyweight metaphysical approach.

Acknowledgments

I would like to thank David Plunkett and Ted Sider for helpful and insightful comments on an earlier version of this chapter. I would also like to thank audiences at the 2016 Chapel Hill Colloquium, at Leeds, at Stanford, and at Rice for helpful discussion of prior versions of this chapter.

References

Appiah, Anthony Kwame. 1992. *In my Father's House*. New York: Oxford University Press.

Ayer, Alfred Jules. 1946. *Language, Truth and Logic*. New York: Dover Publications.

Bell, Clive. 1914/977. Art as Significant Form: The Aesthetic Hypothesis. In George Dickie and R. J. Sclafani (eds.), *Aesthetics: A Critical Anthology* (pp. 36–48). New York: St. Martin's Press.

Blackburn, Simon. 1999. *Think: A Compelling Introduction to Philosophy*. Cambridge: Cambridge University Press.

Brandom, Robert. 2013. Global Anti-Representationalism? In Huw Price (ed.), *Expressivism, Pragmatism and Representationalism*. Cambridge: Cambridge University Press.

Brigandt, Ingo. 2010. The Epistemic Goal of a Concept: Accounting for the Rationality of Semantic Change and Variation. *Synthese* 177 (1):19–40.

Burgess, Alexis, and Plunkett, David. 2013a. Conceptual Ethics I. *Philosophy Compass* 8 (12):1091–101.

Burgess, Alexis, and Plunkett, David. 2013b. Conceptual Ethics II. *Philosophy Compass* 8 (12):1102–10.

Burgess, Alexis. 2013. Keeping 'True': A Case Study in Conceptual Ethics. *Inquiry*. http://dx.doi.org/10.1080/0020174X.2013.851866

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Dasgupta, Shamik. 2018. Realism and the Absence of Value. *Philosophical Review* 127 (3):279–322.

Davies, David. 2004. *Art as Performance*. Oxford: Blackwell.

Dutilh Novaes, Catarina. 2015. Conceptual Genealogy for Analytic Philosophy. In J. Bell, A. Cutrofello, and P. M. Livingston (eds.), *Beyond the Analytic-Continental Divide: Pluralist Philosophy in the Twenty-First Century* (pp. 75–108). London: Routledge.

Eklund, Matti. 2015. Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In Christopher Daly (ed.), *Palgrave Handbook of Philosophical Methods* (pp. 363–85). London: Palgrave Macmillan.

Foucault, Michel. 2006. *A History of Madness*. London: Routledge.

Gert, Bernard, Culver, Charles M., and Danner Clouser, K. 2006. Death. In *Bioethics: A Systematic Approach* (2nd edn) (pp. 283–308). Oxford: Oxford University Press.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them to Be? *Nous* 34 (1):31–55.

Haslanger, Sally. 2012. *Resisting Reality*. Oxford: Oxford University Press.

Horwich, Paul. 1999. *Truth* (2nd edn). Oxford: Oxford University Press.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Ludlow, Peter. 2006. The Myth of Human Language. *Croatian Journal of Philosophy* 6 (3):385–400.

Millikan, Ruth Garrett. 1984. *Language, Thought and Other Biological Categories*. Cambridge, Mass.: MIT Press.

Plunkett, David. 2015. Which Concepts Should We Use? Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 58 (7–8):828–74.

Plunkett, David. 2016. Conceptual History, Conceptual Ethics, and the Aims of Inquiry. *Ergo* 3 (2):27–64.

Plunkett, David, and Sundell, Tim. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosopher's Imprint* 13 (23):1–37.

Preston, Beth. 1998. Why is a Wing like a Spoon? A Pluralist Theory of Function. *The Journal of Philosophy* 95 (5):215–54.

Price, Huw. 2011. *Naturalism without Mirrors*. Oxford: Oxford University Press.

Price, Huw. 2013. *Expressivism, Pragmatism and Representationalism*. Cambridge: Cambridge University Press.

Richard, Mark. forthcoming. *Meanings as Species*.

Ryle, Gilbert. 1949. *The Concept of Mind*. Chicago: University of Chicago Press.

Sider, Theodore. 2011. *Writing the Book of the World*. Oxford: Oxford University Press.

Strawson, Peter. 1963. Carnap's Views on Constructed Systems versus Natural Languages in Analytic Philosophy. In Paul Arthur Schilpp (ed.), *The Philosophy of Rudolf Carnap*, Volume XI: *Library of Living Philosophers* (pp. 503–18). Peru, IL: Open Court Publishing.

Strawson, Peter. 1992. *Analysis and Metaphysics: An Introduction to Philosophy*. Oxford: Oxford University Press.

Sundell, Timothy. 2011. Disagreement, Error, and an Alternative to Reference Magnetism. *Australasian Journal of Philosophy*. DOI: 10.1080/00048402.2011.614266 (pp. 1–17).

Thomasson, Amie L. 1999. *Fiction and Metaphysics*. Cambridge: Cambridge University Press.

Thomasson, Amie L. 2007. Conceptual Analysis in Phenomenology and Ordinary Language Philosophy. In Michael Beaney (ed.), *The Analytic Turn: Analysis in Early Analytic Philosophy and Phenomenology*. London: Routledge.

Thomasson, Amie L. 2015. *Ontology Made Easy*. New York: Oxford University Press.

Thomasson, Amie L. 2016. What Can We Do, When We Do Metaphysics? In Giuseppina D'Oro and Soren Overgaard (eds.), *The Cambridge Companion to Philosophical Methodology*. Cambridge: Cambridge University Press.

Thomasson, Amie L. 2017a. Metaphysical Disputes and Metalinguistic Negotiation. *Analytic Philosophy* 58 (1):1–28.

Thomasson, Amie L. 2017b. Metaphysics and Conceptual Negotiation. *Philosophical Issues* 27:364–82.

Thomasson, Amie L. forthcoming. *Norms and Necessity*. New York: Oxford University Press.

Van Inwagen, Peter. 2016. The Neo-Carnapians. *Synthese*. DOI 10.1007/s11229-016-1110-4

Warren, Mark. 2015. Moral Inferentialism and the Frege-Geach Problem. *Philosophical Studies* 172 (11):2859–85.

Williams, Michael. 2010. Pragmatism, Minimalism, Expressivism. *International Journal of Philosophical Studies* 18 (3):317–30.

Yablo, Stephen. 2005. The Myth of the Seven. In Mark Eli Kalderon (ed.), *Fictionalism in Metaphysics*. Oxford: Oxford University Press.

Index