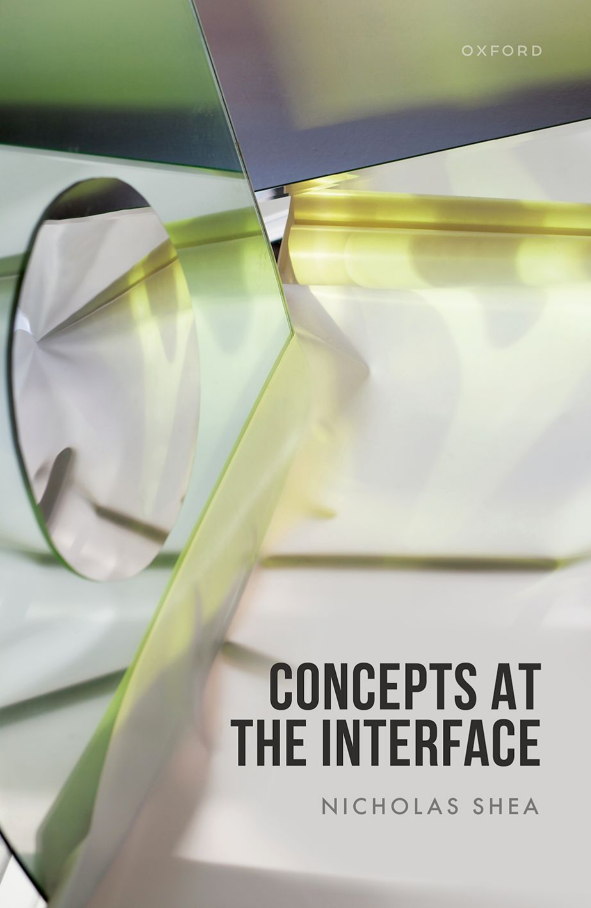


OXFORD

CONCEPTS AT THE INTERFACE

NICHOLAS SHEA

Concepts at the Interface

Concepts at the Interface

NICHOLAS SHEA

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

© Nicholas Shea 2024

The moral rights of the author have been asserted

This is an open access publication, available online and distributed under the
terms of a Creative Commons Attribution-Non Commercial-No Derivatives 4.0
International licence (CC BY-NC-ND 4.0), a copy of which is available at
https://creativecommons.org/licenses/by-nc-nd/4.0/.
Subject to this license, all rights are reserved.

Enquiries concerning reproduction outside the scope of this licence should be sent
to the Rights Department, Oxford University Press, at the address above.

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: 2024936379

ISBN 9780198893660

DOI: 10.1093/9780191997167.001.0001

Printed and bound in the UK by
Clays Ltd, Elcograf S.p.A.

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

To Clara and Adam

Contents

Preface

So—you've turned the first page, wanting to find out what this book is all about. Poised to help, arrayed in your mind you have an immense resource, a huge armoury of concepts. You'll need to pull out the right ones, tweaking here and there, perhaps adding some new ones. Then if you arrange them in the right way, you'll be in a position to understand whatever I throw at you. You are all set to make use of the everyday wonder that is human conceptual thought.

Let's just play with our concepts a little to start with. Imagine *a skiing crab wearing a graduation gown*. Yes, do. Not something you have experienced before I guess. But still, images begin to form. Snow, skis, and legs arrange themselves into some sort of order. You can work out what he is doing with his claws. (Were you, with me, assuming it was a *he*?) Where is all this material coming from? It was there somewhere in your mind, in your concepts of crabs and of skiing. Each concept contributes images, assumptions, and beliefs. Your imagination fills out the picture. Is the crab moving? If so, no doubt he's going downhill. Now look at his gown. Is it flapping in the wind? And we're off, having fun with concepts. The process is both so familiar as to be almost banal and so marvellous as to be almost miraculous.

Concepts are the tools with which we think. Concepts are how we classify things in the world and organise our knowledge about them. Concepts make us take things for granted, quickly parsing our experience into categories and rapidly drawing all kinds of interesting conclusions. But concepts also lead us astray, driving lazy overgeneralisations and housing unpleasant prejudices. Much is implicit, shaping how we see the world in ways we don't even realise—just as you don't see how hazy the air is in the valley until you step into the crystalline clearness of a high mountain peak.

Nevertheless, we all do realise that there is a lot of variation in how people conceptualise the world. It is endlessly fascinating to explore with people from a different culture how they see—that is, conceptualise—the world differently. A straightforward distinction between cutting and tearing seems obvious, for example, but is that distinction universal? More locally, as we teach, we are constantly trying to second guess the student's mindset, what they have misconstrued and what they've understood. The five-year-old hasn't grasped that, however many times you divide a number, you'll never get to zero. The first-year student doesn't understand that ∞ doesn't work like a number. Even everyday conversation calls for some vigilance about whether the other person conceives of things differently.

I have been interested in concepts ever since I started studying philosophy. Beyond the fascination of how we all see the world in our own peculiar way, I found myself drawn—in a move which I suppose diagnoses a philosophical bent—to the deeper question of how we can be thinkers at all. What sort of mechanism enables people to divide up the world into categories and reason about it? How can goings-on in the brain be the basis of this rich and useful mental life?

It was a good time to ask. Although these questions have troubled philosophers for thousands of years, striking progress was being made on them in the last decades of the last century and the first decades of this one. A new idea opened up space for a new answer. That answer is the Representational Theory of Mind (RTM), the theory that thinking consists in a causal process, the interaction of physical particulars, representations, which have meaning or content. Where categorisation and inference appear to us as an unfolding series of thoughts with meaning, RTM promises to explain how this works in terms of a sequence of physical representations interacting in a causal process that is ultimately realized, in some way, in the brain. This idea, inspired by mechanical and then digital computers from the 1940s onwards, initiated the 'cognitive revolution' in psychology, the implications of which are still gradually being worked out.

There are actually two foundational questions here. First, how can goings-on in the brain be the basis of a rich mental life? Second, how can there be meanings at all—how is there space for meaning or representational content in the natural world? The first question looks more psychological, the second philosophical, but in fact both disciplines have contributed to answering both questions. Not only has philosophy liberally made use of results from psychology. Psychology has readily taken inspiration from philosophers like Wittgenstein and Frege (and admonitions from Fodor). As I happily remember from the (psychologist-led) London Concepts Group as a student, this is a field where psychologists are keen to talk to philosophers—as well as our needing to pump them for their results. Concepts research is a parade case for the new enterprise that is cognitive science: a self-conscious alliance between psychology, linguistics, neuroscience, anthropology, computer science, and philosophy, a collaborative endeavour aimed at making progress on shared questions about the mind.

Philosophers initially tried to base an answer to the question of where meaning comes from—a theory of conceptual content—on a theory of categorisation, on how thinkers apply concepts to things in the world. Theories based on definitions were displaced, for philosophical as well as psychological reasons, giving way to prototypes and exemplars. Some objects are typical members of a category—when an object has most of the prototypical properties or matches some stored exemplar. Others are more marginal. Compare a robin with a penguin as an example of a bird. Then it was discovered that sharing an underlying essence was important in many cases, and that other sorts of theoretical knowledge is involved in categorisation and inference.

The empirical data is overwhelmingly drawn from WEIRD populations (Henrich et al. 2010), so questions remain about how well the basic mechanisms generalise. Cross-cultural studies are only just beginning to fill in the picture. However, the empirical story was already getting much richer, with striking data on the role of sensorimotor simulation, many insights from subtle studies with infants and children, better appreciation of how the body and world support cognition, and increasing knowledge of how conceptual abilities relate to the brain and its pathologies. Philosophers realised that a theory of content would not just drop out of an account of how concepts are applied to the world. Plus, they had been working with a too-simple theory of the mechanism. But what to replace it with? The richness of the data—psychological, computational, neurological—makes it hard to see what philosophical theorising should be based on.

Actually, I think a framework has emerged from the science that offers some unity, at a relatively abstract level, and integrates a wide range of empirical results. That is what I aim to present in this book: a framework that incorporates the wider range of phenomena which concepts are now known to be involved with, formulated at a level of generality which is suited to our philosophical questions. This involves treating the issues rather more abstractly than a psychologist would—there is a lot in the book about representations, structures, and types of inference—but in a way that is, I hope, recognisably continuous with the science.

Perhaps this framework will eventually prove to be a useful basis for formulating a theory of content. For myself, it seems premature to launch into the metaphysics of content without getting clearer about how conceptual thought actually works. The connection of content to consciousness also makes the issues a lot harder with concepts than they are in the subpersonal case (Shea 2018). Whether or not it helps with theorising about content, though, the framework is valuable in its own right, as an answer to our first question. It shows us, in outline, how goings-on in the brain, and its interactions with the world, give rise to our rich mental life. It demonstrates that the promise of RTM can be vindicated.

Another aim is to make the person more central. Thinking is something we *do* with concepts. To highlight this, I focus on the use of concepts offline: when, rather than reacting to a stimulus, the agent is thinking about what to do in the future or drawing fresh conclusions about what is the case. Psychological research has understandably focused on how people react to a stimulus, since that is what participants do in an experiment. Philosophy, though, has always had offline thinking in its sights. (We are contemplative folk.) But our paradigm has been reasoning, something like the step-by-step theorem proving of logic. Empirical research shows that we need to cast the net much wider and encompass all the other things that go into our thinking, like conjuring up mental images, running simulations, and relying on a broader range of processes, sensory and motoric, affective and evaluative. A concept provides access to a diverse store of information about a category, only a small part of which is activated and used on each occasion, as the context demands. So we need to make a sharp distinction between the store of interconnected information about a category (which is what some theorists mean by 'concept') and the mental representation that is manipulated in occurrent thinking. Making the offline case central also serves to highlight other features that situate conceptual thinking at the level of the whole person. It is a sphere for mental agency, often effortful, where the thinker appreciates what is going on, and where what they do can reflect their goals and values.

The picture I present is informed both by classic results in the cognitive psychology of concepts and by newer work in neuropsychology and cognitive neuroscience. In addition, advances in computational modelling offer insightful ways of modelling the computations going on in the brain. Some recognisably philosophical questions are centre stage: about the types of representations with which we think, how they are used to build informational models of the world, and the different types of computations they enter into. This throws up a long-standing puzzle about how cognition retrieves only a selection of relevant memories on which to perform inferences—the notorious frame problem. The framework developed in the book shows that, in many cases, the mind has resources for circumventing the frame problem, and for tackling it when it has to. I argue that this hybrid solution shows us why human cognition has proved to be such a powerful resource. The book is concerned with grappling with the details, with making a philosophical contribution to a collaborative enterprise. However, the picture that emerges demonstrates, I think, that cognitive science is realising its promise of delivering a tangible understanding of where our rich mental life comes from.

1
Thinking with Concepts

1.1 Concepts in the Playground of Thought

This book is about thinking, in the everyday sense—what we might call conscious deliberation. How does it work? Why is human cognition so powerful? We have a tremendous capacity for thinking things through. We think about possible strategies and courses of action as we decide how to act. We use the same tools to derive deep conclusions from what we already know. When we learn something new, we can think through the implications, enlarging on what we have learnt. That is, we have an enormous capacity for practical and theoretical inference.

When faced with an immediate choice, although we will often simply respond out of impulse or habit without forethought, we do sometimes take a few moments to think about how to deal with the situation. More typically we find ourselves thinking things through when the world is not pressing: we sit quietly thinking, walk along thinking, or screen out the chatter and just think—in any situation where the ongoing task does not occupy all of our cognitive capacity.

Humans do a lot of deliberation. Other species less so, or not at all. We think things through not only to work out the possibilities for action and their consequences, but also to discover new facts—to infer new conclusions from what we already know. Complex forms of forward planning and elaborate processes of theoretical inference are undoubtedly unique to humans. I want to get a clear picture of the cognitive resources that go into our tremendous capacity for thinking and reasoning about the world in this way.

In philosophy, practical and theoretical inference is standardly thought of as a step-by-step mental process, each step moving from one or two premises to a conclusion, which then forms the basis of the next step. Logical reasoning is the paradigm, although the steps involved need not be deductive. I will call it just 'reasoning'. The steps take place between thoughts that are structured like the sentences of natural language. (They may or may not be subvocalised uses of the language faculty.)

Important as it is, reasoning captures only a part of what we do when we are working out what is the case and how to act. We imagine possibilities and simulate unfolding scenarios using much of the rest of the mind's rich palette of resources. Actions are simulated in sensorimotor systems, relying on their highly-tuned grasp of the sensorimotor contingencies that characterise our physical world. Routes are planned in reliance on the cognitive map in the medial temporal lobe. The values of actions and outcomes are estimated by reinforcement learning systems. Social situations are simulated through the lens of systems that grasp others' mental states and their likely emotional responses. Our overall assessment of a situation is bathed in our evaluative responses to the rich picture we build up using all these special-purpose systems. Intelligent thinking mines all these resources in planning how to act and working out the implications of what we know.

At the same time, our capacity for reasoning is clearly crucial. Reasoning with explicit conceptually-structured thoughts can take us to consequences that go far beyond the expertise of any of our special-purpose systems. We can work out logical consequences that are powerful precisely because of their generality. Reasoning can transcend the assumptions by which special-purpose systems operate. Assumptions, though useful, are also limitations. Reasoning can move beyond them. Conceptual thought can also recombine the information it draws on in new ways.

How does human thinking achieve all of that? At the heart of the story are concepts. A concept is a key that unlocks a rich array of stored information. At the same time, a concept is a component that can figure in reasoning. Concepts allow our thought processes to mine the rich stores of information laid down elsewhere in the mind. Combining a collection of concepts in thought allows the mind to integrate a diverse array of information, much from special-purpose systems, into a coherent thought. Combining existing concepts in new ways allows us to think of, and then simulate, striking new possibilities. The models we build of these scenarios in turn generate new ideas of their own.

* * *

Human cognition underpins many of our species' most remarkable outputs. It has given us the power to construct the complex social groups that we *Homo sapiens* have always depended on for our survival. More recently, it has allowed us to imagine the elaborate social institutions that now regulate our lives: political, commercial, and legal systems that make possible our highly specialised form of obligate cooperation and deep interdependence. Human cognition has created the stories and explanations of organised religions, and then of modern science.

It drives the rich imagining of the arts. Doubtless there are many other factors behind the effervescence of human creation—behind the complexity of the things we produce and the scale of our impact on the world. Our peculiar form of culturally-based inheritance is surely crucial. High-powered vision, manual dexterity, and linguistic communication are probably all part of the story. But so too is the power of human cognition. Conceptual thinking is the motor that drives many of the ways that humans manage to understand the world and plan our complex activities in it.

In piecing together how all this works, the key parts of the jigsaw are concepts. Concepts are elements of the thoughts we have when deliberating. When I think, *what sport shall I play this week?*, I am making use of my concept SPORT, putting it together with other concepts to form the thought. Concepts then connect our thinking to all of the information, of different kinds, that we rely on in thought. The concept SPORT gives me access to semantic memories (e.g. that ultimate frisbee is an easy sport to get into) and to information in many special-purpose systems: what is physically involved in various activities and how I am likely to feel doing them. In the input direction, concepts are applied to the world. This is the process of categorisation. I see a bunch of people running around in a field and categorise the situation under the concept SPORT. In the other direction, thinking with my SPORT concept allows me to recall and entertain perceptual representations of sport, motor representations of doing sport, and emotional and evaluative representations of what it is like to watch or take part.

So the picture I paint will have concepts at the centre. How do they interface between different systems and what kinds of representations and processes are involved in each? The full story needs to tell us how representations are typed and combined, how those representations are processed or computed with, and what kinds of representational models or structures are used for representing and storing information. Many of the elements I rely on are familiar. I have my own standpoint on how some of the elements should be characterised, and also on how they need to be put together. The result is a somewhat distinctive picture. The approach is also distinctive in drawing together points about representational structure, computational process and informational models, attempting to offer a single coherent account of how they work together in concept-involving cognitive processes. I will call this 'concept-driven thinking', or sometimes just 'conceptual thinking' for convenience, but note that these terms are intended to include the special-purpose representations that are involved, which need not be constructed out of concepts. (Note also that conceptual thinking does not cover all uses of concepts—concepts may play roles outside of conscious deliberation, e.g. storage in long-term memory or activation in non-conscious processing: §1.4.)

Building up a cohesive picture of conceptual thought involves bringing together several different literatures. Most prior work on concepts targets only thin slices of our conceptual life. Psychological work has tended to focus on categorisation of perceived stimuli. That is understandable, given the need for tractable experiments, but much of our conceptual thought happens offline. Philosophy has long been interested in the role of concepts, offline, in logical reasoning, but it has tended to neglect the large part of our conceptual life that is concerned with imagination, simulation, or building a model of a real or possible situation. My aim is to offer a more comprehensive account of conceptual thought, combining the different elements into a coherent, unified picture.

* * *

I also want to highlight a feature of our thought processes that is familiar and obvious, but somewhat overlooked in most treatments. Deliberate thought is a way that a range of information is collected and considered together. For example, I may find myself conflicted between helping my son with a problem and leaving him to it, to develop his self-reliance. I weigh these competing concerns in thought. The forecast is dry but the sky looks like rain. I weigh the competing evidence in thought. Thought is a place where we bring together a variety of information and consider it together. We can bring together a series of facts or semantic memories. We can also bring together a collection of sensorimotor expectations and evaluative responses, allowing us to see how each bears on the others. At a party I enter a room and consider joining the nearest group of people. But then I see that the encounter could be emotionally tense, so I select a physically more awkward route through the room. Thinking is a process where we bring these disparate pieces of information together, enabling us to appreciate the relevance of each piece to each piece. That is, thinking takes place in some kind of shared representational space.

A good metaphor for this a workspace. A workspace has the capacity to maintain a set of representations together so that they can be manipulated in respect of one another. This familiar phenomenon is often presupposed in accounts of thinking, but it seldom takes centre stage in theories of conceptual thought. It has been given a strong role elsewhere, as a theory of consciousness—the global workspace theory. The metaphor is apt for our purposes, so I want to appropriate it here, shorn of any commitment to acting as a theory of consciousness. Deliberative thought does indeed typically take place in a workspace, but for all I say here there could be conscious representations outside the workspace, or indeed workspaces that are not conscious. I'm interested in the cognitive phenomenon: a way of entertaining representations such that they are in touch with one another, so that they can be manipulated in relation to one another, and so that the consequences of one for another are typically apparent.

The other field in which the idea of a workspace shows up is in research on working memory. Working memory is a capacity for holding representations online and manipulating them. Experiments on working memory have focused more on the former than the latter—characterising storage and maintenance, more than the capacity for manipulation (Cowan 2008; Hasson, Chen, and Honey 2015, p. 304). The latter is there, though, often obliquely, and is sometimes brought to the fore (Baddeley 2012; Reuter-Lorenz and Iordan 2021, pp. 285–6; Masse et al. 2019). Claire Sergent has recently introduced the idea of a 'global playground' (Sergent et al. 2021). This was also in the context of work on consciousness, but the way she introduces the global playground characterises it as a cognitive phenomenon, more akin to working memory. Representations in the global playground are maintained and shared, offering 'wider cognitive possibilities than automatic unconscious processing', but without being tied to the performance of any specific task (Sergent et al. 2021, p. 12). This is a useful construct for my purposes, both because it allows the contents of the playground to be richer than in typical models of the global workspace, which are usually committed to global 'broadcasting' occurring one stimulus at a time; and because it emphasises that there may be a currently active representation of a scenario (which for me can be suppositional as well as actual) that is not there to guide performance of a particular current task (Cohen et al. 2023). Another merit of 'playground' is that, as a new term, it is less redolent of theories of consciousness.

The cognitive playground is a system in which representations are maintained and manipulated together, such that connections between representations, like relations of support and contradiction, are inherent in the way the information is entertained. A simple example occurs with binocular rivalry, where conflicting information is presented to each eye. The inconsistent information is processed in parallel in different areas of the brain, but other areas are dedicated to representing a coherent scene, which is what the subject consciously experiences (Haynes and Rees 2005). My own view is that this playground characteristic is extremely important. I will use 'thought' liberally, for representations of all kinds in the cognitive playground, whether composed out of concepts or not; and 'thinking' for the unfolding of sequences of thoughts and for the cognitive processes that apply to them (e.g. directing attention). Much of the utility of thinking depends on its taking place in an integrated playground. This makes it pressing to give an account of how the various representational structures, informational models, and computational processes involved in cognition can interface with one another so as to work together in concert. Much of the book will be dedicated to characterising those elements and their interactions. Whilst doing that, I want to ensure that the fact that thinking takes place in some kind of shared representational space, which is so often in the theoretical background, remains fully in focus. There is an arena in which all this thinking takes place. Hence I will freely talk about the cognitive 'playground', with its connection to the cognitive idea of a workspace, even while wanting to disavow its resonance with theories of consciousness. I want to lay claim to the term. It neatly encapsulates a critically important property of concept-driven thought.

1.2 Deliberation

Here is another way to say what the book is about. I'm interested in the role of concepts in executive processes: in processes of 'controlled semantic cognition' (Jackson, Rogers, and Lambon Ralph 2021), under 'endogenous control' (Calzavarini 2022). 'Executive' is sometimes used to smuggle in a homunculus, an unexplained cognitive faculty that is 'me' and does all the crucial thinking and deciding. More often it is unobjectionable: a useful umbrella for a variety of higher level cognitive processes, processes of which we can give a (non-homuncular) computational/functional account. I want to give centre stage to the kind of thinking that depends on working memory, and on which executive processes operate. It has the characteristics associated with 'type 2' cognition (or, misleadingly, 'system 2'), in particular being subject to interference from concurrent cognitive load. (Empirical support for this characteristic does not imply that we need to endorse a type 1/type 2 dichotomy.) When cognitive-load-dependent thinking takes place in a series of steps, that constitutes *deliberation*—although note that 'deliberation' need not be deliberate in the sense of being the result of intending to deliberate or choosing what to think about.

Executive processes operate on representations in working memory. There is much that is controversial about working memory, including whether there are several different working memory systems or just one. Either way, it is clear that there is a tight capacity limit on the amount of information, of whatever kind, that can be actively maintained and manipulated at once.[1] This is the episodic buffer (Cowan, Morey, and Naveh-Benjamin 2021) or focus of attention (Baddeley, Hitch, and Allen 2021). (Separately, verbal rehearsal can take place relatively autonomously in a phonological loop.) A larger collection of information, drawn from long-term memory and/or perception, is activated at any one time, forming a model of the environment (Cowan et al. 2021, p. 49; Reuter-Lorenz and Iordan 2021, p. 285). The focus of attention applies to small parts of this model at a time (Cowan et al. 2021, p. 47). Similarly, it may be that the contents of conscious experience are wider than the narrow set of contents currently at the focus of attention in working memory (Block 2011). Working memory is, however, crucial for maintaining information in consciousness in the absence of input, beyond a very short time window (Baddeley et al. 2021).

The contents of working memory are determined by how attention is directed. Attention is needed to hold items in active working memory, and for dealing with external stimuli, inputting them into working memory or blocking them depending on the thinker's goals (Baddeley et al. 2021). Attention is captured when a

[1] The capacity limit may be, not a resource constraint, but a functional feature that allows the shared use of the same representations: a bound to interference between different tasks that engage the same representations (Musslick and Cohen 2021).

sudden change in an environmental feature is processed. These 'bottom-up' effects compete with the attention directed intentionally, 'top-down', in pursuit of the agent's current goals (Cowan et al. 2021, p. 74). The latter means that the content and manipulation of working memory is partly under executive control. During development, the child's capacity to direct attention improves: to use attention to enter new items into working memory and screen out distraction, and to actively remove from working memory items that are no longer task-relevant. This capacity also varies between individuals. It seems that it is variation in this capacity that accounts for the strong correlation which exists between performance on working memory tasks and standardised tests for fluid intelligence (G), and with educational attainment (Cowan et al. 2021). Thus, a very simple measure of whether a thinker can maintain an internal representation in the presence of distraction can be a strong predictor of these seemingly much more complex traits (Stanovich 2009). The task probing the domain-general element of working memory manipulation can be as simple as measuring whether a participant can resist the urge to saccade towards a flickering stimulus (Engle 2010). It is striking that such a simple test can measure a cognitive capacity that has such wide-ranging effects on life outcomes.

So, on one common understanding of the term, intelligent thought is underpinned by the operation of working memory. The content of the thinker's current active model—a collection of information activated in working memory—depends on the agent's limited-scope capacity for directing attention. Working memory research gives us a functional account of how conscious thinking provides a capacity for manipulating a variety of informational resources, running simulations in existing models of the environment and combining previously learned models in new ways (Langdon et al. 2022).

This is related to the distinction between what are often called 'system 1' and 'system 2' (Evans and Stanovich 2013). Rather than two separate cognitive systems, these are better thought of as two characteristic ways in which different psychological systems can operate, type 1 and type 2. Type 1 processes are fast and automatic, and can operate in parallel with one another with minimal interference. Type 2 processes are slower, operating step-by-step, and tend to interfere with one another. They typically induce a feeling of effort. Interference occurs because type 2 processes depend on working memory. They interfere with one another because the capacity for manipulating items in working memory is limited. We need not think of this as a dichotomy. A cognitive process may depend more or less on the capacity to direct attention to supervise the contents of working memory. The extent of inference or autonomy may lie along a continuum.

A central case of load-dependent thinking is where we think through a problem step-by-step: what I have been calling *deliberation*. Deliberation can be abstract and rule-based; and it can account for how an agent conforms to an explicit norm (Evans and Stanovich 2013). We should not take this to be exclusive: we should not assume, as some dual systems theorists do, that following a norm or making a logically valid inference are things that can only occur in the type 2 way. Automatic processes can also account for how agents comply with normative requirements, and some logical transitions occur automatically (De Neys 2023); conversely, type 2 thinking can be erroneous and irrational. But step-by-step deliberation—load-dependent and effortful—is a key place where we find concepts at work. It serves to highlight the way concept-driven cognition depends on a capacity-limited ability to manipulate representations in working memory. That phenomenon is actually wider. Capacity-limited conceptual thinking doesn't just happen offline. We engage in it online, when thinking about how to respond to a situation, for example, or when trying to understand what someone has said. I emphasise the offline case here in order to highlight the phenomenon.

The book will lay out a picture in which concepts afford access to many kinds of information, conceptual and nonconceptual, so that a thought composed out of concepts in working memory can drive the construction of a multimodal and amodal, cross-domain model of an actual or hypothetical situation, a 'suppositional scenario'. This is constructed in working memory, in the cognitive playground. A series of steps between such overall models, supported by conceptually-structured representations in working memory, will then be a load-sensitive, type 2 process. These are the characteristics of conscious conceptual thinking.

A final cognitive property to mention is the distinction between model-based and model-free learning and decision-making. The distinction has its home in research on reinforcement learning. A model-free system learns, from a history of reinforcement, which actions produce the most long-run reward in which contexts. The system is just learning the value of actions, without appreciating which outcomes in the environment produce those rewards. Thus, when the value of an outcome changes, for example when fed to satiety on a particular foodstuff, the model-free system has to laboriously re-learn new values-in-context for all the actions it could perform. 'Model-based' tends to be something of a catch-all for any system that is not so-limited, but it also has a more positive sense.

In the more positive sense, a model-based system has any or all of the following four properties: (i) it represents something about the causal structure of the environment, (ii) it calculates over those representations in working memory, (iii) it makes choices that are stimulus-independent (not just driven by the current stimulus), and (iv) it produces an immediate change in behaviour when the value of an outcome changes (devaluation sensitivity). The first property comes in degrees and is not particularly connected with executive functions. Special-purpose systems like the cognitive spatial map represent some of the structure of the environment and use it to calculate what to do. (ii) is more sophisticated, since it calls on working memory. This is found in non-human animals that use perceptual- and motor-systems offline, in prospection, to plan a course of action and make a choice (Tomasello 2022, pp. 48–52). (iii) Reliance on working memory allows for choices that are stimulus-independent—to a degree that reflects the extent of the agent's control over the direction of attention. If an agent is planning its actions in this way then, (iv), a change in the value of an outcome can be input in the offline calculation, resulting in an immediate change in what the agent will decide to do. In short, although the collection of systems that can be said to make use of a model, in some sense or other, may be heterogeneous, the richer idea of model-based cognition connotes a collection of features which together characterise intelligent deliberation.

Notice, too, that decisions taken in this way can reflect the current goals and values of the agent. Special-purpose systems operating in automatic ways need not (e.g. stimulus-response tendencies acquired by model-free reinforcement learning). They may presuppose values built in through experience, or canalized in development through gene-based evolution, values that are contextually inappropriate, biased, prejudicial, or otherwise at odds with the reflective values of the agent. Decisions taken by conscious deliberate thinking have access to the agent's goals and to the values which they attach to outcomes entertained in the global playground. Deliberation can help to produce coherence in an agent that has a disjointed set of goals and motivations across different systems (Shepherd 2023).

1.3 Non-Local Inference

In widening our conception of concept-involving thought processes, I need to leave space for inferences that work in a quite different way from reasoning. I will stipulate that *reasoning* involves conceptual representations, moving from one or two premises to a conclusion, the conclusion then forming the basis for further steps, sometimes introducing new premises along the way. *Inference* is a more inclusive category, covering every kind of content-assessable transition between representations. Reasoning is not only found in the logic class. It also captures one way that humans perform everyday inferences. But there is another, rather different, way that people reach conclusions and take decisions. We have processes that can take account of a wide range of considerations at once, weighing or integrating them in reaching a conclusion.

We see examples of that when people make choices for which many different factors are relevant (Usher et al. 2011; Newell and Shanks 2014); also in work on high dimensional category learning (Sloutsky 2010; Ashby and Valentin 2017). In both cases, the output of the process depends on weighing, seemingly in parallel, an array of different factors at once. One case of this that philosophers have been interested in is abduction (or inference to the best explanation). In deciding which theory or hypothesis is most plausible, we can take into account a wide range of evidence and theoretical arguments, weighing each piece of information against the others in order to reach an overall assessment. Or so it seems.

Abductive inference was famously declared by Jerry Fodor to present a deep challenge to classical computationalist versions of the representationalist theory of mind, theories that are committed to the causal role of a representation in cognition being based entirely on its constituent structure. Fodor agrees that inference to the best explanation does indeed call for the judicious weighing of a wide range of evidence, but argues that we don't know how step-by-step reasoning that takes just a few premises at once as input could achieve this kind of overall assessment in a realistic timescale (Fodor 2000). He adds that assessments of confirmation, simplicity, and centrality also have this character. Fodor calls these processes 'global', but the problem arises even if they do not in practice involve the thinker taking into account everything they know and believe. The problem is that these inferences involve the inclusive reliance on a large amount of information at once, seemingly in parallel. In this sense they seem to be non-local, even if not fully global. We do not have realistic ways of modelling these processes as being performed by step-by-step reasoning, one or two premises at a time; so we doubt that is how it is done. We need to enlarge our conception of the kind of computations that could be involved.

In fact we do have plausible computational models of how non-local inferences occur, and we now have good behavioural and neural evidence for their psychological reality in several domains. As we shall see, the cognitive spatial map in the hippocampus may be one case (§2.2). Some computations plausibly take place across the whole coherent map, rather than locally, place cell by place cell (Samsonovich and Ascoli 2005; Khajeh-Alijani, Urbanczik, and Senn 2015). Conclusions are reached in reliance on the structural representation as a whole. The computations are not well-modelled as step-by-step moves between a small number of premises. Nevertheless, the commitments of the representational theory of mind (RTM) are respected: the way chains of representations unfold depends only on syntactic/vehicular properties of the representations. It is just that relations between a whole array of individual representations (place cell activations) are critical to the way computations are performed over the map. An analogy would be looking at a cartographic map and seeing various features that pop out at a glance. The mountaineer performs a holistic assessment over the map, but what they conclude is driven only by the individual marks on the page and their relations. The former makes the process non-local. So Fodor is wrong to suggest that we have no idea how computations could be non-local, consistent with RTM—but that does not yet show how humans actually perform inference to the best explanation and other forms of non-local inference.

In the cognitive map the non-local process takes place outside consciousness and may not involve concepts, but this kind of non-local assessment probably does also take place in the course of conscious deliberation. A range of different facts and considerations are processed in working memory, entered into the cognitive playground, and the thinker makes an overall assessment based on them. The capacity of the cognitive playground is of course limited, so this is by no means a global assessment. If we perform abduction over a collection of evidence held in working memory, then this is far short of being an all-things-considered inference to the best explanation. It does, however, depart from a standard way of thinking of representational transitions, in terms of step-by-step reasoning. We can find many everyday examples when we think about imagination or simulation. When I am imagining walking through the crowded room at that party, my assessment of where to go and what to do is taking on board, at the same time, the physical constraints of the room and the things in it, my physical capacity for moving through it, the social relations between people in the room and between them and me, and the effects on them of various things I might say or do. Individual considerations may grab my attention and be highlighted one at a time, but what I end up doing gets shaped by many different factors all at once.

Admittedly, I am not offering a very precise characterisation of the phenomenon. For now, I want to get by with an impressionistic feel, a sense, offered by a few examples, that there is something here that differs from step-by-step reasoning. The examples suggest a style of inference that contrasts with classical computation. Familiar examples take place within a cognitive playground, but that is neither necessary, as the cognitive map case suggests, nor sufficient, in the sense that step-by-step logical reasoning can also be performed on representations in the playground. Some of the examples involve inferences over a complex structural representation (cognitive map; sensorimotor simulation). Other examples seem to involve the parallel processing of a large amount of information (weighing factors in making a choice; multi-dimensional categorisation; confirmation). The latter are often well captured by computational models that perform parallel distributed processing in artificial neural networks, as we shall see.

In all of these cases the inference is performed on a representation that is more complex than one or two premises. It is some kind of larger model of the situation: a cognitive map of spatial layout; a sensorimotor simulation of potential ways of moving one's body and limbs; a collection of interconnected, potentially relevant evidence; an array of competing considerations and evaluations weighing against one another in making a choice. 'Model' here has to be understood loosely, if it is to cover the phenomenon I am pointing to, but it does suggest the idea that conclusions are being drawn from some sort of coherent representation that cognition has built.

'Model' also has a useful resonance with the contrast between models and theories in philosophy of science. Inferences from theories tend to take place hypothesis-by-hypothesis. A model, by contrast, can be manipulated as a whole. Think of mechanical models of the solar system. But 'model' in philosophy of science does not have quite the same meaning since it encompasses models as simple as a single mathematical equation. Most mathematical models are used in inferences in the step-by-step way rather than the non-local way. ‘Model’ also has resonances with the contrast between model theory and proof theory in formal logic. The connection here is more tenuous since model-theoretic inferences also take place step-by-step. But the intuition behind the label may still be apt, because model-theoretic consequences are things that follow from the model as a whole. They do involve an overall assessment (e.g. truth on every assignment of values to terms), unlike the application of proof rules, even if in practice that overall assessment is often performed exhaustively, piecemeal and step-by-step.

Informational model is a rather inclusive category. Special-purpose systems also involve informational models (the topic of Chapter 4). These can operate outside of the cognitive playground. Informational models are also stored in long-term memory. For now, I am focusing on informational models assembled, under the influence of conceptual thought, in the cognitive playground. When a model is constructed in the playground, that provides us with a way of spotting inconsistencies—inconsistencies between different beliefs (more generally, between descriptive representations), between beliefs and values, or between different values—and of doing something to restore consistency, often by changing what we believe or represent (although not always in a rational way; Harmon-Jones and Mills 2019). This predicts that people should be more likely to achieve consistency between beliefs that they have considered together. That offers little protection against beliefs stored in long-term memory being radically inconsistent with one another. That is, it predicts a degree of belief fragmentation—an implication which has some empirical support (Lewandowsky and Kirsner 2000; Bendaña and Mandelbaum 2021).

1.4 Concepts as an Interface

Concepts, as I use the term here, are a type of mental representation. They are freely-recombinable elements of the thoughts we have when we are deliberating. They can represent all manner of things (objects, properties, events, etc.), from the mundane to the extraordinary: DOG, NUMBER, HEAVY, AND, MONDAY, BUS, MANDELA (small caps denote concepts). Abstractness has sometimes been taken as an indication that a representation is conceptual, but we form concepts of the concrete as well as the abstract. They are sub-propositional, in the sense that a single concept does not make a claim about the world on its own. It is only when two or more concepts are combined together into a complete thought that we have a representation with a truth condition (or satisfaction condition, etc.). When so-combined, a conceptual representation is propositional in the sense that its content is something that can be affirmed as true.

Concepts are found in our judgements, hypotheses, intentions, aims, occurrent desires, and occurrent beliefs. The very same representations might show up elsewhere. Concepts might also be computed with non-consciously. It is possible that they are re-deployed in special-purpose representations, for example as a label in a cognitive map of space. And it could be that they are stored as representations in long-term memory. I want to leave these possibilities open as empirical questions rather than ruling them out by definition. So I am introducing *concept* by indicating paradigmatic cases, rather than defining it. I point to *conscious* deliberation because that is the most convenient way of pointing to this familiar kind of thought process. However, this is not a book about consciousness. I will not be putting any weight on consciousness as such. I am concerned to characterise this kind of thinking cognitively, in terms of representations and functions, remaining neutral as to how (or indeed whether) these properties depend on the nature of consciousness.

We could adopt a wider definition of 'concept' so as to encompass categorical representations of any kind. There are forms of categorical perception (e.g. of phonemes) that doubtless depend on representations that categorise their subject matter but that do not directly show up as freely-recombinable elements of thought. (We can form concepts of phonemes, but categorical perception does not depend on possessing freely-recombinable representations of phonemes.) Much work on whether non-human animals have concepts is concerned with investigating their abilities to categorise, and to generalise in useful ways, for example, to represent the same/different distinction. Those are important and sometimes sophisticated abilities. But I need a term that is tied to our central phenomenon of interest, namely deliberation, and no better term than 'concept' presents itself to talk about the components out of which thoughts of this kind are constructed. So-understood, concepts are not limited to the categorical. They can refer to individuals and stuffs (substances: Millikan 2000); also to properties that are quantitative or graded rather than categorical. What singles out the paradigmatic instances I'm pointing to is not their subject matter, but their general-purpose recombinability in conscious thoughts.

* * *

The next four chapters of the book will flesh out the picture I have just sketched. Conceptual thought often drives the process. I will spell out how concepts held in working memory provide access to various kinds of stored information, conceptual and nonconceptual, entertained in an integrated fashion in a cognitive playground. A thought may bring to mind mental images: visual, auditory, gustatory, tactile, olfactory. It may make us feel a certain way (pleasure, distaste). One thought may prompt another. We can see all of this as building a model of a situation—of an actual or suppositional scenario. It seems that some non-local process allows the thinker to draw conclusions about what is supported by the model, or follows by extrapolating or filling in missing elements. If we understand 'inference' broadly so as to cover all these kinds of transitions between representations, then what we have is a picture of inference in the playground of concept-driven thought. Chapters 2 to 5 will consider elements of this story in turn. The aim is to develop the details of each in a way that coheres with the overall picture.

Chapter 2 examines how representations are structured. Chapter 3 looks at the computations they enter into and how that relates to representational structure. Structural representations enter into computations that depend on the relations in the structure having the particular contents they do. Representations displaying language-like compositional structure will, as we will see, afford computations of a different kind—broadly-logical transitions that depend on the general-purpose compositionality of concepts. Concepts, I will argue, enable cognition in the playground to take advantage of inferences over structural representations, often targeted at a specific domain. At the same time, concepts in working memory are a locus of domain-general recombination and the general-purpose operations of broadly-logical reasoning. (Although that is not the only way of performing inferences with conceptual representations. There is no simple dichotomy here.) We can combine and reason with concepts in ways that are almost entirely unconstrained by their specific contents, allowing us to formulate novel thoughts and derive surprising new consequences from what we already know.

Chapter 4 discusses informational models. Much of the information we rely on in thought derives from special-purpose representational systems of various kinds. Chapter 4 will go into more detail about the examples above, like cognitive maps and sensorimotor simulation, as well as models representing causal relations and other kinds of relational structure, including those used in analogical reasoning.

Using these systems offline, in prospection, is not special to humans. Tomasello (2022) points to the perceptual- and motor-planning inferences that non-human animals perform using perceptual, proprioceptive, and motor systems. Although taking place in special-purpose systems, he treats this form of planning as cognitive, presumably because it takes place somewhat offline and depends on working memory.

Chapter 5 lays out the benefits, for this kind of thinking, of having recourse to freely-recombinable concepts. Each concept affords access to information held in special-purpose systems (characterizations), as well as semantic memories (explicit conceptual representations). A concept held in working memory allows us to rely in our thinking on the deliverances of those models, and to integrate them with information accessed through other concepts. Not all special-purpose representations can be made available in the cognitive playground, but for those which are, conceptual thought plays an orchestrating role. Chapter 5 offers an account—at a relatively high level, but based on empirical results—of how this kind of thinking works and the role of concepts in it. The slogan is that concepts are ‘plug-and-play’ devices.

This discussion will prompt long-standing questions about how concepts are to be individuated and how their reference is determined. In section 5.7, I show how concepts should be individuated so as to fit into this picture—what makes a token representation an instance of the same concept again. The book will not, however, address the question of content-determination. The important question of the right metaphysics of content for concepts is one for another day. To make progress on that difficult issue it will be important first to have a detailed, plausible, and empirically well-supported account of how conceptual thinking works. That is what this book aims to provide.

1.5 Metacognition

A final reason for focusing on executive functions is that this is the sphere where metacognition is thought to operate. As we have seen, executive function encompasses several different cognitive capacities. It has sometimes been called the ‘supervisory attentional system’ (Norman and Shallice 1986), although that tends towards treating it as a pluripotent homunculus. The term does helpfully point out that executive functions include capacities for monitoring cognition itself and controlling how it unfolds. Those are the signature properties of metacognition (Nelson and Narens 1990).

Some theorists treat metacognition as a single cognitive capacity, some kind of inner eye that the mind turns in on itself. A better view is that it is an umbrella term for a variety of ways that cognitive processes are monitored, both explicitly and procedurally, and for the effects of that monitoring on how downstream cognitive processes unfold (control). Despite the label, metacognition is not an additional module or capacity that operates on top of cognition, but a type that characterises some of the executive processes that go on in cognition. Whilst some kinds of monitoring or metarepresentation may occur in special-purpose systems, outside the playground of thought (Shea 2014c), most psychological work on metacognition has focused on conscious thought and metacognition within the cognitive playground (Proust 2013; Koriat 2016; Schneider and Löffler 2016; Ackerman and Thompson 2017). There are functional reasons why monitoring for accuracy, that is assessments of confidence, may be important to the well-functioning of a cognitive playground (Shea and Frith 2019). So there is good reason for thinking that conceptual thought, which occurs at the executive level, will involve metacognition.

I have been talking about informational models of two different kinds: models in special-purpose systems; and models constructed in the playground using concepts. Special-purpose systems use a variety of representational structures to encode information in models of different kinds. In the picture I have sketched, conceptual thinking builds up informational models in the playground. Each concept is connected to its own bodies of information. When held in working memory and put together compositionally into a thought, they drive the construction of a cross-domain model, a model that can put together existing information in novel ways. Metacognitive processes are likely to monitor that model and aspects of the way that conceptual thinking unfolds.

While we do not yet have a clear picture of how this works, Chapter 8 will point to some existing empirical literature that gives an indication of how metacognition applies to conceptual thinking. The upshot is that the thinker probably has a rich metacognitive appreciation of an informational model constructed in conceptual thought: of the dependability of the concepts used to construct it, of the accuracy of the information relied upon, of the reliability of the inferences that take place, and potentially of the overall coherence of the contents of the playground. Conceptual thinking is not just taking place in the playground; it takes place in ways that the thinker themselves appreciates in various ways.

1.6 Re-casting the Distinctions

I am aiming to paint a picture of human practical and theoretical inference, arguing that we need to conceive of thinking broadly so as to encompass all the diverse resources we rely on when deliberating. My picture is that concepts act as an interface between general-purpose reasoning and many of the other information-using systems of the mind. Concept-driven thinking corrals an assorted collection of information in the cognitive playground. How can that work, when these resources are so diverse? They draw on a variety of different kinds of informational models (Chapter 4), which utilise different kinds of representational structures (Chapter 2), and run different types of computations (Chapter 3).

To construct a picture of how these elements all play together, it is crucial to characterise them accurately. Chapters 2 to 4 ground the distinctions we need in a series of empirically well-supported psychological examples. Philosophy has tended to dichotomise the mind. Representations are separated into the iconic versus the discursive; or into the conceptual versus the nonconceptual. Representational transitions are split into the rule-based versus the associative. The mind's architecture is split into modules versus central systems, or the perceptual versus the cognitive. None of these dichotomies provides quite what we need in order to characterise the elements of the picture accurately. Psychology and cognitive science show that the elements are more various than these dichotomies suggest.

In each case, the properties that underlie the dichotomy are perfectly useful. For example, iconic representations can be characterised as analogue, dense, holistic, and concrete; discursive representations as digital, sparse, arbitrary, and abstract. There is debate about which of these properties are necessary, and which if any is basic. Rather than arguing about how best to draw an analogue/digital distinction, I want to go straight to the underlying properties and use them to characterise the elements of my picture. So when I discuss semantically-significant representational structure in Chapter 2, I identify six different properties of the way representations are structured. A range of examples show that these properties are exhibited piecemeal, in different combinations, in different cases.

Many of these representational systems are compositional. Compositionality is usually seen as the core tenet of the language of thought hypothesis. The cases show that the landscape is more complex. Compositionality is exhibited in a range of different ways. Chapter 2 distinguishes the kind of compositionality at work in cognitive maps and other kinds of structural representations, which can support separate singular and general terms, and allow for rich recombination, from the language-like form of compositionality exhibited by the conceptual representations involved in conscious deliberation. The latter employ a general-purpose and semantically-neutral mode of combination, like predication in natural language. Rather than debating whether this means that the mind really uses a language of thought, or not, I concentrate on pinning down the properties exhibited by the different cases. Calling these other systems 'nonconceptual' is potentially misleading. First, it may be that concepts—representations that figure as recombinable components of deliberative thoughts—can also be deployed in special-purpose systems, for example in object files involved in tracking visual objects, or to label locations in the cognitive map. Second, many of these putatively nonconceptual systems actually exhibit, piecemeal, properties often taken to be characteristic of the conceptual. So I will mainly avoid relying on a conceptual/ nonconceptual distinction.

Similarly, when the mind is dichotomised into modules versus central systems, most of the properties used to characterise modularity are real features of some psychological systems. But rather than adding another epicycle to the extensive literature on whether the mind is modular and how modularity should be defined, it is more useful for our purposes to work directly with some of the underlying properties. Many of the informational models discussed in Chapter 4 are domain-specific, for example being specialised for visuo-motor coordination and control. 'Domain-specific' is not quite the right distinction, however. The map-type structures in the medial temporal lobe that are used to represent spatial locations are seemingly also deployed to represent other kinds of relational structures, domains where the content can be more abstract. These systems are special-purpose, in that they are suited for representing only some kinds of content, but they are not specific to a particular domain (cp. 'functional specificity'; Margolis and Laurence 2023). A related distinction is modality-specific vs. amodal. Some of the resources that I argue are involved in concept-driven thinking have been taken be modality-specific (and, further, to support a modality-specific theory of concepts; Barsalou 2008; Prinz 2002). While there are clearly amodal elements in my picture of conceptual thinking as well, I refrain from characterising the special-purpose systems as modality-specific since many, like the cognitive map, are clearly supra-modal (that may be true of all special-purpose systems; Calzavarini 2022). Accordingly, I will mostly talk about *special-purpose* informational models, to contrast with the general-purpose mode of representation exemplified by the language-like combinatorial structures of conceptual thought.

My view also differs from the Fodorian dichotomy between modules and central systems in other respects. Concepts can occur as representations in special-purpose informational models, for example for representing natural kinds, mental properties, or natural numbers (§4.7). Conversely, conscious conceptual representations are involved in transitions that are automatic and content-specific (Chapter 3).

Other properties associated with modularity I also treat piecemeal. We have seen that deliberation draws on working memory and so exhibits the characteristics of type 2 cognitive processes (§1.2). These are features associated with Fodor's central systems, contrasted with the fast and automatic operation of modules. However, some of the special-purpose informational models I discuss in Chapter 4 also exert cognitive load. They are not all simply fast and automatic. Many of these special-purpose systems rely on dispositions to make transitions between representations that effectively build in assumptions about the nature of the environment. (They perform a 'content-specific' type of computation, in the sense to be defined in Chapter 3.) This means that their operation is often somewhat encapsulated from information in other systems. But this does not imply complete encapsulation. The picture is compatible with there being a range of top-down and contextual effects on how special-purpose systems operate (and hence with the empirical evidence to that effect). Conversely, there can be fast, automatic transitions between conceptual representations. Finally, I completely avoid any claims about innateness. My picture is compatible with many different views about which capacities and representations are learned and unlearned, and the extent to which their adaptiveness depends on gene-based evolution or cultural evolution (Shea 2012). My picture has a collection of components and capacities that make up the mind, but these are not modules in anything like the classical, Fodorian sense.

Nor does my picture require a clear distinction between perception and the rest of cognition. Cases like the systems of 'core cognition' described by Susan

Carey look to be interestingly intermediate between the two. These systems deploy representations on the basis of input in a relatively automatic way, building in assumptions about the domain (e.g. of agents), but they also seem to be richly entwined with more cognitive processes. To the extent that my picture appeals to anything like a perception/cognition distinction the special feature is at the other end, the non-perceptual end, where we find deliberation. I will argue that there is something special about the way concepts combine in conceptual thought and the way deliberation has access to a general-purpose mode of non-content-specific, broadly-logical, reasoning. But 'cognition' is not a good label for this capacity. On my picture, cognition is much broader, bringing in and marshalling information from across the mind.

Perhaps the most problematic distinction is the dichotomy between rules and associations. 'Associative' is used to mean many different things. Sometimes it is about the way a disposition to transition between representations is learnt or acquired (or lost: whether it is susceptible to counter-conditioning). Other times it is about those dispositions themselves, for example 'associative' is used to talk about the way activation of one concept can prime another, like SALT-PEPPER, where the transition is not inferential or faithful to content. These are real phenomena, but they do not support a neat division into modes of cognition, the rule-based versus the associative. It is also problematic to treat classical computation versus parallel distributed processing as a matter of rules versus associations. One of the properties in the mix here is indeed computational, related to the ability to compute with variables, or to perform computations that depend on what Lake et al. (2017) call model-building as opposed to pattern matching. In Chapter 3 I argue that the most fundamental contrast is between computational transitions that are content-specific and non-content-specific, in a sense I spell out there.

In short, in laying out my account of how concepts mine information from throughout the mind and drive the construction of rich scenarios in the playground of thought, I will mostly eschew contested dichotomies and instead use the underlying properties—some of which I also re-cast—to characterise the variety of representational structures, computational processes, and informational models involved.

1.7 What's New?

My account colonises territory that is under-explored in existing work on concepts. Many have focused on the domain-generality of conceptual thought (Evans 1982; Fodor 1998, 2000). They have tended to overlook the role of concepts in deploying and organizing special-purpose resources. Other theorists have emphasised the domain-specific aspects of conceptual thought (Prinz 2002;

Barsalou 2008) while under-estimating the importance of being able to marshal these elements in general-purpose ways. My account shows that conceptual thinking decisively involves both.

The idea that concepts are an interface is not new. I draw heavily on Liz Camp's work on the variety of resources, in addition to explicit conceptual representations (psychologists' semantic memories), that concept-driven thinking makes use of (Camp 2015, 2019). She emphasises ways in which sensory, affective, and evaluative responses also characterize the subject matter a person is thinking about, whether it be a person, a foodstuff, or a favourite pet. These 'characterizations' offer us a much richer conception of the range of information accessed through a concept. But the hypothesis that concepts act as an interface does not yet answer our question about how real-world theoretical and practical inference actually works. It raises a knot of thorny problems that need to be addressed if the hypothesis of concepts-as-interface is going to fly.

Special-purpose systems encode information about the world in a variety of ways: a suite of sensorimotor expectations; relational structure in a spatial map; similarity structure in a social-semantic space representing the people we know. Can the deliverances of diverse informational models be integrated? How do the various representational structures inter-operate? Many special-purpose systems perform computations of a different type than those driving conceptual reasoning. Can one cognitive process have recourse to both? And what kind of cognitive process is thinking, exactly? Some parts seem to be fast and automatic (type 1), but the overall phenomenon is deliberate and controlled (type 2). Working memory plays a role, as does metacognition. Which cognitive capacities feed into concept-driven thinking and how do they work together?

The aim of the book is to answer those questions and assemble the answers into a coherent picture. The answers are empirically driven. There is also theoretical work to precisify, and in some cases re-cast, existing distinctions. The way I put the pieces together is, I think, novel.

The picture is of course only tenable if it is empirically well-supported. But it should also be assessed as a whole, for its theoretical coherence, and for what it can explain. So the book is not a point-by-point argument that my theory is superior to other views. The book is focused on constructing the picture. The aim is to synthesise elements that appear in different streams of current work and arrange them into a more full-bodied account of conceptual thinking. That picture will only succeed if it forms an attractive whole. If it does, it shows that the hypothesis of concepts-as-interface is tenable. So the first aim is that the overall picture should make good on the interface hypothesis. It also has explanatory payoffs: it can help us deal with some other important problems. That is the focus of the second half of the book.

So, having expended so much effort on the details of the view in the first half of the book, the aim of the second half is to showcase some of the things this new picture allows us to explain. First off, it explains how deliberate thinking manages to dance around the notorious frame problem (faced by classical AI systems), while at the same time transcending the limits imposed by the lesser-known 'if-then' problem (Chapter 6; see e.g. Gallistel and King 2009). Concept-driven thinking can make use of the built-in assumptions of special-purpose informational models, while at the same time having access to general-purpose compositional power and content-general computational processes. Acting as an interface, concepts allow us to take advantage of the benefits of each and to circumvent their complementary limitations.

Chapter 7 turns to the commonplace but elusive idea that the unfolding process of thinking draws on the meaning of the thoughts involved. That has proven difficult to reconstruct in the context of the representational theory of mind, with its commitment to transitions between representations being implemented in causal transitions between representational vehicles caused by their non-semantic properties. While RTM has proven to be an extremely fruitful framework and has strong empirical support, it has left a puzzle about what role semantic contents are playing in thought, for the thinker. My picture shows that thinkers are right to think that much of their thinking draws substantially on the meaning—referential content—of the representations involved (while not on its own answering questions about the causal efficacy of content). A conceptually-constructed informational model in the cognitive playground supports content-driven, abductive inferences.

The picture makes more space for the role of the person in conceptual thinking. While not being homuncular, by incorporating metacognitive processes we get a slightly clearer glimpse of the thinker in the process of conceptual thinking (Chapter 8).

Finally, painting on an even larger canvas, the account explains why human concept-driven thinking is an especially powerful way of inferring new facts and planning for the future. The unreasonable power of human cognition lies in our ability to go beyond, as well as to perform, reasoning (Chapters 5 and 9).

In short, the ambition of the book is to characterise the elements of thinking (Chapters 2–4) so as to construct a picture of how concepts act as a psychological interface—keys that unlock the mind's many informational resources and assemble them in an integrated way in the cognitive playground (Chapter 5); and then to show that the resulting picture has significant explanatory payoffs (Chapters 6–9).

Chapter Summary

1.1 Concepts in the Playground of Thought

The book is about how conscious deliberation—practical and theoretical inference—works. Much is done offline, when there is spare cognitive capacity; sometimes also in a few moments before responding to a situation. Humans do a lot of deliberation. Some thinking is a matter of step-by-step reasoning, logic being the paradigm. (p. 2)[2] But much thinking involves simulation and imagination, drawing on the mind's other resources. Reasoning is important for its generality and the ability to transcend the assumptions of special-purpose systems. The way this happens depends crucially on concepts, which can both figure in reasoning, and also integrate many other forms of information.

Human conceptual thinking underpins many of the remarkable achievements of our species. (p. 3) Concepts are crucial parts of the jigsaw, combining to form a thought but then connecting to information stored both in conceptual representations, and in special-purpose systems: sensory, motoric, evaluative, and affective. This picture calls for an account of how representations are typed and combined, computed with, and used to form informational models of the world. The psychology of concepts has tended to focus on categorisation, philosophy on reasoning; the book aims for a comprehensive account.

(p. 4) Thinking takes place in a shared representational space where different kinds of information and competing evaluations can be considered together. These are the functional features of a global workspace (bracketing any connection with theories of consciousness). A better idea is the cognitive 'playground'. (p. 5) The cognitive playground has the functional property that representations are maintained together and manipulated in relation to one another (glossing: 'playground', 'thought', 'thinking').

1.2 Deliberation

(p. 6) To put it another way, the book is about the role of concepts in executive processes (glossing 'deliberation'). Executive processes operate on representations in working memory, with a tight capacity limit on how much information can be manipulated at once. The contents of working memory are determined by how attention is directed; doing so skilfully is a cognitive capacity with wide-ranging consequences. (p. 7) Working memory offers a functional characterisation of how thinking has the capacity to manipulate diverse informational resources.

Using the type 1/type 2 distinction, deliberation works in the type 2 way—it depends on our limited capacity to direct attention to supervise the contents of working memory. Step-by-step or type 2 reasoning is a clear instance of capacity-limited conceptual thinking. (p. 8) A thought constructed out of concepts in working memory drives the construction of a suppositional scenario in the cognitive playground. This is related to the model-based/model-free distinction. 'Model-based' connotes four features which, although separable, come together

[2] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

in intelligent deliberation. (p. 9) This process has access to the current values and goals of the agent, from which special-purpose systems and model-free decisions may be disconnected.

1.3 Non-Local Inference

As well as reasoning between conceptual representations, step-by-step, we can perform inferences that take account of a wide array of considerations at once (glossing 'reasoning' and 'inference'). Example: multi-factor choice; high-dimensional category learning; and, in philosophy, abduction. (p. 10) To account for non-local inference, we need to enlarge our conception of computation (that is the truth behind Fodor's critique). We do now have models of non-local computations, taking place over an array of representations, that respect RTM's stricture that transitions occur in virtue of vehicle properties. Non-local inference also occurs in conscious thinking, for example deciding how to walk through a room at a party. (p. 11) For now, I am simply characterising the phenomenon roughly by reference to a few examples.

Such inferences occur over a more complex representation or set of representations, some kind of 'model' of a situation. 'Model' has an intended resonance with scientific models (contrasted with theories), and with model-theoretic entailment in logic. (p. 12) An informational model in the cognitive playground is subject to some kind of check for consistency.

1.4 Concepts as an Interface

Concepts (glossed here) are sub-propositional freely-recombinable elements of the thoughts we have when deliberating. (p. 13) This is not a definition, but a way of pointing to paradigm instances. It does not cover categorical representations that do not display general-purpose recombinability in deliberate conscious thought.

The book will build up a picture of how conceptual thought drives inferences taking place over a collection of representations, of different kinds, in the cognitive playground. (p. 14) Chapter 2 examines how representations are structured and Chapter 3 looks at the computations they enter into and how that relates to representational structure. Chapter 4 surveys different kinds of informational models. Some other animals can use informational models in prospection. Chapter 5 shows how conceptual thought makes this more powerful: concepts act as a locus for accessing and combining information from a wide array of different special-purpose informational models (they are 'plug-and-play' devices). (p. 15) Chapter 5 offers an account of concept individuation; the fundamental issue of content-determination is not addressed in the book.

1.5 Metacognition

Executive functions include the capacity for metacognition: for monitoring and controlling how psychological processes operate. Metacognition covers a variety of ways that concept-driven thinking is monitored and controlled. Metacognitive processes can monitor the current cross-domain model in the cognitive playground. (p. 16) Chapter 8 points to some existing empirical literature; the upshot is that concept-driven thought is something that the thinker themselves appreciates in various ways.

1.6 Re-casting the Distinctions

How can concepts act as an interface between diverse kinds of informational models, representational structures, and computational processes? We need more subtle distinctions than the standard philosophical dichotomies provide. (p. 17) My tactic is to characterise a variety of psychological cases by reference to the useful properties that underlie the common dichotomies, for example six different properties of semantically-relevant representational structure (Chapter 2).

Compositionality itself comes in different forms, for example in cognitive maps vs. the general-purpose language-like recombination of concepts. Related to modularity, many of the informational models in Chapter 4 are special-purpose (some not all specific to a particular domain, but probably not modality-specific). (p. 18) My view also differs from Fodor's modules/central-systems dichotomy in the way properties are distributed. Special-purpose systems can be fast and automatic, but may instead exert cognitive load; they have built-in assumptions and can be partly encapsulated, but admit a range of contextual and top-down effects. I do not rely on a clear perception/cognition distinction—general-purpose concept combination (Chapter 2) and content-general transitions (Chapter 3) are probably proprietary, but cognition goes much wider. (p. 19) Nor do I work with a dichotomy between rules and associations, although the distinction in Chapter 3 is in the vicinity. In short, I aim to use underlying properties directly to characterise the variety of representational structures, computational processes, and informational models involved.

1.7 What's New?

Previous theories of concepts have focused either on the domain-general or on the domain-specific; the account here decisively involves both. (p. 20) I rely heavily on Camp's idea that concepts access a wide range of conceptions; the aim is to show how these are integrated in theoretical and practical inference. I ask: which cognitive capacities feed into concept-driven thinking and how do they work together? Any novelty is in the way I put together the pieces (and perhaps in how I re-cast some of the distinctions). The picture should be assessed as a whole, including for its explanatory payoffs.

The picture developed in the first half of the book shows how concept-driven thinking manages to circumvent the frame problem and its converse, the 'if-then' problem (Chapter 6). (p. 21) It vindicates the idea, compatibly with RTM, that much thinking draws substantially on the meaning of the representations involved (Chapter 7). By including a role for metacognition, it brings the thinker somewhat better into sight (Chapter 8). Most ambitiously, the picture shows why human concept-driven cognition is an especially powerful way of inferring new facts and planning what to do (Chapters 2, 5, and 9). In short, the book characterises the elements of (Chapters 2–4), and sets out (Chapter 5) a framework, and shows what it can explain (Chapters 6–9).

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0001

2
Representational Structure

2.1 What is Semantically-Significant Representational Structure?

This chapter is about the different kinds of representational structure exemplified by the representations involved in concept-driven thinking. I start off by saying what representational structure is, and identifying six aspects of semantically-significant representational structure. Section 2.2 shows that these six features are exhibited, piecemeal, by mental representations of various kinds. Section 2.3 sets out what it takes to be a structural representation, and how this differs from representational organization, with which it is sometimes elided. Section 2.4 examines whether the general-purpose compositionality of language and concepts is an instance of structural representation. I argue tentatively that it is not; or, if it is, that conceptual representations and sentences are structural representations of a special kind.

Representational structure interacts with the way that representations are processed. Chapter 3 distinguishes two broad kinds of computational process. We will see that, although there are no necessary connections, there are reasons why representational structure tends to align with computational process. Different kinds of representational structure also form the basis of the different varieties of informational model described in Chapter 4. These are the elements that need to fit together in order to give an account of concept-driven thinking (Chapter 5).

First off, what is representational structure? External representations give us familiar examples. A public language sentence is structured out of words. A cartographic map is structured out of marks and symbols arranged spatially in two dimensions on the page. Some simple signals have no representational structure at all: think of the ringing of a fire alarm bell or the tail-slap of a beaver to signal danger. The physical vehicle does have structure—a fire alarm is a series of rings—but its structure has no representational significance. It does not affect the content carried by the representation. (It could, but in these cases does not.)

In the case of a sentence, many ways of dividing up marks on the page do not correspond to representational structures (Burge 2018). The representational structure of a sentence is grammatical or syntactic. In the sentence 'the dog bit the man', 'the d' is a part, but not a syntactic or representational part. Nor is 'dog bit the', given the way the sentence is parsed grammatically; whereas 'the man' is a representational part. A map can be divided into spatial parts. Most contiguous regions are representational parts; but not, for example, a region that bisects a symbol on the map. The symbol for a pub could be split into handle and glass. The half symbol is not a representational unit on the map. Not every way of dividing a representational vehicle into parts produces semantically-significant units.

This point makes for terminological difficulties: a representation can have structure that is not semantically-significant, so does not count as representational structure in the standard sense. The ringing of a fire alarm is a representation which, while having structure, lacks semantically-significant representational structure. I will follow the standard usage, so that representational structure (unmodified) is taken to be semantically-significant representational structure, disambiguating explicitly when necessary. This is one of several places where the terminology becomes complex. So, as well as explaining how I am using each new term when I introduce it, at the end of the chapter I have drawn together a list which recaps how I am defining or using each term (§2.5).

A further terminological difficulty is that 'structural representation' is the standard term for a specific kind of representational structure: representations that rely on a structural correspondence between representation and world (Swoyer 1991; Ramsey 2007, pp. 77–92; Shagrir 2012). Maps are one example. This is just one kind of semantically-significant representational structure. Sentences have a different kind of representational structure (although whether this really is different is a matter of controversy, see §2.4). I will not attempt to legislate a new term for structural representation, so it will be important to bear in mind that a representation's having semantically-significant representational structure does not entail that it is a structural representation.

In the case of maps and sentences, representational structure is part-whole structure. The whole is a complex representation. Other representations are parts. The meaning, content, or semantic significance of the whole is systematically related to the meanings of the parts and their mode of combination. One way of being 'systematically related' is that the meaning of the whole is determined by the meaning of the parts and their mode of combination. That is the standard account of sentence meaning. It is not the only option. Another way of being systematically related is that the meaning of the parts derives from the meaning of the whole, given the mode of combination. For example, in certain kinds of map the points on the map pick out locations in virtue of their relations to other points in the structure (Shea 2014a, p. 131; 2018, p. 125).

Representations can have constituents in a way that is not a matter of part-whole structure (certainly not spatial part-whole structure). For example, a distributed pattern of neural activation may be decomposable into separate activation vectors (Smolensky 1988, 1995; Shea 2007; Eliasmith 2013). One component may carry information about colour and another about direction (Mante et al. 2013). If the separate activation vectors are functionally significant, then it may be right to think of the distributed pattern of activity as constituted by the combination of the component vectors. The distributed pattern of activation is a superposition of component patterns. The vectors are not spatially separable. Each neuron's activity is affected by contributions from both vectors. In what sense, then, are they separate? There is a strong analogy with the way two waves add up when they pass through the same space. Consider two searchlight beams crossing in an X. At the point where they cross, the light intensity at any moment will be the sum of that from both beams. (The waves can cancel out as well as combining into higher peaks.) Nevertheless, each beam retains its identity and continues onwards towards its target. In the same way, a distributed pattern of neural activation may be the sum of component vectors, components that act independently in the way processing unfolds. In the strongest case, where the component vectors are orthogonal, operations can be performed on one component without having any effect on the others. Thus, superposed patterns of activation can be representational constituents.

It will be convenient to call representations—the entities that carry representational content—'vehicles' of content. 'Vehicle properties' are non-semantic properties that are or may be relevant to the way representations are combined and processed. Since vehicle properties can carry semantic values, they too should be considered to be vehicles of content. So 'vehicle' can be used both for particulars (representations) and for their (non-semantic) properties.

Semantically-significant representational structure is a functional notion (Burge 2018; Lande 2021). It depends on how a putative representation is formed and processed, and also on which (non-semantic) properties are causally relevant to those transitions. Written words are an obvious case. With words and sentences, representational structure starts from the sequential order in which they are written. The syntactic or grammatical structure of a sentence divides up words in the sequence into a higher level pattern (classically, a tree structure). But vehicles in the brain are quite unlike written words. In general, vehicles are individuated functionally. As we have just seen, constituents need not be spatially separable.

In the case of mental representation, the entities that count as vehicles may be quite abstract by reference to activity in the brain. What RTM is committed to, however, is that vehicles and vehicle properties are a legitimate non-semantic way of characterising the causal dynamics of the system (Shea 2018, pp. 37–41).

The components out of which a representational structure is composed must be psychologically real: they must figure in psychological processes and competencies (Burge 2018). It is constituency relations amongst the entities and properties so-characterised that are the basis for representational structure.

Representational structure is a matter of how vehicles and vehicle properties are combined. The representational structure of the whole is fixed by vehicle properties of the constituents and how they are put together. The structure of a representation is semantically significant to the extent that vehicle properties and relations have semantic import; that is, carry semantic values or make a difference to the content of a complex representation. For example, the compositional structure of the sentence 'man bites dog' makes it the case that its content concerns a dog being bitten rather than a dog doing the biting. Other vehicle properties, like the typographic font, have no semantic significance.

The cognitive sciences have developed a variety of techniques for investigating representational structure. It can be inferred from: which entities and properties are and are not distinguished; patterns in reaction times and errors; indirect facilitation and interference effects like priming or neural repetition suppression; and patterns of breakdown under cognitive load, as a result of pathology, or by direct interference (e.g. via transcranial magnetic stimulation, TMS). There is also an increasing range of techniques that measure vehicles directly, for example through recording neural activity with electrodes or by multivariate pattern analysis of fMRI data.

Representational structure is also inferred by investigating constraints on what can and cannot be represented together. For example, a cartographic map cannot be used to represent the spatial relation between Almaty and Bishkek, and between Bishkek and Chimkent, without also representing the spatial relation between Almaty and Chimkent. That is, representational structure determines 'distributional properties' (Lande 2021): which representations can, cannot, and must co-occur. Lande (2021) shows how this logic works in vision science: experiments involving adaptation have allowed researchers to discover the distributional properties of visual representations and thereby to infer their semantically-significant representational structure.

There is much philosophical work taxonomizing representational structure into broad kinds, usually dichotomies. These broad kinds are known as representational *formats*. The most prominent format distinction is between iconic and discursive representations (Quilty-Dunn 2020). As explained in the first chapter (§1.6), these dichotomies are not ideal for my purposes, either being too contested, or running together different properties that we will need to separate.[1]

[1] Coelho Mollo and Vernazzani (2023), preprinted on arXiv just before the book went into production, make a similar argument. Further, they make a persuasive case that existing work on representational formats in cognition has relied too heavily on analogies with public external representations.

I will instead proceed by working directly with six aspects of semantically-significant representational structure that are identified in this literature. These features are exhibited, amongst other places, by the representational structure of natural language sentences:

(a) They have semantically-significant components (e.g. words) (unlike an unstructured alarm call).
(b) The components can be tokened separately, one without the other.
(c) When tokened together, the components are typically 'bound' together by a device whose semantic significance is that their contents should concern the same state of affairs, rather than being a simple list of contents.
(d) Each constituent is 'incomplete' in the sense that it does not, on its own, make a claim or pick out a worldly condition.
(e) Different components make different kinds of contribution to the overall content, for example some act as singular terms that represent particulars, others as predicates that represent properties or relations.
(f) The representation makes use of a general-purpose device of concatenation: predication, or an even more general-purpose form of concatenation (like Merge in natural language).

These features are also exhibited, piecemeal, by representations of other kinds. I will go on to argue that the conceptual representations entertained in conscious deliberation have all six features. There have been many attempts in philosophy to distinguish between conceptual and nonconceptual representations. Several different ways of drawing the distinction are on offer. Since the structural features of paradigmatic conceptual representations do not come and go together, but are found piecemeal in other cases, any way of drawing a dichotomy risks being misleading. It is better to accept that there is a plurality of types of representation, exhibiting a variety of kinds of representational structure (Camp 2007, 2018). I have instead pinned down what I mean by *concept* by reference to their paradigmatic occurrences—concepts are generally-recombinable constituents tokened in deliberative thinking—leaving open that concepts may also figure in other places (e.g. in non-conscious processing, as constituents of some representations in special-purpose systems, or in long-term memory). The next section examines how these structural properties are exemplified across a range of psychological cases.

2.2 Varieties of Semantically-Significant Representational Structure

This section surveys a variety of forms of representational structure exemplified by mental representations of different kinds. We will see that the features just listed, derived from language-type structure, are in fact found, piecemeal, in mental representations of various other kinds.

The base case is where there is no structure at all. This may be rare when it comes to mental representations. With external symbols, a paradigm is the vervet alarm call signalling system (Seyfarth, Cheney, and Marler 1980). There are three different calls, not systematically related to one another, each arbitrarily related to a different kind of predator. Vervets have evolved and/or learnt to react appropriately to each type of call. The animals do not register or make use of any relation between the different calls. These are *nominal signs* (following Godfrey-Smith 2017; see also Planer and Godfrey-Smith 2021).

Nominal signs lack any semantically-significant representational structure. Nor are the relations between them of any semantic significance. Planer and Godfrey-Smith (2021) give an example of a nominal sign system established by explicit convention: as used by Paul Revere, one lantern meant *the English are coming by land*, two lanterns meant *the English are coming by sea*. The latter sign is built of out two lanterns, but the spatial structure has no semantic significance. It is not semantically-significant representational structure. Nor are the relations between the two signs, for example that the second is twice as bright as the first, of any significance.

Most examples of nominal sign systems are non-psychological. They concern signals sent between organisms, or internal signalling within the body, for example by hormones. When it comes to mental representations, even the simplest representations tend to be more than a simple on/off signal. They usually come in degrees. A neural signal that registers surprise at a novel stimulus correlates with how unusual the stimulus is (Polich 2007). Similarly, the dopamine signal registering reward reflects the degree to which the reward was unexpected (Rushworth, Mars, and Summerfield 2009). There are many types of mental representation that track quantity, like the much-studied analogue magnitude system that tracks the number of discrete objects or events encountered (Nieder and Dehaene 2009). In these examples there is a system of interrelated representations, of which only one is tokened at a time. For example, the level of neural activation may represent numerosity. Different levels of activation represent different numerosities. Only one level of activation is tokened at a time.

These cases differ from nominal sign systems in that there are relations between the different representations. Since no two representations are tokened as a part of the same structure, these are not structural representations. The representations do not have semantically-compositional components. Nevertheless, the way the representations are interrelated in a family can be computationally useful, as we shall see (§2.3). They display what Peter Godfrey-Smith calls 'organization' (Godfrey-Smith 2017; see also Planer and Godfrey-Smith 2021). Such families of organized representations are one step up from nominal signs.

The next case is where a representation has more than one semantically-significant dimension of variation. For example, the honeybee nectar dance has two dimensions of variation: angle to the vertical and number of waggles. Neither dimension can be tokened without the other. A dance must occur at a certain angle and will always consist of a certain number of waggles. However, these two vehicle properties have semantic significance independently of one another. A dance at θ degrees to the vertical means: *there is nectar along a bearing θ degrees ahead of the direction of the sun*. A dance of *n* waggles means something of the form *there is nectar x metres away*. Consumer bees could be set up to forage based on one of these dimensions while remaining oblivious to the other. The semantic significance of the representation does not depend on their being combined (as it does with predication in natural language). Each dimension carries its own correctness condition. (Contrast predication: *Fa* has a truth condition but *F* and *a* do not.) The representation has two semantically-significant components, tokened together, but the two components are not bound together using a semantically-significant mode of combination.

Some mental representations work in the same way. Mante et al. (2013) studied how macaque monkeys performed a task that involves looking at arrays of moving coloured dots (see also: Thura et al. 2022; Langdon, Genkin, and Engel 2023). The researchers discovered a distributed pattern of activation in prefrontal cortex that carries both colour and motion information. The two properties are registered by independent dimensions in (high dimensional) activation space. Which dimension the animal relies on to guide behaviour depends on whether colour or motion is relevant to the judgement they have to make on a particular trial. Any pattern of distributed activation necessarily has a component along both dimensions. Neither can be tokened without the other. But they have separate semantic significance, each carrying its own correctness condition (Shea 2018, pp. 100–3), rather than being combined (e.g. one being predicated of the other).

By contrast, in early visual processing it is likely that representations of attributes like colour and motion can be physically tokened separately, so that one attribute can be represented without representing the other (Gazzaniga, Ivry, and Mangun 2019, p. 198). For example, if there are separate feature maps for colour and contour curvature, then the way the colour of a location is represented does not require anything about its curvature to be represented. They are not part of the same representational structure, as they are in the Mante et al. example in prefrontal cortex.

Something more is needed if two attributes are to be represented as instantiated together, for example as features of the same object or of the same location. They need to be 'bound' (Treisman 1996). Binding is a matter of vehicles for the two attributes being combined together in a representational structure. For example, there is good evidence for *object files* in visual processing, or in visually-driven working memory, in which various properties are predicated of a perceived object that is being tracked through space, including as it moves in and out of sight behind other objects (Pylyshyn 1989; Quilty-Dunn 2016, 2020). Visually-registered properties may also be co-attributed to the same spatial location, or each attributed to the other in some way. The semantic significance of any of these forms of binding is that both features are attributed to the same object, event, or state of affairs. Consider someone seeing a blue circle at the same time as seeing a green triangle. The visual system is registering both the circle and the triangle. There are also visual representations of the two colour properties, blueness and greenness. The significance of binding is that the blueness is attributed to the circle and not to the triangle. Some device of concatenation connects the representation of blueness with the representation of circularity. (This may be by predicating both of a location or of a visual object representation.) The semantic significance of the way the two representational vehicles are combined is that blueness and circularity are represented as properties of the same object. Because of binding, the visual system is able to represent the difference between encountering a green circle plus a blue triangle, on the one hand, and a blue circle plus a green triangle, on the other.

In this case the representational constituents do not make a claim about the world on their own, but only when combined into a complex representation. Each is *incomplete*. In the same way, in the sentence 'Layla runs', both the singular term, 'Layla', and the predicate, 'runs', are incomplete. Only the whole sentence represents what we can call a 'complete condition'. In this case the complete condition is a truth condition. In other cases it is a correctness condition, accuracy condition, or satisfaction condition. I am stipulating that a *complete condition* is found at the level of facts or states of affairs, something that could be the case, or against which the world could be assessed (also at the level of a proposition, but calling a content propositional often carries further theoretical baggage).[2] With a nominal sign, its content concerns a complete condition even though the representational vehicle has no semantically-significant representational structure. With incomplete representations, they must be bound together in order to represent a complete condition.

Investigating what kind of concatenation is at work in the visual system is an important question for vision science. Feature representations may combine with an object representation to make a claim about an object (*that object has a green surface*).[3] Or the attributes may be represented in a feature map that ascribes properties to an array of locations (Treisman and Souther 1985; Clarke 2021), with each colour representation combining with the map so as to represent the

[2] The term is slightly awkward in that being incomplete does not imply that the system forming the representation has more work to do. A perceptual mechanism that has the function of identifying (Burge 2010) has not fallen short when it outputs a representation with a content like a demonstrative noun phrase (*that F*). However, we need a term for the contrast, and alternatives (e.g. unsaturated, sub-propositional) have disadvantages of their own.

[3] Contrast a perceptual attributive, e.g. *that green surface* (Burge 2018, p. 90).

colour at a particular location. Either way, the constituents are incomplete. They do not individually concern a complete condition, a way that the world could be. In this respect they contrast with the bee dance case, and with the PFC colour-motion case, in which two complete conditions are represented, separately, by two different aspects of a complex vehicle.

The constituents of a sentence ('Layla runs'), in addition to being incomplete, are also physically tokenable separately ('Layla'). The bee dance is a case where two components are not tokenable separately, but each represents a complete condition. (Also in the Mante et al. 2013 colour-motion case.) There are other psychological cases where there are incomplete components which are not tokenable separately.

An illustration of that is found in the hippocampal place cell system. This is an example we will return to repeatedly, so it is worth introducing in some detail. The medial temporal lobe contains a system which represents spatial locations and is used for navigating through space, a 'cognitive map' (O'Keefe and Nadel 1978; O'Keefe and Burgess 1996). The cognitive map relies on various components: place cells, grid cells, head direction cells, landmark cells, object-vector cells, border cells, etc. (Grieves and Jeffery 2017). Structural representations in the medial temporal lobe have been most extensively studied in rats, in relation to space, but there is converging evidence in humans where, in addition to spatial locations, the same neural structures can also come to represent more abstract relational structures (Schuck and Niv 2019; Liu et al. 2021). An important component of the cognitive map is the array of place cells found in the hippocampus. Place cells show remarkable sensitivity to spatial location. When active during movement, place cell firing correlates with the animal's location in a given arena. Irrespective of which way the animal is facing or what it can see, a given place cell fires when and only when the animal is at a certain location. Place cells retain their spatial sensitivity in the dark, updated based on feedback from the animal's self-generated movement. In this 'online' mode of operation, place cells correlate in a highly reliable fashion with current location.

The representational structure we are interested in is observed when the cognitive map is taken offline—when it is not being driven by current sensory input. In this offline mode of operation, chains of place cell activity correspond to potential routes through the environment (Dragoi and Tonegawa 2011; Wang, Foster, and Pfeiffer 2020). This is not a matter of the way place cells are spatially arranged within the hippocampus. Rather, as a result of the experience of moving around, neural interconnections form between place cells that correspond to neighbouring locations (Fig. 2.1). Place cells for nearby locations that 'fire together', 'wire together' (or were already wired together). This pattern of synaptic connections means that, when place cells are active offline, one cell activates another in a way that reflects the spatial proximity of the places to which they correspond. We will focus on the patterns of activity that occur when place

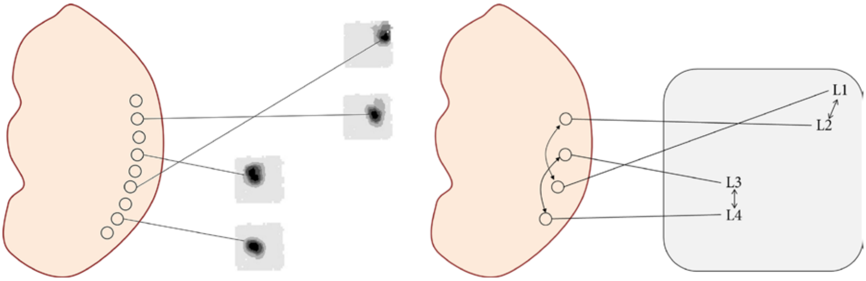


Fig. 2.1 The left panel schematically illustrates an array of place cells in the CA1 layer of the hippocampus. Heat maps are shown for four of the place cells, indicating where each is active when the animal is moving around freely in a square arena (viewed from above). Cells corresponding to nearby locations (black blobs that are close together in the square arena) need not be close together in the hippocampus. Nevertheless, as shown in the right panel, cells that correspond to pairs of places that are near one another (L1 and L2; L3 and L4) are interconnected, and thus, in offline processing, tend to activate one another.

cells are active, leaving aside the standing network of synaptic connections. Co-activation of vehicles reflects proximity of locations.

These relations of co-activation are used to calculate potential routes through the environment. Routes through the environment are assigned value based on how readily they lead to rewarding outcomes (Mattar and Daw 2018; Krausz et al. 2023). There are various ways this could happen, but for simplicity we can think of the system in offline mode as simulating various potential routes. It starts by activating a target location to be reached and runs through sequences of co-activation that trace back to the current location. It picks the shortest of the simulated sequences and, switching to online mode, follows that sequence in reverse in order to reach the target. The role of place cells in route planning gives us good reason to think that place cell activity represents spatial locations, and that the activation of one place cell by another represents that the two corresponding locations are near one another. I won't attempt to reproduce that argument here (see Shea 2018). For the sake of the example we can simply take it that the activation of a place cell represents a location and co-activation of two place cells represents spatial proximity between two locations.

This, then, is an example of a representational structure consisting of incomplete components that cannot be tokened separately. The vehicles of content are place cell activations. The relation of co-activation between vehicles represents the relation of spatial proximity between locations. Where cell 1 represents location 1 and cell 2 represents location 2, the activation of cell 1 by cell 2 represents that location 1 is near to location 2. In this offline mode of operation, the activation of a single cell in isolation does not represent a complete condition. It is incomplete. Only in combination does the activation of two or more place cells represent a complete condition (e.g. that location 1 is near location 2). There is also an incomplete representation of proximity. The vehicle property that carries the content *nearby* is co-activation. Unlike in the case of a sentence, however, the relation (co-activation) cannot be tokened without tokening the incomplete constituents between which that relation obtains.

The way space is represented in offline place cell activations exemplifies many but not all of the above-noted structural features of natural language sentences (§2.1): semantically-significant representational components, which are incomplete, making semantic contributions of different kinds (some representing particulars, others relations), with contributions that are semantically bound to one another—i.e. (a), (c), (d), and (e) above, but not (b) (being separately tokenable). The same is true of the way cartographic maps represent spatial relations between places. A map cannot token a representation of a spatial relation, like *being 10 km to the west of*, without tokening representations of the entities (e.g. towns) between which that relation obtains.

The cognitive map is an example of a structural representation (on which more in a moment). It uses a device of semantically-significant concatenation: activating one place cell after another. But this is a special-purpose device of concatenation. It has the significance of predicating spatial proximity of two locations. This is unlike the general-purpose device of concatenation by which words are connected in a sentence; or the general-purpose way in which concepts are combined in conscious deliberate thoughts (meeting the 'generality constraint'; Evans 1982). In the visual system we saw a case where incomplete constituents can be tokened separately. There, too, however, it appears that the device of concatenation is probably special-purpose: a mechanism that binds together visual features into bundles, or binds visual features to perceptually tracked objects. Neither the place cell system nor visual feature binding makes use of a domain-general device of representational concatenation, of the kind found in natural language sentences.

This brief survey of just a few selected examples is enough to give a sense of the variety of types of representational structure exemplified by the mental representations studied by psychology and cognitive neuroscience. The features that are exemplified together in natural language sentences and canonical cases of conceptual representation are also found at work, piecemeal, in mental representations of other kinds.

2.3 Structural Representation

The aim of the next two sections is to further characterise two broad kinds of representational structure: the structure exhibited by maps and the structure exhibited by sentences. To start with, this section focuses on the representational genus of which maps, both cognitive and cartographic, are a species. The genus is structural representation.

A structural representation makes use of a structural correspondence between representation and world. With a cartographic map, the correspondence is between spatial structure on the page and spatial structure in the world. As we saw, in the cognitive map in the hippocampus, the correspondence is between the relation of co-activation on place cells and the relation of spatial proximity between places. Because the correspondence is representational, this is a case of structural representation. A relation on representational vehicles represents a relation on entities in the world:

> *Structural Representation*
>
> A complex representation in which a relation on representational vehicles $v_{1,}\ldots, v_n$ represents a relation on the entities represented by $v_1,\ldots, v_n$

It is hard to get clear evidence that mental representations are structural representations. The cognitive map is a notable exception; we will see some further plausible examples below. But the clearest cases arise with public external representations. For example, when I list the students in my class by test score, I have created a structural representation. The relation of higher/lower in the list represents the relation of better/worse academic attainment. I could do the same by reading out the names in order (as in an old fashioned classroom). Then the relation of temporal order would represent relative academic attainment. Many familiar examples use spatial arrangement or temporal sequence to stand for other relations. We could use spatial relations to represent relative wealth, for instance. I can plot my students on a graph with parental income on the horizontal axis and average test score on the vertical axis. Relations between points along the horizontal dimension then represent relative affluence.

Colour is also often used to represent property values on a map or graph. To represent the weather, a temperature heat map uses colour to show the current temperature across the country. Colour relations between points can be taken to represent temperature relations between places. The same device could be added to our graph of student test scores. Colour the points from yellow through orange to red depending on the students' high school exam grades at admission. Colour relations allow us to read off, at a glance, students' relative academic standing when they started the course. The resulting graph is a structural representation of relations in all three properties: test scores, parental income, and high school performance.

There is considerable controversy about whether a theory of content can rely on structural correspondence to play a content-determining role. The basic worry is that structural correspondence and related notions (isomorphism, homomorphism) are too liberal for the obtaining of a correspondence to be the reason why a representation represents as it does. I have argued that it can, when suitably constrained (Shea 2018, pp. 111–26). I won't recapitulate that argument here. The stronger claim about content-determination is not needed for our purposes. The definition of structural representation does not require that the correspondence between relations is doing the content-fixing. It simply requires that a relation carries content (a relation represents a relation). In the case of our student graph, for example, the representational contents are set up by stipulation or convention. That is why colour representations represent relative school performance. What makes the graph count as a structural representation is that they do so represent.

With endless ingenuity, people have used spatial relations on the page to represent all kinds of relations, concrete and abstract, mundane and transcendent. Very many relations correspond in some way to the structure of space. However, there is a strong limitation once a representational scheme has been established (by convention or use, cp. Shea 2018, pp. 120–37). In my list of students, above/below on the list represents relative test score. I can use those relations to perform computations, for example to sort the students into similar-ability groups. There is a certain amount of compositionality. For any two students, I can represent that they have similar scores (by putting them one above the other on the list) or that they have very different scores. The representational system we have established allows us to represent any relation of relative test performance between any two or more students. What it cannot do, without being changed or supplemented, is to represent other relations between these students. Contrast the way we represent relations in natural language. Using the structure of a sentence to represent one relation between individuals ('Aisha loves Milly') does not preclude using sentence structure, deploying other words, to represent other relations between those individuals. If a structural representation uses spatial relations to represent relative test scores, that spatial relation cannot be used at the same time to represent other relations. The correspondence is giving the content, and that limits the representational significance of the semantically-significant relations within a given system of structural representation. This limitation will prove to be important in the next chapter, when we consider how representational structures enter into computations of different kinds.

Structural representations often have another notable feature: holism (Camp 2018). Consider a cartographic map showing the location of the towns in a region. Adding a new town to the map simultaneously represents its relations to all the other towns on the map. In the spatial cognitive map, if offline activity involved spreading activation across the whole array of place cells, as in the computational model in (Corneil and Gerstner 2015; Khajeh-Alijani et al. 2015) (Fig. 2.2), then this would be a holistic representation in the same way (Camp 2022). Even short

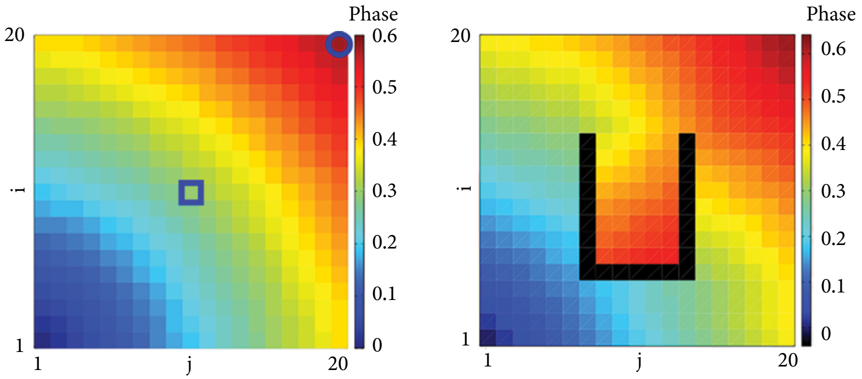


Fig. 2.2 Plots from a computational model of a way in which a whole array of place cells can be used in parallel to calculate long-range distance relations. The model is based on the fact that all place cells are involved in background oscillations at a certain frequency (theta oscillations). The phase of an individual cell's firing relative to the ensemble oscillation (e.g. firing 0.5ms after the peak of the background wave) can encode spatial information. (In experiments when an animal moves along a linear track, this is observed in the form of 'theta phase precession'.) In the model, the phase offset of each cell (0.0 to 0.6, blue-red/grayscale) correlates with the distance of its corresponding location from the origin (bottom left). In the right-hand panel, a barrier (shown in black) increases the length of the shortest route to locations in the middle of the space. This scheme displays holism: the activity of each cell simultaneously stands in the relevant vehicle relation (phase offset) to all the other cells in the array. From Khajeh-Alijani et al. (2015). See the open access online edition of the book for the full colour figure.

chains of activation display holism, albeit on a limited scale: for each location represented, its spatial relations to all other locations represented by other place cells in the chain are also represented (Shea 2022a).

* * *

One kind of representational structure that does not amount to being a structural representation is being an organized representational system (introduced in the last section). (The cognitive map displays organization, but also has the stronger property of being a structural representation.) Organized representations come in families, for example neural signals representing surprise, reward, or numerosity. While only one representation in the family is tokened at a time (in the simplest cases), there are systematic relations between the representations in the family. Lande calls this extrinsic, inter-representational structure (2021, p. 667). As I have argued elsewhere, being an organized representational system is not sufficient for being a structural representation (Shea 2013, p. 127; 2018, 2023c). Briefly, although the representations tokened on different occasions are related, they are never composed together into a combined representational structure.

We can see that clearly with the bee dance. There is a nice systematic relation between number of waggles and distance to nectar (e.g. distance = number of waggles multiplied by a constant). But consumer bees responding to the signal are not acting on a relation between dances. As classically described, they never compare two dances or take the relation between a pair of dances as input. Relations between dances do not enter into downstream computations.

Nevertheless, the systematic relation between different vehicles is a 'good making' feature of an organized representational system. It is error-tolerant and extends non-accidentally to novel cases. Families of related representations may be easier to implement and easier to learn. The relations make it easy for downstream operations to respond systematically to different vehicles in the range.

Furthermore, a collection of different organized representations can interact to perform computations. Different analogue magnitude signals, for example, can interact to perform analogue computations. Computation of visual distance from ocular vergence may work like that (Banks et al. 2016; Shea 2018, pp. 137–8). Two signals correlating with the viewing angle of the two eyes can easily enter into a computational step which generates a third signal that correlates with the distance to the object being fixated. The activity of the third signal is caused by the activity of the first two signals. But the first two signals are not composed into a complex representation, nor is the relation between them acting as a vehicle of content. The vehicle for the content about object distance is the activity of the third signal.

The brain deploys various devices for systematically manipulating patterns of activation, which prove to be computationally useful. For example, one unit may sum the activity of some other units, or accumulate the activity generated by a range of units over time, or perform exponentiation, linear filtering, or divisive normalisation (Carandini and Heeger 2012). Where the so-connected vehicles form an organized representational system, the whole set up is potentially computationally useful. Representational systems that are organized, in this sense, can therefore form the basis of what I have snappily called 'computationally useful processing structure' (Shea 2023c). While this kind of organization is a property worth noting, we should be careful to distinguish it from structural representation.

Organized representational systems are analogue, in one sense of that term.[4] If by 'digital' we mean representations for which the only principle of vehicle individuation is same-different (LeCun, Bengio, and Hinton 2015, p. 441), then organized representational systems are non-digital because vehicles are individuated partly in terms of their interrelations within a family. For example, a firing rate of 10Hz is more similar to a firing rate of 11Hz than it is to a firing rate of 40Hz; and those similarities map on to similarities in content. When organized

[4] This relates to the 'mirroring' conception of analogue representation, variously developed by Maley (2011, 2023); Beck (2018); and Lee, Myers, and Rabin (2023).

representational systems are connected together in a computationally useful processing structure we have a case of analogue computation, in one sense of that term (cp. Peacocke 2019).

There are several other features that have been used to draw an analogue/digital distinction. One is that the vehicles should be continuous (Camp 2007; Peacocke 2019, p. 55). That may seem possible if vehicle types are individuated by a continuous quantity like firing rate or inter-spike interval, but it is almost certainly an idealisation since some fine-grained differences between such states may make no difference to downstream processing.[5] Organized representational systems need not be continuous. A related criterion is that the system of representations should be dense, in the sense that between any two representations there is a third, at least down to a level beyond which differences make no difference to downstream processing (Goodman 1968, pp. 160–2; Peacocke 2019, pp. 62–3). An organized representational system need not be dense, but our examples have been (e.g. analogue magnitude representations; semantic dimensions in an activation state space). So our examples count as analogue in several senses. Being organized is a property that captures one aspect of what theorists have meant by analogue representation.

The examples so far have been representations that vary along one dimension, like a neuron's firing rate. An organized representational system can also be organized along several different dimensions at once. The waggle dance has two dimensions, one correlating with distance, the other with direction. Distributed patterns of activation can lie in a many-dimensional space of potential variations. These are found both in the brain and in artificial neural networks. We noted above that directions or vectors in activation space may have semantic significance. For example, perceived faces are represented in such a space. It seems that any face we encounter is located in a quality space that rates it on dimensions like masculine-feminine, trustworthy-untrustworthy, and dominant-submissive (Todorov et al. 2015; Lee and Kuhl 2016). This is an organized representational system. The way it is organized builds-in various properties for free (properties that, in another representational system, would have to be inferred). For example, if a face is masculine then it cannot be feminine. That just drops out of the way those properties are represented, i.e. by a family of vehicles of which only one can be tokened at a time.

* * *

[5] A related criterion is that a causal description of how vehicle processing unfolds should be stated in terms of continuous quantities, even if the vehicles that are tokened do not or cannot realize all those values (cf. Peacocke 2019). Perhaps a better criterion is that the semantic mapping—from vehicle properties to contents—should be continuous in the relevant vehicle property (even if not every instance of that mapping is or could be realized).

Consider the case of representations that consist of patterns of activation in a high dimensional state space (p. 29). State space representations are not, without more, a case of structural representation. However, it is common that where representations are organized in a state space the thinker is able to make comparisons between them. Two different faces can be represented in the same space and the thinker can rate how similar they are in their dominance. If that judgement is made by relying directly on distance between points in state space, without forming a downstream representation of dominance, then relations between the points are plausibly representing dominance relations and this would be a case of structural representation.

A second property to distinguish from structural representation is illustrated by some iconic representations. While some iconic representations qualify as structural representations (e.g. many maps), others, like a collection of colour chips representing colours, are mere organized representations. Interestingly, some icons have semantically-relevant structure without qualifying as structural representations. Consider computer icons, the graphical symbols we click on to operate computer software. Arguably a graphic icon has its content partly in virtue of a correspondence between its structure and the structure of the entity it represents. That may be why an icon that looks like an antiquated cardboard folder from a filing cabinet represents a directory in the computer's file structure. Convention is clearly also at work, but focusing on the contribution made by structure, that does not obviously require that the parts of the icon are also representations in their own right. Many users may have no idea why the folder icon looks the way it does. They just recognise the picture and know its conventional meaning. Nevertheless, the structure is relevant to fixing the icon's content. What makes this interesting is that, within the representational system for interacting with the computer, the parts of the icon do not compose to form its semantic value. They make no semantic contribution to the semantic value of the icon.

The case of computer icons is complicated by the fact that their conventional meaning trades on what they look like. A potentially cleaner example is the use of a set of arbitrary tokens to represent number. A set of n arbitrary tokens can be used to check numerosity by seeing if a collection of individuated objects maps one-to-one to the set. Plausibly, such a set, of three arbitrary tokens, say, could be used to represent *threeness* (the number property exemplified by sets of three objects). Susan Carey argues that mental representations of this form are used to represent the numerosity of small sets of objects, up to a set size of four objects (Carey 2009). For example, a canonical collection C of three object files can act as a representation of threeness. The system checks whether a set of objects it encounters has three members by seeing whether it can be put into one-to-one correspondence with C. That is a reliable way to track threeness. C as a whole is acting as a representation of a property (the property of having the cardinality *three*)—it is not a compositional representation (e.g. conjunctively representing 1+1+1). The three elements of C need not, in principle, represent anything at all, for the one-to-one mapping from the elements of C to sets of three objects to be part of what makes it the case that C represents *threeness*. At a later developmental stage, the count words, learnt by rote and at that stage no more than meaningless symbols for the child, become a means for expanding the capacity of the system beyond four items (Carey 2009). The n internal entities that together represent the number n, when they are mere uninterpreted words, may have no representational contents individually.[6] The internal set is a representational vehicle, representing what it does in part because it is composed of n individuated elements, but its so-representing does not require that the components are representational constituents. The structured vehicle represents as it does in part because of its structure, but this is not a case of structural representation.

The examples in this section serve to outline the contours of one important category of representational structure, structural representation. In the next section we turn to the kind of representational structure exhibited by natural language sentences and by conscious deliberate thoughts composed out of concepts.

2.4 General-Purpose Compositional Structure

The aim of this section is to characterise a second important category of representational structure: the kind of structure exemplified by natural language sentences. The parade case of language-like compositionality is predication: 'Aisha loves Mildred'. The component terms are incomplete and play different semantic roles. The meaning of the whole is structure-dependent: 'man bites dog' means something different from 'dog bites man'. To contrast with the last section, a central question will be whether language-like compositionality differs from the kind of compositionality at work in a structural representation or, if not, what distinguishes it as a special kind of structural representation.

Language-like compositional structure is plausibly exhibited by the conceptual representations which are tokened in episodes of conscious deliberation (occurrent beliefs, desires, hypotheses, intentions, etc.). The compositional principle at work is general-purpose, in the sense that any concept can be combined in thought with any other concept or concepts of the right kind (there may be adicity restrictions). This is why concepts meet Gareth Evans's generality constraint (Evans 1982).[7] It forms the centrepiece of Jerry Fodor's account of what it is to think in a language of thought (Fodor 1975; Fodor and Pylyshyn 1988). There is good evidence that people use mental representations exhibiting language-like compositionality for a range of tasks, for example: following and remembering geometric patterns of movement (Amalric et al. 2017), thinking about shapes

[6] In legalese, they are jointly but not severally responsible for carrying content.

[7] For Evans, a constraint on what it takes to be a concept; for me, an empirical fact about the type of representation to which I pointed when introducing the term (§1.4).

(Sablé-Meyer et al. 2022), learning new concepts (Piantadosi and Jacobs 2016), combining thoughts in new ways (Frankland and Greene 2020); as well, of course, as understanding natural language (Pietroski 2018).

Having language-like compositional structure corresponds to one understanding of what it takes to be a *propositional* representation. There is a thin sense of 'propositional' which just means that a representation has a content which is assessable for truth and falsity. A stronger sense requires a representation to have sentential structure, with object-predicate structure being the paradigm. The conceptual representations entertained in conscious deliberation are propositional in this sense. Since I want to allow that concepts may also show up as constituents of representations of other kinds (§1.7), for example as labels in a cognitive map, not every representation involving concepts need be propositional in this sense. This stronger sense of propositional goes along with being the type of representation that supports logical reasoning, specifically reasoning involving conjunction, negation, and disjunction. (Strictly, this is a further condition, requiring terms for conjunction and negation and/or disjunction, and the capacity to use language-like compositional principles to combine those terms with conceptual representations.) We will see in the next chapter that there is an explanatory connection between having language-like compositional structure and supporting logical reasoning involving conjunction, disjunction, and negation (although no necessary connection between the two).

Some argue that the language-like compositionality of conceptual thought derives, in all cases, from the language faculty (Chomsky 2017). As against this, there is good evidence that performance of many non-linguistic tasks which seem to involve conceptual reasoning, like solving arithmetical and logic puzzles, can survive the near complete loss of linguistic abilities (Fedorenko and Varley 2016). While this is an important question (Dehaene et al. 2022), it need not detain us here. The distinctively language-like way of composing concepts in deliberate thought might depend constitutively on a capacity that is fundamentally part of the language capacity; alternatively, it may be entirely independent of the language faculty; or there may be instances of each. My account just turns on concepts being able to enter into general-purpose compositional structures meeting something like Evans's generality constraint.

A related question is whether there really are any concepts. Maybe we just have words, plus representations of other kinds—kinds that do not qualify as conceptual. Edouard Machery has mounted a prominent challenge of this type (Machery 2009). Machery argues that psychologists take a concept to be something more than just a constituent of thought. A concept has to be a body of knowledge about a thing X that is used by default in categorisation and inference. Since there is evidence that, in thinking about any given X, people will sometimes deploy a prototype, sometimes rely on exemplars, and sometimes draw on theoretical knowledge and information of other kinds, there are no concepts in this sense.

My project calls for no such restriction (Vicente and Martínez Manrique 2016). I allow that a concept—which may or may not be a mental use of a linguistic token—will give the thinker access to a rich body of information, of many different kinds, with different types of information used on different occasions (Weiskopf 2009a). Nor need the information connected to a concept of X be the same, as between different thinkers (Millikan 2017).[8] It is this plurality that raises the problem of explaining how these different kinds of representational structure (this chapter), informational models (Chapter 4), and computational processes (Chapter 3) can play together in thought (Chapter 5).

While many kinds of representation exhibit forms of compositionality, as we have seen, predicate-argument structure is a special kind of compositionality, often taken to be characteristic of a 'language of thought' in the classic Fodorian sense. Predication is an extremely general-purpose device of concatenation. It allows any predicate to be put together with any singular term or terms (according to the predicate's adicity). Although very general, it is just one of several forms of language-like composition. For example, two nouns can be composed to make a new predicate ('ice sculpture'); and devices of negation can be composed with a predicate to create a new predicate ('non-fiction', 'not green'). I will start with predication and bring in other forms of compositionality later.

To give a proper philosophical analysis of language-like compositionality would be a book-length project in its own right (Collins 2011). There is a large literature, in philosophy and linguistics, on the unity of linguistic meaning or the unity of the proposition. It will be impossible to do justice to those debates here. My aim is more modest: to bring out some of the key characteristics of language-like compositional structure without aiming to reach firm conclusions about the thorniest philosophical issues. As we will see, compositionality in language-like cases works differently from our examples of structural representation. There is a deep issue as to whether it is in fact an instance of structural representation, of a very general kind. I will venture only a tentative view about that. Whichever way that question is resolved, predication differs in an important way from paradigmatic examples of structural representation. The point of this section is to highlight that difference.

I am taking predication to be a relation between representations. A term representing a property is predicated of a term representing a particular. The result is a complex representation with the content that the particular instantiates or exemplifies the property. ('Predication' is also sometimes used to talk about a relation between an object and a property in an abstract object, a proposition. I am setting that aside here.) In first-order predicate logic, writing a predicate followed by

[8] That is why I talk in terms of *a thinker's* concept of X, rather than *the* concept of X.

a singular term thus—Fa—counts as predication. In natural language, the relation between words that amounts to predication is much more complex. It depends on the grammatical structure of the sentence, which is marked by word order and morphology. Words are concatenated in a sentence, and various grammatical (and perhaps also semantic) features determine that predication is at work (or other forms of compositionality).

Something has to play the role of determining the semantic significance of concatenating words into a sentence (or of concatenating concepts into a complete thought). A sentence is more than simply a list of words ('under dogs trees running love'). Similarly, to form a conceptual thought is to do something more than to entertain a collection of concepts (UNDER DOG TREE RUN LOVE). In very many cases, hearers can understand the meaning of a sentence without relying on word order, even for a language like English where word order carries much of the burden of specifying compositional structure. Just hearing 'wants Fido dinner', you can probably work out what I mean. Experiments show that this is true of up to 90 per cent of English sentences. For the remaining 10 per cent, word order is crucial (Mahowald et al. 2022). That is still a lot of sentences—enough that understanding compositional structure is critical to successful communication.

Fodor argues forcefully that semantic constituency in thought is something more than one concept's being tokened and then another (Fodor 2003, pp. 91–4). In fact, he makes a stronger claim. He argues that tokening two representations in succession could not have the significance of predication (Fodor 2003, p. 93). That is too strong. A representational system could be configured, by convention or as a matter of use, so that the mere succession of a predicate by a singular term is the way that predication is marked. Something close to that is true of a standard notation for first order predicate logic, as we saw. (We also saw that succession—of a causal kind—has semantic significance in the place cell system. That was because of the way succession is relied on in calculating routes.) In any case, there would have to be something that makes it the case that succession has semantic significance. Fodor is doubtless right that, as a matter of fact, in the conceptual system, mere succession between concepts does not have the semantic significance of predication. Something else is at work when one concept is predicated of another. Similarly with language: it is the syntactic and morphological features giving the grammar of the sentence, not the mere succession of words, that signifies predication.

* * *

What is the semantic significance of predication? A straightforward view is that predication ascribes a property to an object (King 2009). Frege's account was in terms of function application: a predicate is a function to truth values and predication saturates that function with an object so as to produce a truth value (Pietroski 2016). Fodor argues that predication represents that an object instantiates a property (Fodor 2003, pp. 91–4). In common between these views is the claim that the significance of predication lies in some object or objects instantiating or exemplifying some property (Rescorla 2009; Camp 2018).

Many go further. They take instantiation or exemplification to be a substantive relation between an object and a property and they take predication to represent that relation. As we just saw, Fodor (2003, pp. 91–4) claims that syntactic constituency (at the level of vehicles) expresses the instantiation relation (at the level of reference). That raises the problem of the Bradley regress (Orilia and Paolini Paoletti 2020). If instantiation is a substantive relation that unites object and property, then presumably the object, the property, and the relation of instantiation can all exist without the object and the property being related by the relation of instantiation. So it looks like we need the relation of instantiation to come in again to unite them. And so on up a potentially infinite hierarchy.

One way to forestall the Bradley regress is to deny that instantiation really is a relation. Metaphysically there is just an object having a property. Objects have some properties and that is metaphysically basic. No further relation is needed. That move is coupled with a claim at the level of representations: predication, while being a relation between vehicles, is not a relation that carries semantic content. It is a way of concatenating vehicles which has representational significance—by making a complex representation different from a mere list of terms—but its representational significance does not consist in representing or referring. I will return to this insight in a moment.

Another way to diffuse the Bradley regress is to accept it but argue that it is not vicious. Yes indeed, when Leyla runs, Leyla stands in the instantiation relation to running; Leyla, running, and the instantiation relation stand together in the relation of instantiation; and so on up the hierarchy. However, for all those things to be true just is for it to be true that Leyla runs. Nothing more is needed for all these facts to exist (all at once, as it were). This second view has the advantage of making it easier to understand how we can sometimes refer explicitly to the relation of instantiation. But note that it shares something important with the first answer. In standard cases of predication using ordinary syntactic devices (rather than when referring to instantiation explicitly) there need be no representational constituent or device that serves to stand for or represent instantiation.

I won't attempt to resolve the Bradley regress here, but I will adopt two insights that are common to the solutions we have just seen. First, when F and a are combined by means of a form of syntactic constituency that has the significance of predication, the result is a complex representation with the content Fa. Second, this can be done without there being a representational constituent that expresses instantiation.

The constituency of predication contrasts with the constituency at work in structural representations. In a structural representation, a relation between representational constituents represents a relation between the entities those constituents represent. A relation serves to concatenate two representational constituents: points on a map are related by a relation of spatial separation; place cell activations are related by the time it takes for one to activate the other (or by phase offset, as in Fig. 2.2). The relation serves to combine two representational constituents into a complex representation. It also carries semantic content: it represents the distance between the two corresponding locations. The vehicle property doing the combining is also a carrier of semantic content. But, unlike predication, a given scheme of structural representation is limited in the entities and relations it can pick out and combine. As we saw in the last section, although a given structure could be used to represent many things, once a representational scheme has been established, that limits what can be represented.

The source of the limitation is that, in a structural representation, the relations between representations themselves have semantic values. In both cases the structure has semantic significance, but with a structural representation this is because something stronger is true: the relations that define the structure have semantic values. The content of the whole is constrained by a structural correspondence—when correct, the structure of the representation corresponds with the structure in the world it represents.[9] In the cognitive spatial map, for example, the time it takes for one place cell to activate another represents the distance between their corresponding locations. The structure so-constituted is correct just in case the locations represented by the two place cells are so-separated in space. The relations between vehicles which define the structure have a specific representational significance (spatial proximity). It is a way of combining vehicles so as to represent spatial relations between locations, not a means for predicating properties of locations (Rescorla 2009; Camp 2018). By contrast, predication provides a way of combining any n-ary predicate with any collection of n singular terms. The form of constituency does not place strong constraints on what relations can be represented of the singular terms that are combined into a complete content. The syntactic principle for combining words in a sentence is *neutral*: it has ‘minimal representational import beyond mere combination’ (Camp pers. comm., elaborating on Camp 2007).

But why isn’t predication just a very abstract kind of structural representation? The syntactico-morphological feature that signifies that a given concatenation has the significance of predication (e.g. the word order in ‘Layla runs’) is a relation between vehicles (‘Layla’ and ‘runs’). Why isn’t this a structural representation in which the relevant relation stands for instantiation?

My first answer is that this view raises the problem of the Bradley regress. That is avoided on the view where predication has semantic significance—it makes it the case that the content of the whole representation is Fa—without itself being a

[9] Recall: that may be, but need not be, because the obtaining of a correspondence in certain circumstances is what determines content: Shea (2018, chapter 5).

vehicle carrying representational content. One obvious difference is that there are many different determinate relations (distances) that relate points on a map, each with a different semantic content—each represents a different distance between locations. Predication is not a relation which varies in this way. Even leaving that aside, recall the point that it is not possible to token the representational constituent which represents the relation of spatial proximity without tokening representations of the locations which it relates. Contrast the predicate and singular term, F and a, each of which can be tokened separately. Something further is needed to form a representation whose semantic content is Fa. With points on a map, nothing further is needed to concatenate R, a, and b so as to represent the complete content aRb, beyond tokening the vehicles which represent a, R, and b. No further mode of combination is needed.

My second answer to the challenge is more concessive. Maybe the relation of predication does carry content, referring to the instantiation relation. Even so, there is a clear contrast with paradigmatic examples of structural representation. There, the relations which define the structure have specific representational contents. They refer to relations like spatial proximity, relative exam performance, relative dominance, etc. If predication refers to instantiation, this places very little constraint on what can be represented by the structure. It is an extremely general scheme of concatenation (Camp 2018, p. 25). The simpler view, it seems to me, is that predication is not a case of structural representation, but even if it is, there is still a clear difference in generality. In paradigmatic cases of structural representation, the relation that fixes the structure carries specific representational content; predication, by contrast, is extremely general. Predication is much more semantically neutral than the mode of combination at work in paradigmatic structural representations.[10]

Concepts appear to combine in the same general-purpose way. It has long been thought that the way concepts combine obeys a generality constraint (Evans 1982). The conceptual system allows any two or more concepts to be combined into a coherent thought. The concepts may need to be of the right kind to be so-combined, for example singular term with predicative concept (but see Magidor 2009). But the principles of concatenation are not restrictive as to the subject matter represented, for example constraining the kinds of relations that are allowably combined with two singular concepts. (There will also be constraints of memory, processing capacity, etc. that prevent certain thoughts being entertained; but the generality constraint is a *ceteris paribus* principle capturing a characteristic that flows from the nature of the constituency relations.)

[10] This concessive answer has the merit that it offers a straightforward account of how predication could develop or evolve from something less semantically neutral. A mode of combination that evolved to represent *cause*, say, as in e_1 *caused* e_2, could start to be used for other purposes (gives, loans, owes), with the specific semantic significance fixed by another marker. Greater semantic neutrality could thus evolve gradually through a series of intermediate cases. (Thanks to Liz Camp for this idea.)

A much-celebrated merit of the general-purpose compositionality of concepts is that it enables a very useful form of productivity. It allows us to represent things we have never encountered, considered, or represented before. A limited vocabulary of representational types can be used to represent an extremely large number of states of affairs (perhaps without bound). Any form of compositionality, whether special- or general-purpose, has a combinatorial payoff: the total number of representations that can be formed is a multiplicative function of the number of representational constituents. This means that, when it is not specified in advance which things will need to be represented, there is a benefit in terms of representational efficiency with using a combinatorial system (Frankland, Webb, and Cohen preprint). With general-purpose compositionality, because the mode of combination is not restricted by subject matter, the number of representations that can be formed combinatorially is very large indeed. The capacity to form a very large number of novel representations makes this an efficient way to use representational resources when the nature of the to-be-represented information is not well specified in advance.

To sum up, when we move beyond the simplest representations that lack any semantically-significant structure, then compositionality, of various kinds, is at work. Natural language and conceptual thought are compositional, but so too are structural representations. The meaning of the whole is systematically related to the meanings of the parts and their mode of combination. All of these types of representation display semantically-significant representational structure. Language and conceptual thought are special, though. They make use of a general-purpose mode of combination. To a first approximation, they can put together any two representational elements to make another meaningful representation. Less approximately, ‘any’ = any of the right type (e.g. adicity), and ‘two’ = a small number (cf. locations on a map).

I have focused on predication as the parade case of a general constituency relation, but it is actually just the most prominent example of a family of general constituency relations. Concatenating words or concepts can have other kinds of semantic significance. As we have seen, we can put together two concepts to form a new concept (ICE SCULPTURE); and we can use predicate negation to form new concepts (NON-FICTION, NOT GREEN). Sentences can be concatenated to form a new sentence (e.g. through propositional conjunction: p and q); or concatenated with a sentential negation operator (‘It is not the case that it’s raining’). There is also quantification: the way quantifiers combine with other terms, in language or thought, has semantic significance, but is not a matter of predication.[11]

In typological grammar there is a single operation, ‘Merge’, for putting together any two elements (Moortgat 2010). There is a calculus, NL, that gives a logic for

[11] If predication is taken to express the relation of instantiation, does each of these modes of combination express some (previously unrecognised) relation between the objects or properties referred to by the terms being concatenated?

Merge with no restriction on its interpretation.[12] Merge is something more than bare concatenation. It imposes a hierarchical structure that goes beyond the linear order of words in a string. This allows for the many asymmetries that are essential to language: subject-predicate (importantly), but also antecedent and anaphor, operator and variable, internal argument and external argument, specifier and adjunct, etc. (Collins 2011, p. 106). The semantic significance of joining two items via Merge will usually depend on various syntactic and morphological markers. For example, in 'red is a colour' the 'is' signifies predication (cf. 'red square').

These examples remind us that the semantic significance of the way words are combined into a sentence is not limited to predication. There are a variety of forms of representational combination, each with its own semantic significance. Each has a very general domain of application, not being limited to a specific subject matter. It is a further claim to argue that there is a single mode of combination, Merge, that encompasses all these cases. Conceptual thought may also have a combinatorial device like Merge: a way of concatenating any two representations, whose semantic significance is fixed by further syntactic markers.[13] That would substantiate my claim about the generality of language-like compositional structure in a particularly strong way, since Merge is a completely semantically-neutral device of concatenation.

The existence of Merge in conceptual thought would serve to illustrate the point that the generality of language-like compositionality contrasts with the content-specificity of the compositional principles at work in paradigmatic cases of structural representation. However, I do not need there to be a mental equivalent of Merge to establish my conclusion. Just focusing on one of these forms of combination, like predication, is enough to demonstrate a clear contrast. There is an important difference between the content-specific way representations combine in a structural representation and the content-general way representations combine in language and conceptual thought.

2.5 Terminology

Complete condition: a way the world could be, or a condition against which the world could be assessed, for example a truth condition, correctness condition, accuracy condition, or satisfaction condition.

Complete content: representational content at the level of a complete condition.

[12] Another calculus, LG, distinguishes two Merge operations, both meeting the generality constraint, each having further restrictions.

[13] That would explain the recursive nature of thought, since any representation so formed could be subject to further concatenation (subject to memory and processing constraints).

Computationally-useful processing organization: a processing structure that can be used for performing certain computations; derivatively, the computations for which a processing structure can be so used.

Concatenate: link or join together items in any way (not limited to serial order); thus, predication is a form of concatenation.

Concept (non-definitional: pointing at the phenomenon): freely-recombinable constituent of the thoughts that occur in deliberation.

Conceptual representation: a representation, constructed out of concepts, with a complete content.

Deliberation: thinking in a series of steps and subject to cognitive load.

Incomplete representation: representation that does not represent a complete condition on its own, but only when combined with other representational constituents into a complex representation.

Inference: a content-based transition between representations.

Language-like compositional structure: the type of compositional structure exhibited by natural language sentences, for example involving predication (see §2.4 for details); also exhibited by at least some forms of conceptual thought.

Organized representational system: a collection of representations where the function mapping vehicles to contents applies systematically across all vehicles in the family, mapping similar vehicles to similar contents. (All our cases concern similarity structure, but the phenomenon may generalise to systematic relations of other kinds.)

Paradigmatic structural representation: a structural representation that does not have language-like compositional structure (if that turns out, *pace* the first position argued for in §2.4, to count as a case of structural representation).

Processing structure: the way a collection of vehicles or vehicle families interact causally in internal processing.

Reasoning: step-by-step inferences from one or a few conceptual representations (premise or premises) to another (conclusion).

Representation: content-bearing entity (can be used for the vehicle, or to pick out a content-bearing entity in terms of its content).

Representational structure: shorthand for semantically-significant representational structure.

Semantically-compositional components: narrower than semantically-significant representational structure: component representational vehicles and vehicle properties that have a semantic value.

Semantically-significant representational structure: aspects or components of a representational vehicle that carry or determine semantic content. (So the predicational structure of a sentence counts as semantically-significant representational structure, even though predication does not carry semantic content—it contributes to determining semantic content. The organization of an organized representational system counts as semantically-significant representational structure, when it is part of a computationally useful processing organization that is used, since it contributes to determining semantic content. The structure of an iconic representation whose parts are not representational (e.g. computer icon; set of object files for number) counts as semantically-significant representational structure since it contributes to determining reference.)

Structural representation: a complex representation in which a relation on representational vehicles represents a relation on the entities they represent.

Structure of a representation: any kind of structure of a representation, i.e. aspects or components of a representational vehicle, and their relations.

Thinking: the unfolding of a sequence of thoughts, and the executive processes that apply to them.

Thought: an inclusive term for any kind of representation figuring in the cognitive playground, whether conceptual or not-conceptually-compositional, including structural representations and organized representations.

Vehicle of content: entity, property, or relation bearing content. Used narrowly for the particular (e.g. word, pattern of activation), more broadly to include properties that carry content, i.e. properties to which a semantic value attaches ('vehicle properties'). Most often used in respect of a particular picked out in terms of content-bearing properties: for example, when marks on the page are picked out as words, that picks out vehicles in terms of a type to which a semantic value attaches.

Chapter Summary

2.1 What is Semantically-Significant Representational Structure?

This chapter is about the different kinds of representational structure exemplified by the representations involved in concept-driven thinking. These form the basis of different computational processes (Chapter 3) and informational models (Chapter 4), and fit together in my account of concept-driven thinking (Chapter 5).

All representations have physical structure; some have no representational structure. (p. 28)[14] Not every way of dividing a representational vehicle into parts produces semantically-significant units. I use 'representational structure' (unmodified) to mean semantically-significant representational structure, which

[14] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

is what concerns us here. 'Structural representation'—where there is a structural correspondence between representation and world—is just one kind of representational structure. In maps and sentences, representational structure is part-whole structure, with the meaning of the whole systematically related to the meaning of the parts. (p. 29) Representations can have semantically-significant structure that is not a matter of part-whole structure, for example component activation vectors that superpose into a distributed pattern of neural activation. 'Vehicles' are the entities that carry content; 'vehicle properties' are non-semantic properties that are or may be relevant to the way representations are combined and processed. Semantically-significant representational structure is individuated functionally. Vehicles of mental content are found in the brain; while they may be highly abstract, they must figure in psychological processes and competencies.

(p. 30) The structure of a representation is semantically significant to the extent that vehicle properties and relations have semantic import; that is, carry semantic values or make a difference to its content. The cognitive sciences have developed a variety of techniques for investigating representational structure. Representational structure determines 'distributional properties': which representations can, cannot, and must co-occur. Rather than a dichotomy (like iconic vs. discursive), I will work with six different aspects of representational structure (all exhibited by natural language sentences). (p. 31) List (a)–(f). These six features are likewise exhibited by conceptual representations; also piecemeal by representations of other kinds (so using a conceptual/nonconceptual dichotomy can be misleading).

2.2 Varieties of Semantically-Significant Representational Structure

This section surveys a variety of forms of representational structure exemplified by mental representations of different kinds. (p. 32) The base case is *nominal signs*, where there is no representational structure (e.g. vervet alarm calls). In Paul Revere's lantern signalling system, neither the physical structure of the representation (e.g. two lanterns), nor the relations between representations (e.g. brighter than), have semantic significance. With mental representations, even simple signals are rarely nominal—often the magnitude of the signal will track some quantity represented. These are not structural representations, but the different representations tokened at different times form a family that displays representational 'organization'.

(p. 33) Next, a representation can have more than one semantically-significant dimension of variation; these need not be 'bound' to one another—they can have semantic significance independently. Mental representations can work in the same way, for example with colour information and motion information carried by independent dimensions in dynamic neural activation space, neither tokenable without the other. Contrast representations in early visual processing, where colour and motion are represented separately, each tokenable independently.

More is needed for *binding*: combining two vehicles in a representational structure such that two attributes are represented as instantiated together. (p. 34) In this case, the representational constituents are 'incomplete': neither makes a claim about the world on its own, but only when combined. Whatever way concatenation of visual features works, it contrasts with cases where two complete conditions are represented separately, by different aspects of a complex representational vehicle. (p. 35) Incomplete components may or may not be separately tokenable.

A central example is representation of locations and their spatial relations by place cells in the hippocampus. Co-activation of place cells represents that the corresponding locations are near to one another spatially. (p. 36) The way place cells are used to calculate routes supports the conclusion that the activation of a place cell represents a location and the co-activation of two place cells represents spatial proximity. The vehicle property representing the spatial relation (namely, co-activation) cannot be tokened without tokening the incomplete constituents between which that relation obtains. (p. 37) The way place cell activation represents spatial relations exemplifies features (a), (c), (d), and (e) of representational structure, but not (b) (separate tokenability). Nor (f): the cognitive map, and visual feature binding, both deploy a special-purpose device of concatenation, unlike the general-purpose concatenation at work in natural language.

This brief survey shows that the features that are exemplified together in natural language sentences and canonical cases of conceptual representation are also found at work, piecemeal, in mental representations of other kinds.

2.3 Structural Representation

This section is about structural representation, the kind of structure exhibited by maps; the next is about the kind of structure exhibited by sentences. (p. 38) A structural representation makes use of a structural correspondence between representation and world. Definition: *structural representation.* The clearest examples of structural representation are public external representations, for example using spatial arrangement or temporal sequence to represent relations in the world. When colour is used to represent property values (e.g. temperature), colour relations can represent worldly relations. The definition of structural representation does not require that the correspondence between relations is doing the content-fixing. (p. 39) The representational significance of the semantically-significant relations within a given system of structural representation is limited. Structural representations also often display holism.

(p. 40) A system that displays representational organization need not amount to being a structural representation. (p. 41) Organization is useful for a number of reasons: error-tolerance, extension to novel cases, ease of implementation. Furthermore, and less-recognised: organized vehicles can readily interact so as to perform useful computations. The brain deploys various devices for systematically manipulating patterns of activation, which are computationally useful: they form a computationally useful processing structure. Organized representations are analogue in the sense that vehicles are individuated partly in terms of their similarity to other vehicles; they can enter into analogue computations. (p. 42) Organization differs from other properties that have been used to draw an analogue/digital distinction. An organized representational system builds in for free the fact that different determinates of the same determinable property (e.g. trustworthiness) exclude one another.

(p. 43) Representations in a state space are organized and not automatically structural, but when two representations are tokened in the same state space, the relation between them can carry representational content (e.g. relative dominance). The structure of a representation can play a role in fixing content without the parts each having their own semantic significance. For example, a set of three arbitrary tokens can be used to represent a numerical property: having the cardinality *three.*

(p. 44) The examples in this section have served to outline the contours of one important category of representational structure, structural representation, to contrast with general-purpose compositional structure in the next section.

2.4 General-Purpose Compositional Structure

This section characterises a second important category of representational structure: that exemplified by natural language sentences. There are mental representations with language-like compositional structure, exhibiting general-purpose compositionality. (p. 45) Having language-like compositional structure is what makes a conceptual representation propositional (in one sense), and supports logical reasoning. My account can remain neutral on whether conceptual compositionality is underpinned by linguistic compositionality, or the converse, or whether they are independent. My account does not require that a concept consists of or gives access to a default body of information (cf. Machery 2009).

(p. 46) Predication is the most prominent, but not the only, form of language-like compositionality. Predication differs in an important way from paradigmatic examples of structural representation; tentatively, it is not a kind of structural representation at all. Predication is a relation between representational constituents, a way of concatenating representations. (p. 47) A sentence is more than a list of words—its compositional structure is crucial (and required for understanding a significant proportion of sentences). Mere succession of one concept by another is not enough to predicate one of the other—concepts are combined by some device that has the semantic significance of predication.

In common between theories of the semantic significance of predication is that the result involves an object or objects instantiating or exemplifying a property. (p. 48) Some theorists claim, further, that the predication relation represents instantiation, potentially launching an infinite hierarchy of instantiation (the Bradley regress). Some deny that instantiation really is a relation; coupling this with the view that the vehicle-based relation of predication does not itself refer or represent. Or we can accept the whole infinite hierarchy; and still deny that there is a device which serves to stand for or represent the relation of instantiation. Without resolving the whole issue, I endorse the view that combing F and a by the vehicle relation of predication has the effect that the content Fa is represented; and that doing so does not require there to be a representational constituent that expresses instantiation.

The constituency of predication contrasts with the constituency at work in structural representations. (p. 49) Predication is semantically neutral: it places no strong constraints on what relations can be represented of the entities referred to. Objection: why isn't it a structural representation, one in which the relation of predication represents instantiation? Answer: to avoid the Bradley regress. (p. 50) But for present purposes it is enough that there is a clear difference in generality between the general-purpose compositionality of predication and the compositional principles at work in paradigmatic structural representations. Concepts combine in the same general-purpose way. (p. 51) The general-purpose compositionality of concepts underpins a powerful form of productivity; which is an efficient way to use representational resources when the nature of the to-be-represented information is not readily specified in advance.

To sum up, all but the simplest representations display forms of compositionality; language and conceptual thought are special in making use of a general-purpose mode of combination. There are other general-purpose constituency relations, for example predicate negation and nominal compounding. There may be a single, most-general concatenation operation, Merge. (p. 52) There may instead be a variety of different forms of general-purpose concatenation (in language and conceptual thought). Just focusing on predication is enough to establish a contrast between the content-specific way representations combine in a structural representation and the content-general way representations combine in language and thought.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0002

3
Computational Processes

3.1 Transitions Faithful to Content

This chapter is about the computational processes that mental representations enter into. The foundational claim of the representational theory of mind (RTM) is that representations are physical particulars that undergo causal processes in the service of generating appropriate behavioural outputs—behaviour that is appropriate to the current situation. RTM's insight is that transitions between representations can be configured in such a way that the transitions are faithful to representational contents. That can in fact be achieved in two quite different ways, as I will argue. Some transitions are 'content-specific', others 'content-general' (or 'non-content-specific'). This chapter will define that distinction and set out the characteristics of these two different kinds of computational process.

The content-specific/content-general distinction is based on two ways of implementing the fundamental insight of representationalism. Machines were originally designed to be physically useful, to weave cloth, say. The big discovery underpinning the advent of information technology was that a machine could be designed to do something intellectually useful. A mechanical process can be configured so as to multiply two large numbers together, for example. Charles Babbage's difference engine was designed to achieve that using cogs and wheels.[1] A digital computer does it using currents in semiconductors. The basic principle is the same in both cases. There are physical particulars with content, representations of numbers, which interact causally in ways that are faithful to their contents.

[1] Babbage and Ada Lovelace built on this to develop plans for a general-purpose mechanical computing device (the analytical engine).

In the difference engine, transitions between representations are faithful to content in the sense that, given two numbers at input, the output will represent their product. The most straightforward way for a transition to be faithful to content is for it to be truth preserving: if the inputs are true then the outputs will be true. Logically valid inferences, for instance, are truth preserving; necessarily so. To use the standard example:

(1) All humans are mortal
(2) Socrates is human
∴ (3) Socrates is mortal

A digital computer is built on ramifying schemes of valid inferences. It deploys logic gates to make truth-preserving transitions. Operations specified in a programming language are compiled so that they can be carried out by complex arrangements of these basic components. That is how a digital computer can multiply two very large numbers together. Operations are faithful to content when they correctly perform the functions they are designed to implement. A transition that is designed to take the diameter of a circle as input and output the circumference will be faithful to content if the content represented by the output is π times that of the input.

There are many different ways that a transition can be faithful to content. Computations can involve both descriptive and directive representations. Given a represented goal O (directive) and a descriptive representation of how to produce it (*doing A causes O*), the computation outputs an intention (*do A*). That transition is faithful to content provided the intended action is likely to bring about the goal if the descriptive premise is true. For present purposes we can focus on systems that are engaged in working out what is the case, broadly speaking. So only descriptive representations will be in play. Necessary truth preservation, as in the case of a deductively valid inference, is a strong form of faithfulness. Transitions that sometimes go wrong but mostly go right are also likely to be useful. For example, we may be disposed to draw categorical conclusions from probabilistic information:

(4) p is more than 95 per cent probable
(5) p

Being disposed to make inferences of this form, the thinker won't go wrong very often, provided the premise is accurate. (We can see how often.) This is the form of faithfulness that will concern us most: transitions where the output is likely to be correct if the inputs are correct.

Not every transition that occurs need be faithful to content. A system can misrepresent and a transition can misfire. Furthermore, a system can have a disposition to make a transition which is systematically mistaken. People may, for instance, affirm the consequent in situations where that leads to false conclusions. In a different kind of case, the inferential dispositions encapsulated in the visuomotor system can be mis-trained, for example by wearing prism goggles (Redding and Wallace 1997). Afterwards the transitions it is disposed to make go wrong, outputting conclusions that are false in normal environments. Furthermore, some transitions are not held to this standard at all, such as the way some thoughts cause others when the thinker is free-associating. So this notion of faithfulness to content does not require that every transition between representations should be faithful to content.

It is a nice question just how much faithfulness is required. Of the transitions that count, enough of them must be faithful to content so that RTM can get a grip—so that the operations of the system and its behaviour in the world are explicable in terms of representational content (Shea 2016). How much is that? Fortunately, we don't need an answer here. For our purposes what matters is that the faithfulness of a transition turns on the contents of the representations involved. For a transition where faithfulness is a matter of the conclusion's being likely to be true if the premises are true, we ask which contents underpin truth transmission. The distinction I will draw between content-specific and content-general transitions turns on which contents the faithfulness-to-content of a transition depends on.

In many of the computational processes studied by cognitive neuroscience, the transitions involved are nothing like the necessarily truth-preserving inferences of formal logic. Internal processes are wired up by experience to work well enough in the context of the organism's normal environment, but the outputs are only accurate often enough to be useful, and they only get things right, to the extent that they do, by trading on presuppositions about statistical regularities in the environment. The visual system, for example, may be set up to make transitions from a contrast map, to representing line segments, to an edge map (Wolfe et al. 2018). These dispositions effectively presuppose that certain arrangements of contrast are a reliable sign of the location of edges—which they are, in normal environments. Deep neural networks (DNNs) work in the same way. Through a vast amount of training, transitions between distributed representations in hidden layers come to reflect statistical regularities in the training environment. So DNNs too learn to make transitions from distributions of contrast to representations of edges, shapes, and patterns (Güçlü and van Gerven 2015) (Fig. 3.1). Indeed, there are interesting similarities between the way processing unfolds across the layers of a deep neural network and in the human visual system (Yamins et al. 2014; Cichy et al. 2016; Cao and Yamins 2021).

The exact form of these transitions is not yet well understood, either in the human visual system or in artificial neural networks. Complexity presents a challenge in both cases, with additional difficulties in the brain with collecting

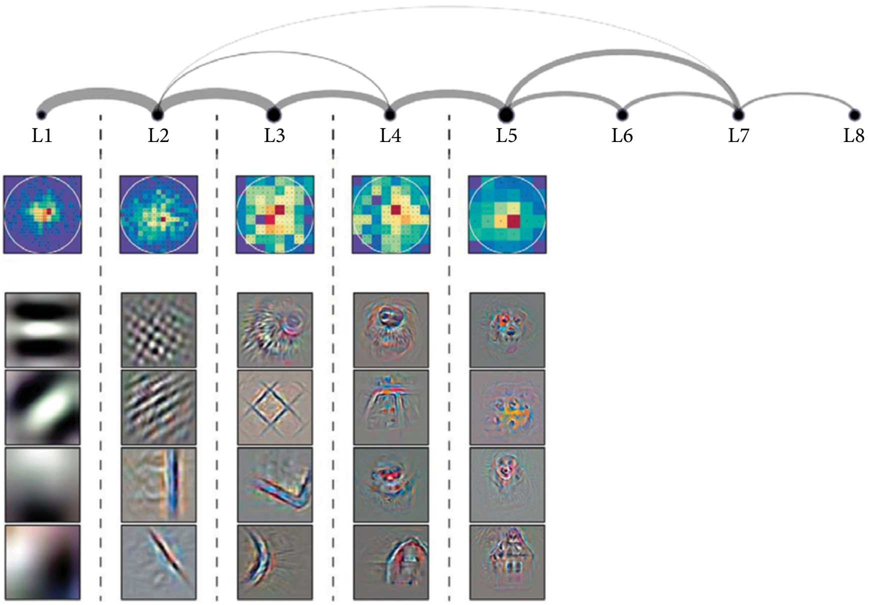


Fig. 3.1 Transitions between representations in a deep neural network trained to perform image classification. The network has learnt transitions from contrast patterns to edges and textures, and then to parts of objects. From Güçlü and van Gerven (2015). See the open access online edition of the book for the full colour figure.

detailed data about how processing unfolds. Preliminary work in DNN models suggests that some of the learnt transitions are highly specific to the subject matter on which the system was trained. For example, in InceptionV1, a DNN trained to categorise visual images, there is a transition from detecting left-oriented and right-oriented fur (in layer 3b), to detecting left-oriented and right-oriented dog heads (layer 4a), to detecting orientation-invariant dog heads (layer 4b) (Olah et al. 2020). This analysis focused on contents carried by individual units. It is likely that many computations take place over distributed patterns of activation (in DNNs and in the brain). The key transitions may, in particular, take place between components of activation vectors, or between dimensions of the neural manifolds that capture the activity of a population of neurons in the brain (Langdon et al. 2023). (Operations can occur independently to the extent that components are orthogonal.) For example, Nanda et al. (2023) were able to discover the internal transitions that a simple neural network had learned in order to perform modular addition. If these analyses give us any indication of what is being represented by distributed patterns of activation, they suggest that the transitions the system makes, en route to achieving its trained outputs, are highly tailored to the subject matter on which it was trained.

In many systems which are described as performing analogue computations, representations interact to carry out a computation, and the interactions are faithful to content only because of the specific contents involved. For example, desert ants are able to navigate by path integration, maintaining a representation of the distance and direction back to their nest even as they follow a tortuous, winding outbound route. There is evidence that they achieve this by keeping track of their angle of travel at each step and carrying out a simple multiplicative calculation (Mueller and Wehner 1988). This approximation is not perfect, but it works well enough, given the way the ants behave. The transitions between internal states are faithful to content enough of the time to be useful. The contents to which they are faithful concern angles and distances. Faithfulness to content turns on the components representing these quantities.

Analogue computations can be more abstract, for example representing number. Children can perform addition, subtraction, and multiplication on the set size of large arrays of objects (Barth et al. 2006; McCrink and Spelke 2010). It seems likely that this is achieved through the interaction of analogue signals in the brain tracking numerosity (Nieder 2016). These interactions are suited to many different tasks, since doing (approximate) arithmetic is useful in many different circumstances. They exhibit considerable generality, taking them closer towards the content-general end of the spectrum. Nevertheless, the faithfulness-to-content of these transitions depends on the fact that it is quantity that is being represented (number, numerosity, or some such). They would not be even approximately correctness-preserving for any arbitrary contents.

Computational models in cognitive neuroscience often have this characteristic. Here are some examples. Keeping track of eye position by integrating inputs encoding velocity (Shagrir 2012). Integrating instantaneous direction of motion in order to calculate the overall direction of motion of a stimulus and to program a corresponding saccade (Beck et al. 2008; Shea 2014b). Anticipating the trajectory of a limb in response to a motor command, and updating based on feedback (Miall and Wolpert 1996, pp. 183–5; Shea 2018). Calculating distance from ocular vergence (§2.3) (Banks et al. 2016). Our go-to example of the spatial cognitive map also has this characteristic. There are internal vehicles (place cell activity and co-activation) that interact in performing offline route calculation. Picking the shortest internal chain of co-activation is a way of calculating the shortest route, it is a content-faithful transformation, but this faithfulness turns on the fact that place cells represent locations and co-activation represents spatial proximity. It is an elegant computational mechanism, but its faithfulness to content depends on those specific contents.

The cognitive map has a further characteristic. It is a structural representation, as we have seen. The faithfulness to content of the way representations are processed turns on how the structure represents and is used. I will return to this point shortly (§3.3).

All these examples contrast with logical inference. A logical inference's being faithful to content does not turn on the subject matter. It takes a form such that the conclusion is bound to be true if the premises are true, irrespective of the specific subject matter involved. The next section will pin down this contrast.

3.2 Content-Specific and Content-General Transitions

So far we have seen examples of two different ways in which transitions between representations can be configured so as to be faithful to content. First there are broadly-logical transitions. Speaking very roughly, these transitions are faithful to content whatever contents are involved. Second, there are transitions that operate within a particular domain, working well enough to be useful, but only because of the specific contents being represented. The impetus for RTM was the realisation that a mechanical process could be set up to make transitions of the first kind. In fact, however, when it comes to mental processes, transitions of the second kind may be more common. We saw examples ranging from visual processing, through analogue magnitude arithmetic and probabilistic information integration, to motor control.

This second category generalises Wilfred Sellars's idea that, in language, there are material inferences as well as logical inferences (Sellars 1953). The inference from 'it is raining' to 'the streets will be wet' is not simply a disguised logical inference with a suppressed general premise. It is its own type of inference, widely used in practical reasoning. The aim of this section is to develop this distinction in a way which is applicable, beyond language, to a broad range of cases from the cognitive sciences.[2] My more inclusive use of the term 'inference' covers any kind of computational or content-based transition. It applies to representations of any kind: nominal representations, mere organized representations, structural representations, and representations displaying general-purpose compositionality. My aim is to endorse Sellars's insight and show that it can be given general application.

Although content-specific transitions are the more basic case, it will be convenient to start with content-general transitions. Experimental work finds that deductive reasoning operates differently from other forms of inference, like inductive reasoning, that depend on world-knowledge and context (Heit and Rotello 2010). We are disposed to make some logical transitions automatically, just in virtue of representing the relevant premises, irrespective of the plausibility of the conclusion (Ball, Thompson, and Stupple 2018; De Neys 2023). For example, Reverberi et al. (2012) found that modus ponens transitions are made

[2] Nor need the distinction be tied to a use theory of meaning.

automatically, but not modus tollens. I have said that logical inferences are truth preserving no matter what contents are involved, but that is too quick. The validity of an inference does of course depend on the contents of the logical terms. A content-general transition is one whose faithfulness to content only depends on the content of the logical terms, not on the content of the non-logical terms. Consider some examples:

(6) p and q

∴ (7) p

(8) For all x, if Fx then Gx

(9) Fa

∴ (10) Ga

Inferences of these forms are deductively valid. Other examples, while still being formal, are not deductively valid. The conclusion is likely to be true if the premises are true, given the form of the transition. We saw that with (4) to (5): moving from high probability to a categorical conclusion. Another example is:

(11) All observed Ks have been P

(12) Ka

∴ (13) Pa

This pattern works well enough in most ordinary applications, but it famously won’t work for arbitrary predicates K and P (Goodman 1955, p. 74; Quine 1969).

These forms of transitions are all characterised by their generality. The examples involving ‘and’ and ‘if…then’ are completely general as to their subject matter. (6) to (7) is faithful to content whatever complete contents are substituted for p and q. (8) to (10) is faithful to content whatever singular term is substituted for a, and whatever predicates are substituted for F and G. (11) to (13) is somewhat less general; nevertheless, it is faithful to content for very many of the predicates in ordinary use. There is probably no sharp cut-off between content-general patterns of inference and the rest. The difference is a matter of degree, marked by how wide-ranging the class of representations is over which the disposition operates and is faithful to content.

In all of these cases the inference does depend on the content of the logical terms. What works for ‘all’ would not work for ‘some’. In our examples faithfulness to content depends on: ‘all’, ‘and’, ‘if…then’, ‘% probable’, and ‘observed’. I will call these ‘broadly-logical’ terms: terms on whose content the validity or truth-conduciveness of a generally-applicable transition pattern depends. Broadly-logical terms contrast with open class terms like ‘Socrates’ and ‘mortal’.

Content-general inferences involve terms with specific contents, but their being faithful to content does not turn on the content of these terms. Think about the Socrates inference again, (1)–(3). It would be meaningless without using terms that refer to properties and particulars (e.g. Socrates). *A fortiori*, it would not be truth preserving without them. But it is content-general, not content-specific. Substitute other concepts for SOCRATES, HUMAN, and MORTAL and you get another truth-preserving inference. This transition's being faithful to content turns only on the contents of the broadly-logical concepts involved. By contrast, the faithfulness to content of content-specific transitions does depend on the content of the non-logical terms (of the non-broadly-logical-representations). So we can define the distinction as follows:

Content-specific transition: a transition between representations such that whether or not the transition is faithful to content depends on the content of representations other than broadly-logical terms.

Content-general transition/non-content-specific transition: a transition between representations such that whether or not the transition is faithful to content depends at most on the content of broadly-logical terms.

Content-general transitions do depend on content, but their faithfulness to content turns only on the content of broadly-logical terms (if any[3]). Something stronger may also be true. On certain plausible theories of content, the content of broadly-logical concepts is determined by the role they play in patterns of inference. For example, the content of AND is plausibly fixed by its role in inference patterns like (6)–(7) (Peacocke 1992). By contrast, the content of concepts like SOCRATES and MORTAL, which figure in the variable positions in these schemata, depends on relations between the thinker and their environment. If that is correct, then the faithfulness to content of broadly-logical transitions only turns on their form, which fixes the content of the concepts on whose meaning faithfulness-to-content depends. In any event, content-specific transitions turn on their particular subject matter: contrast and edges; dogs and fur; locations and proximity; etc. The sense in which content-general transitions are non-content-specific is that their faithfulness to content does not depend on the content of any non-broadly-logical terms. If these are indeed different kinds of transitions then we should expect to find functional differences between the way people identify logical contradictions (*x is red and not red*) and semantic contradictions (*x is red all over and blue all over*). That is an empirical prediction of my account.

My way of defining broadly-logical terms relates to one of the ways that logicians have attempted to define the logical constants. The idea there is that an

[3] Trivially, the inference from p to p does not depend on the content of any non-broadly-logical terms, but it is valid and content-general. There are also truth-conducive ways of introducing a conclusion *de novo*, without a preceding premise, e.g. 'Either it's raining or it's not raining'.

operation is logical just in case it is invariant under permutations of the objects in the domain of interpretation (Bonnay 2008). This is to characterise logical notions in terms of their generality (Bonnay 2008, p. 33) and formality (p. 34). There are various objections to this definition, for example it implies that 'most' counts as a logical constant. Fortunately, for our purposes we don't need a resolution of the issue in formal logic. Logic is concerned with the nature of various abstract structures. We are after a distinction between different kinds of transition, transitions that actually occur between representations in biological brains and computing machines. Our distinction can be a matter of degree and accommodate intermediate cases. Terms like 'most', and indeed other quantifiers, are at the broadly logical end of the distinction. They can underpin content-general computational steps.

Another example of a content-general transition is a disposition to update beliefs through exact or approximate Bayesian inference (Rescorla 2024). Bayesian inference works well no matter what the subject matter is. It is, however, difficult to implement, so real cognitive processes resort to approximations. How well these work may depend on the domain; nevertheless, schemes of approximate Bayesian inference are typically very generally applicable. They will still lie towards the content-general end of our distinction. A more controversial example involves our capacity to detect and punish cheaters. If this depends on a domain-specific module that identifies situations in which a benefit is received, and no reciprocation is made or cost paid (Gigerenzer and Hug 1992), then the transition to social censure would be content-specific. However, the core reasoning may be performed in a content-general way (Oaksford and Chater 1994), supplemented by content-specific biases in salience or attention. Content-generality or content-specificity turns on the particular empirical facts of the case.

Content-general computation is related to the computer science idea of variable binding. It has long been recognised that a key attribute of classical computers is their ability to compute with variables—representations which can be 'bound' to specific values, but which can be computed with without specifying their values (Penn et al. 2008; Gallistel and King 2009; Kriete et al. 2013; Bottou 2014; Graves et al. 2016; Santoro et al. 2017). Content-general computation is a generalisation of this idea (which is already very general). Our canonical example of a content-general transition, the Socrates inference (1)–(3), does not use variables. It takes place with a particular singular term and particular predicates. Nevertheless, its truth conduciveness does not turn on the nature of these specific contents. It is content-general, not content-specific.

The patterns of inference at (6)–(7), (8)–(10), and (11)–(13) were specified in terms of variables. One way to implement a disposition would be to store these schema using variables and then to match them to current inputs. If so, a particular instance like our Socrates inference might be performed using bound variables. But a disposition can conform to a pattern without the system having to represent the pattern explicitly. In that case a variable-involving schema like (8)–(10) would accurately capture a disposition without the system needing to have access to variables. The schema would describe patterns in what the system does with specific, non-variable representations.

A system with access to variables might perform these inferences without calling up the specific values of the variables. We can see that clearly in an inference that is performed over free variables:

$$\begin{array}{lll} & (14) & (x - 1)^2 \equiv x^2 - 2x + 1 \\ \therefore & (15) & 2x \equiv x^2 + 1 - (x - 1)^2 \end{array}$$

So variables can underpin content-general transitions. On the other hand, variables might also occur in content-specific transitions. A domain-specific cheater detection module might store an inference scheme encoding variables with a restricted range of application (people and social exchanges, say). Nor does content-generality require the use of variables, as we have seen. Content-generality is the more inclusive category.

* * *

There is a notion of computation where only content-general transitions count as computational. If computations have to be defined wholly in virtue of syntax, and to be neutral about what subject matter is being represented (or whether any subject matter is being represented), then only broadly-logical inferences will qualify as computational. (Even more narrowly, the capacity for variable binding might be thought to be required for computation.) That is a possible terminological choice. However, content-specific transitions are an equally good way of meeting the core RTM commitment to transitions which are faithful to content. Many putative analogue computations fit in this category, as do the examples of content-specific transitions I have drawn from computational cognitive neuroscience. So it is reasonable to think of both types of transition as computational.

In the computers invented in the twentieth century, content-general transitions have been to the fore. The computers in use in most practical applications are built on ramifying schemes of broadly-logical inference. Where they have built-in mathematical computations, these are somewhat content-specific, since they concern numerical quantities, but they lie towards the content-general end of the spectrum. The earliest computing machines were somewhat special-purpose, and there have been experiments in analogue computing machines deploying content-specific transitions, but it is only with the advent of powerful DNNs in the last decade that practical applications place widespread reliance on content-specific transitions, ones which are learnt incrementally from a large training set (as in the image classification example above). However, phylogenetically, it seems most likely that the order is reversed. Content-specific transitions seem to be the basic case, ubiquitous in representational processing in the animal kingdom.

It may be that the capacity for content-general computation is an ontogenetic achievement. Whether this is so turns on tricky empirical issues about whether the capacity for language depends on content-general computations (on which the jury is still out), and on whether developing that capacity depends on learning—whether the capacity for content-general computation is learnt from the environment rather than canalized in development. While phylogenetically it is reasonably clear that humans engage in more widespread and systematic logical reasoning than other animals, ontogenetically the issue is less clear-cut. On the one hand, Catarina Dutilh Novaes has recently made an impressive case that the capacity for deductive reasoning is socially learnt (Dutilh Novaes 2020), having evolved through cultural evolution as a 'cognitive gadget' (Heyes 2018). On the other hand, the cross-cultural evidence for a language of thought quoted in Chapter 2 includes tasks that seem to call for relatively content-general computational dispositions (Dehaene et al. 2022), suggesting less variability than might be expected from a cognitive gadget. Either way, we still need to think of the capacity for content-general inference as a cognitive achievement, with content-specific computation the more basic way that representational processing manages to be faithful to content.

Is this a new distinction? It is not new to point out that there are important distinctions in this area. It is widely recognised that computational models divide into significantly different types. Theorists have drawn many different distinctions. There is the difference between classical computational architectures and artificial neural networks (connectionist systems), sometimes characterised as a difference between symbolic and subsymbolic computation. Lake et al. (2017) argue that underlying the symbolic/subsymbolic dichotomy there is in fact a deeper distinction, that between model-building systems and those that just perform pattern recognition. The distinction I am offering, similarly, identifies something fundamental that, amongst other things, does separate paradigmatic classical computational systems from paradigmatic artificial neural networks. But it does not align perfectly: both types of system could in principle perform both content-specific and content-general transitions.

The distinction is not a matter of whether representations do or do not display compositionality. There is a connection with modularity: content-specific transitions feature in familiar modules, as standardly understood. But my distinction is not equivalent to modular vs. non-modular. In one sense I am drawing a new distinction in the vicinity of these existing dichotomies; in another sense I am simply characterising a familiar distinction in a somewhat novel way. My aim is to precisely specify a distinction that is sufficiently deep and general that it can do the explanatory work I need—that I need for the wider project of explaining how various representational structures, informational models, and computational processes play together in the course of concept-driven thinking.

The content-specific/content-general distinction concerns two broad types of computational processing. The last chapter drew a distinction between two broad types of representational structure. The next section explores whether the two are connected.

3.3 Types of Transition Go with Types of Representational Structure

In the last chapter, we saw that there is a contrast between paradigmatic cases of structural representation, on the one hand, and predication in language and thought, on the other (§2.4). The general-purpose compositional principles of language contrast with the special-purpose compositional principles at work in a map (physical or cognitive). In a structural representation, the relation by which components are combined into a complex representation has specific representational content. In the spatial cognitive map, the relation is co-activation and the content represented by that relation is spatial proximity.

Structural representations work by there being a correspondence between a relation on representational vehicles and a relation on the entities represented by those vehicles. The way that a structural representation enters into computations trades on that correspondence. For instance, internal computations over chains of place cell activation are used to calculate the relative length of potential routes through the environment. That computation works as a way of calculating route lengths because there is a correspondence between place cell co-activation and spatial proximity. The relation which defines the structure has representational content and the way it is computed with depends on that specific content. These are content-specific transitions.

What is coming into view here is a connection between the distinctions discussed in this chapter and those in the last. On the one hand, non-content-specific transitions are characterised by the generality of their operation, their content-faithfulness turning only on the content of broadly-logical terms; similarly, language-like compositionality is characterised by the generality of its application. On the other hand, content-specific transitions are faithful to content only because of the specific content of non-broadly-logical terms; and the compositionality of a structural representation is special purpose: its compositional device has specific representational significance (e.g. it represents spatial proximity). Special-purpose compositionality seems to line up with content-specific computation, and general-purpose compositionality with content-general computation.

However, as we will see shortly, there is no necessary connection here. Conceptual representations can enter into content-specific transitions (§3.4). It is also possible for structural representations to be operated on in content-general ways. But there is an underlying alignment between type of transition and form of compositionality. Plausibly, what gives the compositional device at work in a structural representation a specific content (e.g. representing spatial proximity) is the way the representations are used computationally—in transitions whose content-faithfulness depends on that specific content. Alternatively, the specificity of the content means that stronger conclusions can be drawn in inference, content-specifically, than would be possible if the compositional principle did not have specific representational significance. Either way round, there is a reason why special-purpose composition should go hand-in-hand with content-specific transitions.

However, once a structural representation is established there is nothing to stop it being operated on in content-general ways. To do that in rich ways may require the representational system to include some broadly-logical terms or devices. Even without that, inferences could be drawn from a map that depend only on the vehicle relation representing a transitive relation (not necessarily spatial proximity); or even just on the fact that it represents a two-place relation. Something similar is going on with Venn diagrams. In a Venn diagram, spatial relations stand for set inclusion. Reasoning with a Venn diagram can establish various set-theoretic conclusions. That case is more complex, since Venn diagrams are used by an interpreter rather than being computed over directly. (There are more examples of analogical inference in §4.6.) But it illustrates how a structural representation can be computed with in less content-specific ways. In short, structural representation goes with content-specific computation in the sense that structural representations are likely to enter into content-specific computations and forming structural representations does not require a capacity content-general computation.

Similarly, content-general computation aligns with general-purpose compositionality, but again with no necessary connection. Plausibly what makes a language-like compositional principle like predication general-purpose is that it is operated on in ways that do not depend on its having any specific content, that is, in content-general ways. In both 'Socrates is human' and 'all humans are mortal', the mode of combination serves to predicate biological properties of people. But the logical conclusion we draw does not depend on a restriction to biological properties. It would work for any objects and properties. Alternatively, in a representational system that includes a general-purpose mode of combination, that expressive power would be idle if representations were only ever computed with in ways that depend on specific contents—if we only reasoned with 'Socrates is human' in ways that depend on the combination concerning biological properties of people. Only when the sentence enters into broadly-logical inferences do we take advantage of the general-purpose nature of the compositional principle. Nevertheless, language-like representations can be operated on in content-specific ways, as we will see in the next section, so there is no necessary connection here. Rather, general-purpose compositionality tends to go along with having the capacity for content-general computation.

In short, there are good reasons why the distinctions in this chapter and the last tend to align. Content-specific transitions over representations displaying special-purpose compositionality is the base case. Content-general transitions over representations displaying general-purpose compositionality is a more sophisticated computational achievement.

This all relates to a sense in which the conceptual representations used in conscious deliberation are propositional. We saw in the last chapter that there is a stronger sense of *propositional*, beyond merely being truth-assessable, according to which a propositional representation has sentence-like structure. This is sometimes related to the idea that propositional representations support logical reasoning. We can now see why. A representational system that supports language-like compositionality, and that includes terms for conjunction and negation/disjunction, is one over which broadly-logical inferences could be performed. A representation's being propositional in this stronger sense enables logical reasoning involving conjunction, negation, and disjunction.

Content-general computation also goes along with another property often thought to be distinctive of classical computation, and of a language of thought, namely role-filler independence (Hummel et al. 2004; Quilty-Dunn et al. 2023). A representational system shows complete role-filler independence when the meaning of the individual elements in a combinatorial structure does not vary depending on how they are combined. In some maps, the location picked out by a point depends on its relations to other points on the map. The points do not, then, display role-filler independence. Or if, in natural language, the predicate 'green' means something different in the compound 'green leaf' than it means in 'green tomato' (Travis 1997), then that mode of combination would not exhibit role-filler independence. The truth-conduciveness of broadly-logical inference requires that the meaning of terms does not shift during the course of the inference—at least not too much. If 'human' means something very different in the first and second premises of the Socrates inference (1)–(3), then the truth of the premises will no longer make probable the truth of the conclusion. (With logical inference, strict validity depends on there being absolutely no equivocation when terms are reused.) So content-general computation requires a good degree of role-filler independence, at least within an inferential step. To turn to an example in Chapter 5, it could be that DOG means something different in the thought DOG BITES MAN than it does in the thought MAN BITES DOG, but if these were both relied on in a single chain of inference, the shift could introduce some inaccuracy.

Having seen that the structural and computational distinctions tend to align (but with exceptions), in the final section I want to elaborate on an important exception: the way conceptual representations are involved in content-specific transitions.

3.4 Content-Specific Transitions Involving Concepts

In the examples we have considered so far, content-general transitions have involved deliberate reasoning, taking place over conceptual representations; and content-specific transitions have occurred within special-purpose systems, involving non-conceptually-compositional representations (representations that do not have language-like compositional structure). Are concepts restricted to the sphere of content-general computation? In this section I will argue that they are not. There are also content-specific transitions involving concepts. These come in two kinds, those linking conceptual representations to each other and those linking concepts to representations of other kinds.

Starting with the latter, dispositions to categorise are often content-specific transitions. We move from seeing a certain arrangement of shapes, colours, and textures to forming the thought THAT IS A DOG.[4] Although we can be misled by how things look, this disposition works pretty well most of the time. It is faithful to content because the features represented in perception are a reliable (but fallible) sign of the presence of a dog. That is to say, it is a content-specific transition. Although concepts can, of course, be applied by reasoning from other concepts, theories of concepts have to leave space for concepts to be applied based on perception or other special-purpose resources. Elisabeth Camp has an expansive account of what she calls 'characterizations', on the basis of which we can apply a concept (or characterize its referent). Christopher Peacocke's theory of concepts, with a narrower focus, has transitions from non-conceptual representations to concept application. For example, the concepts SQUARE and DIAMOND are applied on the basis of perceptual representations of a four-sided figure and its bisectors (Peacocke 1992, pp. 74–7).

There is empirical evidence for distinguishing between applying a concept by reasoning from other conceptual representations and applying a concept based on content-specific transitions from non-conceptual representations. Experimental evidence shows that there are two routes to categorisation (Ashby and Valentin 2017). One route to applying a concept is automatic and relatively insensitive to concurrent cognitive load. A concept is applied based on a weighted sum of many sensory dimensions. The other route is deliberative, drawing on working memory and sensitive to cognitive load (Smith and Grossman 2008). It involves application of a rule that the thinker represents explicitly and can report

[4] The property of being a dog may already be represented in perception, but there are almost certainly some cases where the act of categorisation introduces a new content: applying the concept INGRATITUDE, say, or (legal) CONTRACT. Even if the property of being a dog is represented in perception, the transition from that representation to application of a concept, although less substantive, would also be content-specific. A disposition to transition from a feature-placing or non-conceptually-compositional content involving dogs to applying the predicative concept DOG to a particular is only faithful to content because of the specific contents of the representations on each side of the transition.

(e.g. that Cs have long legs and no spots), using a process that can only rely on one or two dimensions at once.

The two processes have distinctive psychological and neural signatures (Ashby and Maddox 2011). Learning a multi-dimensional categorisation requires feedback and works better if the category label comes after the sample being categorised (but only shortly after). With rule-based learning, a category label is not essential and, when given, helps even if it comes before or long after seeing the exemplar. Learning a multi-dimensional categorisation is impaired by switching the location of the response key and is minimally compromised by concurrent cognitive load. This dissociation predicts that when people categorise stimuli on the basis of multiple dimensions they may not be able accurately to report the basis on which they sort stimuli into categories (as found by Hampton and Passanisi 2016; and by Frith and Frith 1978). People may know many of the reasons an animal is categorised as a dog, for example, but this need not accurately reflect the features relied on by their own categorisation dispositions.

We can caricature these two types of transition as follows:

(16) That object has long legs and no spots
(17) If an object has long legs and no spots, then it is a flug
(18) That is a flug

(19) [That is high on three, and low on four, specific perceptible features]
(20) It's a dax

The route involving reasoning, (16)–(18), is a content-general transition. (Reasoning can also encompass content-specific transitions between conceptual representations *inter se*—an idea we return to later.) The multi-dimensional route, (19)–(20), is a content-specific transition. (The description in square brackets is shorthand for some collection of perceptual representations.) It is a content-specific transition from other resources to a conceptual representation.

Concepts are also involved in content-specific transitions in the other direction, from a conceptual representation to other resources. When I think about dogs using my DOG concept, that brings to mind the characteristic sight, sound, feel, and smell of dogs. Theorists talk about this information as being 'associated' with the concept, but these are not bare associative links, in the way that SALT makes us think PEPPER, say. They encode information about the subject matter of the concept. As Camp puts it, they serve to characterize the referent (Camp 2015). This can be quite important to the way a concept works. For example, many think it is central to the way that moral concepts work that when we categorise a situation as falling under a concept like UNFAIR, or categorise a person as falling under a concept like CRUEL, that produces a characteristic emotional response (a reactive attitude: Strawson 1962). In short, we move in thought from conceptual representations to a whole host of sensory, motoric, affective, and evaluative representations. These transitions tend to be appropriate to the subject matter. To the extent that they are faithful to content, that is because of the specific contents involved. They too are content-specific transitions.

Often special-purpose representations act as intermediates in a process that starts with one conceptual thought and eventually takes us to another. I am trying to work out whether a chair I want to buy will fit in the car. Having asked myself the question (conceptually), I imagine rotating the chair at various angles to see if it is likely to fit. When I conclude that it won't, that is a conceptual judgement. The thought process that took me there proceeded via some special-purpose, not-conceptually-compositional representations. This relies on content-specific transitions at every stage: from concept to visual image (of the chair), amongst visual images and motor commands (mental rotation), and back to a conceptual conclusion (categorising the resultant image under DOES NOT FIT). I call the process of getting from one conceptual thought to another in this way a *mediated* content-specific transition ('mediated-CS transition').

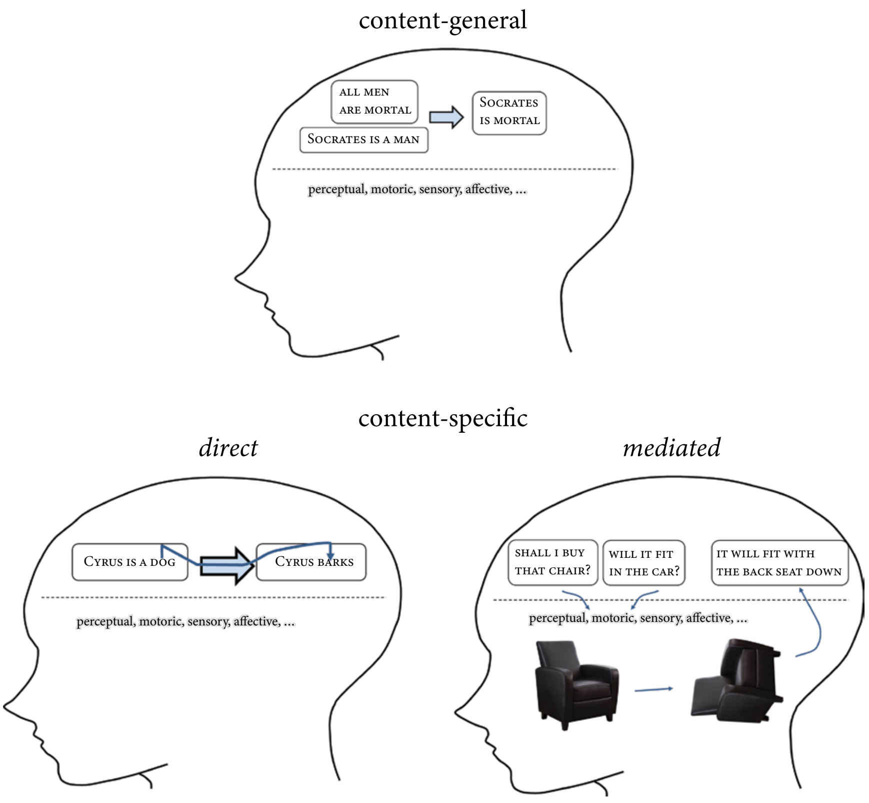
* * *

We will examine mediated-CS transitions in detail in Chapter 5. I mention them now only briefly, in order to distinguish them from *direct*-CS transitions (see Fig. 3.2). These are transitions that occur between conceptual representations directly, taking place within conceptual thought. Our existing examples of reasoning have been broadly-logical and content-general. But there are also plausibly links between concepts *inter se*, links that underpin transitions which, without being logical or deductively valid, work well enough to be useful (Machery 2017, p. 222). For example:

(21) Moby is whale
(22) Moby is a mammal

(23) Cyrus is a dog
(24) Cyrus barks

The idea is that the thinker is disposed to make transitions of these forms whatever singular concept is substituted for MOBY or CYRUS. They are not necessarily truth preserving, but if a thinker is disposed to make direct DOG-BARK inferences like (23)–(24), they won't go wrong very often. It is a worthwhile disposition to have. The transitions are faithful to content in the sense that the conclusion is likely to be true if the premise is true. The above examples count as direct-CS transitions provided the thinker has the disposition to make transitions of the following forms, without depending on some further premise:



content-general

ALL MEN ARE MORTAL
SOCRATES IS A MAN
SOCRATES IS MORTAL
perceptual, motoric, sensory, affective, ...

content-specific

direct

CYRUS IS A DOG
CYRUS BARKS
perceptual, motoric, sensory, affective, ...

mediated

SHALL I BUY THAT CHAIR?
WILL IT FIT IN THE CAR?
IT WILL FIT WITH THE BACK SEAT DOWN
perceptual, motoric, sensory, affective, ...

Fig. 3.2 Types of representational transition: content-general and content-specific (*direct* and *mediated*). (The cartoon head is intended to show that these things are going on the mind. There is no correspondence to parts of the brain. The dotted line serves to separate representations in conscious deliberation that are constructed out of concepts from the rest. This is not a matter of levels of processing.)

(25) x is a whale
(26) x is a mammal

(27) x is a dog
(28) x barks

We need examples where the transition occurs when the thinker is in 'factual mode', working out what is the case or what to do, and not simply free associating. A direct-CS transition is not just a matter of spreading activation (cf. Vicente and Martínez Manrique 2016), like the way TREE activates PALM which in turn activates WRIST (Marcel 1980). Furthermore, the disposition is specified at the level of complete representations. For example, if we substitute *x is not a whale* for premise (25), the thinker would not be disposed to transition to (26). Notice that the faithfulness to content of these transitions depends on the specific contents involved. A disposition to move between two arbitrary predicates, x is F to x is G, would make no semantic sense.

It is an empirical question whether there are in fact any direct-CS transitions between conceptual representations. The most promising examples include cases like whale→mammal, kill→die, and red→coloured (Laurence and Margolis 1999).[5] In particular, although hierarchical category inclusion relations are sometimes calculated using typical features (typical dogs have features of typical mammals), it may be that inferences between atypical instances, like whale(x) → mammal(x), are stored directly (Murphy 2002, p. 209; citing Glass and Holyoak 1974). There is evidence that, when certain concepts are tokened, related property concepts are activated, whatever the context, for example RABBIT→FUR and GUN→TRIGGER (Whitney et al. 1985; see also Machery 2015). To the extent that these reflect transitions between complete conceptual representations, and not bare associations between concepts, they are also evidence of direct-CS transitions. Research on 'shallow' semantic processing (Solomon and Barsalou 2004) provides further evidence (which we return to below).

Importantly, these are cases where the transition does not depend on any further premise or other input. They contrast with cases where the thinker will only draw the relevant conclusion when they are explicitly representing the general premise. Lea, Mulligan, and Walton (2005) illustrate one way of testing whether people are relying on an explicit premise in making an inference. They use inconsistency to probe the representations people build up when reading a text. The same sentence will take longer to read when it conflicts with the discourse representation the reader has formed up to that point. If a text has said that Nathan broke down in his car and missed the chance to give his best man's speech before the start of the wedding meal, readers will slow down when reading the sentence, 'The groom's mother complimented Nathan on his wonderful speech'. In one condition, rather than being told explicitly that he missed the chance to give his speech, the reader had to infer that from the premises 'If Nathan does not arrive before the food is served, he will not be able to give his speech', and 'By the time he got back, people were halfway through the meal'. The measure of whether people infer that he did not make his speech is that they read the same sentence (complimenting him on the speech) more slowly in the conflict condition. In this experiment, participants only made the inference if the conditional premise occurred shortly before the antecedent, or if they were reminded of it by a cue, not if they had read it ten sentences before. This suggests

[5] A rival explanation for these transitions would be that the determinate-determinable or class inclusion relation forms an explicit premise and the conclusion is reached in reliance on a broadly-logical rule, e.g. F is a determinate of G, x is F, therefore x is G. That would then be a content-general transition.

that drawing the conclusion depended on explicitly representing the conditional premise. The case contrasts with inferences based on ‘pragmatic world-knowledge’, which require no explicit premise. On reading, ‘The angry husband threw the delicate porcelain vase against the wall’, readers will slow down over a subsequent sentence in which the vase is intact. There is no need to tell them explicitly, ‘if a delicate porcelain vase is thrown against a wall, then it will break’.

In the same way, the transition from (27) to (28) would count as a direct transition provided it does not depend on explicitly representing that *dogs bark*. The inferential disposition is ‘built into’ the concept DOG. Just tokening the concept in an appropriate syntactic frame in factive mode disposes the thinker to make the transition. In this particular respect it is like the content-general transitions in the last section: just tokening the concept IF…THEN in an appropriate syntactic structure when reasoning about what is the case disposes the thinker to make a modus ponens inference.

The direct-CS disposition works by effectively presupposing the generic content that *dogs bark*, but there is no vehicle for that background assumption. The agent relies on a fact about their situation without articulating it explicitly (cp. ‘situated inference’: Barwise 1986). The content *dogs bark* is represented only implicitly—represented by means of a disposition to make transitions between explicit representations (27) and (28) (Shea 2015). The transition is not relying on a representation that *dogs bark* that could enter into computations in its own right. Assimilating direct-CS transitions to logical inferences with a further premise would elide a psychologically real difference (e.g. explicit LTM vs. implicit LTM: Smith and Grossman 2008) and obscure the distinction between content-specific and content-general computational processes.

Although some direct-CS connections could be innate (*sensu* unlearnt), most are doubtless built up from experience. For an illustration of how direct inferential dispositions are overturned by experience, think of the way students start reasoning with the concept of infinity when they first learn the symbol ‘∞’. We start using INFINITY as if it were a natural number. We soon learn that this leads us into falsehoods and contradiction, so we acquire a disposition to infer with INFINITY in more circumscribed ways. That is an experience-based change to patterns of direct-CS transition.

There is a history to the idea of direct-CS transitions. At one time it was thought that a lexical concept encoded a definition, represented explicitly. In response to strong evidence against explicit definitional structure, Fodor turned for a while to ‘inference rules’ to account for semantic inference (Fodor, Fodor, and Garrett 1975; but cf. Fodor 1998, pp. 108–12). Inference rules are in fact a type of direct-CS transition between conceptual representations. This was in effect to adopt what Laurence and Margolis (1999, p. 5) call an ‘inferential model’ of how a concept is ‘structured’ (as opposed to the earlier ‘containment model’, which was a matter of representational structure in our sense). The idea was that transitions like the whale–mammal inference, (25)–(26), are made quickly and automatically as a result of tokening the concept. Putative examples include:

(36) x is a bachelor → x is an unmarried man
(37) x causes y to die → y dies
(38) x is red → x is coloured

The conclusion is reached in each case, not by recalling a piece of information from memory, but by executing a procedure that is built into the rules for processing the focal concepts (BACHELOR, CAUSE, RED).

These transitions were called meaning postulates (Montague 1974). Meaning postulates should be necessary, analytic, and partly individuative of the concepts involved. Meaning postulates have now been rejected by most theorists because it turned out to be difficult to identify a privileged set of inferences that are necessary, analytic, or plausibly individuative of a lexical concept. Nevertheless, having given up the claim that there is a privileged set of analytic inferences that individuates a concept, we can still distinguish between procedures that involve a concept and information explicitly represented using a concept. For example, relevance theorists treat concepts as atomic and then distinguish between two ways that the contents connected with a concept are stored (Carston 2010; Allott and Textor 2012). Most information is stored in the form of 'encyclopaedia entries' in memory: propositionally-structured assumptions and beliefs, and non-conceptual imagistic and/or sensory-perceptual representations; but there are also 'logical entries'. These are inference rules involving the concept (Carston 2010, p. 246).

Inference rules work like meaning postulates in terms of how they are processed, but they are not analytic or individuative of a concept, nor need they be deductively valid (Block 1993; Machery 2017, 222). They are revisable, and a concept's identity may survive changes to the direct-CS transitions in which it is involved. They can encode ordinary empirical information (like the fact that *dogs bark*). While my thesis is neutral about the issue, it seems very likely that most are acquired through learning. The key point for our purposes is that they are procedural and not explicitly represented. They subsist directly between premise(s) and conclusion, without a further explicit representation being relied on.

Cognitive science has a long history of drawing a distinction between rules and associations (Pinker 1991; Quilty-Dunn and Mandelbaum 2019). Rules are often thought to go with symbolic representation and classical computational systems. Associations align with subsymbolic representation and artificial neural networks or connectionist computational systems. The distinction I have drawn here problematizes the rules/associations dichotomy. From one point of view, direct-CS transitions are a kind of rule that operates over symbolic, conceptually-structured representations. They are inference rules. From another point of view, direct-CS transitions are more like associations. They reflect statistical structure and are likely acquired through entities being connected in experience; they are unlike the theorem-proving of classical computers and more like the trained content-specific dispositions of an artificial neural network. So while accepting that what were called 'inference rules' are a kind of direct-CS transition, I want to eschew the unhelpful rules/associations dichotomy in favour of various better distinctions—the different kinds of representational structure enumerated in the last chapter, and the distinction between content-general and content-specific transitions defined in this one.

Triggering a direct transition disposes the thinker to token a certain conclusion, but whether that conclusion is in fact reached may depend on competing dispositions and other facts. The disposition may have exceptions or be cancelled in certain contexts (e.g. in a high-stakes context). And whether the conclusion is in fact drawn will depend on what else is being represented concurrently (as in classic dichotic listening experiments, e.g. Lackner and Garrett 1972). For example, the conclusion may not be drawn if it contradicts something that is already represented to be the case. Still, triggering the disposition does not depend on further premises being represented.

Digital computers use direct-CS transitions to encode some types of information. Most processing steps draw on explicitly represented information retrieved from memory. Some, however, are built in. Gallistel and King (2009) give the example of a 'literal': where the value of a constant like π is not represented explicitly but is implicit in a processing disposition. Think of a procedure set up so that it outputs 3.14159 times the number given as input. The procedure can be used to calculate the circumference of any circle, given its diameter. But the value of π is not looked up or explicitly represented anywhere. It is purely implicit in the processing disposition.

In the examples so far, the direct links have existed between conceptual representations piecemeal, a pair at a time. However, there may also be more systematic connections between a whole collection of concepts. Artificial neural networks trained in natural problem domains often develop similarity spaces in their hidden layers (Laakso and Cottrell 2000; Khaligh-Razavi and Kriegeskorte 2014). There is growing evidence that the same thing happens with conceptual spaces in the brain (Cichy et al. 2016; Mok and Love 2019). Paul Churchland argues that concepts are represented in a semantic state space which has the property that tokening one conceptual representation disposes the thinker to move to tokening nearby conceptual representations (Churchland 1998, 2012). If that is right, some direct transitions between concept-involving representations are underpinned by the topography of the state space in which they are represented.

Reaching a conclusion in virtue of a direct-CS transition between conceptual representations is significantly different from reaching a conclusion by relying on further premises. While the latter could be content-general, depending only on the content of the broadly-logical concepts involved, direct concept-concept transitions are content-specific. They are only faithful to content in the light of the specific contents of the open class concepts involved.

Direct-CS transitions also contrast with the mediated-CS transitions mentioned above (which we will return to at length in Chapter 5). Solomon and Barsalou (2004) performed experiments that serve to contrast the two. In a property verification task, participants were given two words and they had to say whether the first refers to something that is a physical part of the second: given HORSE-*mane* the correct answer is 'yes'. As lures they were given entirely unrelated pairs like PLIERS-*river*, and also associatively-related pairs like DONKEY-*mule*. Participants gave correct responses more quickly when faced with unassociated pairs than with associated but false pairs: for the associated pairs, making the true-false judgement took longer. Only these difficult cases led to neural activation in perceptual systems (Kan et al. 2003). These results suggest that there are 'shallow' semantic connections, consisting of the presence or absence of a direct link, that obtain between conceptual representations directly, without relying on activating associated sensorimotor representations.

A plausible interpretation is that participants were relying on direct transitions to rapidly differentiate pairs like HORSE-*mane* from PLIERS-*river*, but went via mediated transitions to differentiate HORSE-*mane* from DONKEY-*mule*. There is evidence that these direct transitions are activated earlier in the processing sequence (Lea et al. 2005). They are selectively spared when other semantic relationships are impaired in Alzheimer's disease (Glosser et al. 1998). These results thus show that there is an empirical contrast between (at least some forms of) direct and mediated transitions.

In short, there is an important difference between direct- and mediated-CS transitions between conceptual representations. In one way, direct-CS transitions pattern with content-general transitions, since both consist of moves between conceptual representations directly. In another important respect, however, direct-CS transitions pattern with mediated-CS transitions, since both are content-specific.

To sum up, there is an important difference between two ways that representations can be computed with—two ways that transitions between representations can be configured so as to be faithful to content. Faithfulness to content may be content-specific, depending on the specific contents of the non-broadly-logical concepts involved, or content-general, depending only on the content of the broadly-logical concepts involved. The various ways that information is stored in the mind—'informational models' (Chapter 4)—may operate in either way.

Structural representations, relying on special-purpose compositional principles, tend to enter into content-specific transitions. Language-like representations, like the conceptual representations involved in conscious deliberation, relying as they do on general-purpose compositional principles, are well suited to content-general computations. It is these types of representational structure and computational process that need to be integrated in an account of concept-driven thinking (Chapter 5).

Chapter Summary

3.1 Transitions Faithful to Content

This chapter argues that RTM's commitment to representations being processed so as to be faithful to content can be achieved in two quite different ways: content-general and content-specific transitions. In computing machines, mechanical or electronic, physical particulars representing numbers interact in ways that perform useful computations. (p. 60)[6] A basic way to make transitions faithful to content is to implement logically valid inferences. These basic components are arranged so as to perform useful operations, like multiplication.

There are many ways of being faithful to content; it is sufficient if the conclusion is likely to be true, given true premises, often enough to be useful. This notion does not require that every transition is faithful to content—there can be errors and transitions that are systematically mistaken in some contexts. (p. 61) We don't need to say here how much faithfulness is required and in which circumstances; our question is: which contents does faithfulness turn on?

In many cases studied by cognitive neuroscience, processes are wired up by experience just to work well enough in the organism's normal environment. Research on DNNs and neural activation patterns suggests that the internal transitions are highly specific to the subject matter on which the network was trained. (p. 63) Similarly, desert ants navigate based on analogue computations involving representations of angles and distances. More abstract analogue computations, for example approximate arithmetic, are much more generally applicable, but still turn on certain specific contents being represented (numerosity or magnitudes). Many computational models in cognitive neuroscience are also like this, including the cognitive map. The cognitive map is also a structural representation; faithfulness to content turns on this (§3.3). (p. 64) All these examples

[6] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

contrast with logical inferences, which are faithful to content irrespective of the specific subject matter involved.

3.2 Content-Specific and Content-General Transitions

This section pins down the difference, introduced in the last section, between two different ways in which a transition can be configured so as to be faithful to content. This develops Sellars's distinction between logical inference and material inference. A content-general transition is one whose faithfulness to content depends only on the content of the logical terms (examples). (p. 65) They need not be deductively valid. They just need to work well enough to be useful. They are characterised by being very generally applicable as to subject matter. 'Broadly-logical terms' are those on whose content the truth-conduciveness of a generally-applicable pattern of inference depends. Content-general transitions = faithfulness to content turns only on broadly-logical terms (if any). (p. 66) Definitions: content-specific transition; content-general (or non-content-specific) transition.

Content-specific transitions depend on their particular subject matter; content-general transitions do depend on content, but only on broadly-logical content; that may be fixed just by the form of the transition. This relates to defining logical operations as those that are invariant under permutations of objects, accepting that terms like 'most' are broadly logical. (p. 67) Bayesian inference is content-general; approximations to Bayesian inference are likely to lie at the content-general end of the graded distinction; cheater detection could be either.

Content-general transition is a generalisation of the idea of variable binding, as used in computer science. Content-general transitions need not involve operations on variables. (p. 68) Variable-based inferences need not be content-general; being content-general is the more inclusive category.

The notion of computation should not be restricted to the broadly-logical. Until recently, the vast majority of computers depended on content-general transitions; phylogenetically, it is likely that content-specific transitions came first. (p. 69) Whether the capacity of broadly-logical transitions depends on cultural learning is an open question.

My distinction is not completely novel—it is related to, but deeper than, the distinction between classical and connectionist architectures, symbolic vs. sub-symbolic computation, or model-building vs. pattern recognition. Nor is it a matter of compositional vs. non-compositional, or modular vs. non-modular, although there is some rough alignment—it is a new way of drawing a deep distinction in the vicinity of existing dichotomies.

(p. 70) Next: connections between this distinction and the difference in representational structure highlighted in the last chapter.

3.3 Types of Transition Go with Types of Representational Structure

In the last chapter we saw that the general-purpose compositional principles of language-like representations contrast with the special-purpose compositional principles at work in a structural representation. In a structural representation, the relation that defines the structure has content and the way it is computed with depends on that specific content. Special-purpose compositionality lines up with content-specific computation, and general-purpose compositionality with non-content-specific computation.

There are no necessary connections, but there are reasons—based on what gives the compositional principle its specific content; and/or the strength of the conclusions that can be drawn—why the special-purpose compositionality of structural representations tends to go hand-in-hand with computational transitions that are content-specific. (p. 71) Once a structural representation is established, there is no bar on its being operated on in content-general ways. Similarly, there are good reasons—based on what makes the compositional principle general-purpose; and/or use of its expressive power—why content-general transitions tend to align with general-purpose compositionality.

(p. 72) Performing content-general transitions over representations displaying general-purpose compositionality is a more sophisticated achievement. The argument here shows why two senses of 'propositional' tend to go together: having language-like structure; and supporting logical reasoning involving conjunction, disjunction, and negation. Content-general transitions require a good degree of role-filler independence, at least within the scope of a chain of inferences. The next and final section will elaborate on an important exception to the alignment between transition and structure discussed here.

3.4 Content-Specific Transitions Involving Concepts

(p. 73) Concepts also take part in content-specific transitions: between conceptual representations and other representations, and amongst conceptual representations themselves. Categorisation is often a content-specific transition. Categorisation can also occur by reasoning from other conceptual representations—experiments show functional differences between these two routes. (p. 74) They have different psychological and neural signatures. Stylised examples of each. Moving from a conceptual representation to sensory, motoric, affective, and evaluative representations is also a content-specific transition. (p. 75) When special-purpose representations act as intermediaries from one conceptual thought to another, we have a *mediated* content-specific transition (more on this in Chapter 5).

Links between concepts *inter se* can underpin *direct* content-specific transitions between conceptual representations, for example: x is a whale → x is a mammal. Without a further explicit premise, the transition does not exhibit necessary truth preservation, but it is still a worthwhile disposition to have. (p. 76) This is not just a matter of associations between concepts: these transitions occur when the agent is thinking 'factually'.

(p. 77) There is some evidence for direct-CS transitions (i.e. where there is no explicit general premise). Experiments can show whether a further explicit premise is involved in drawing an inference. (p. 78) The DOG→BARK transition counts as direct provided just tokening the concept in the appropriate syntactic frame in factual mode (e.g. Fido is a dog) disposes the thinker to make the transition. The general premise, for example that *dogs bark*, is merely implicit—it is not a representation that could enter into computations in its own right. Direct-CS transitions are doubtless built up from experience; an example is changes with experience in the way students use the concept of infinity (and the symbol '∞').

Transitions of this type were once relied on by Fodor as 'inference rules'. (p. 79) As 'meaning postulates', these were taken to be necessary, analytic, and individuative of concepts; dropping those commitments, there is still space for inference rules (relevance theorists' 'logical entries', contrasted with 'encyclopaedia entries'). Direct-CS dispositions encode ordinary empirical information and are revisable. They are poorly classified by the traditional rules/associations dichotomy. (p. 80) Whether the conclusion is drawn will depend on what else is represented in context, but triggering the disposition does not depend on further premises. Digital computers use some direct-CS transitions, for example in having a 'literal' (procedure) encoding the value of π. Some direct-CS transitions between conceptual representations may be underpinned by a representational state space, rather than occurring piecemeal. (p. 81) The absence of an explicit general premise means that the direct transition is only faithful to content because of specific contents (e.g. *whale* and *mammal*), and so contrasts with content-general transitions between conceptual representations.

Experiments support the distinction between direct-CS transitions and mediated-CS transitions. Direct transitions are activated earlier in processing and are selectively spared when other semantic relationships are impaired in Alzheimer's disease. Direct-CS transitions are like content-general transitions, in that they occur between conceptual representations directly; but they are like mediated-CS transitions, in being content-specific.

To sum up, there is an important difference between two ways that transitions can be configured so as to be faithful to content: content-specific and content-general transitions.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0003

4
Informational Models

4.1 Ways of Storing Information

The overall aim of the book is to present a picture of how concepts are involved in the processes by which people work out what is the case or what to do. Prospection or forward planning requires the agent to have some kind of model of the world, a model which they use to calculate the likely consequences of their actions. Conscious deliberation draws not just on semantic memories—explicit representations composed out of concepts in a general-purpose way—but also on information encoded in many other forms.

In work on reinforcement learning an apparently simple contrast is drawn between model-based and model-free decision-making—decisions which do or do not rely on a world model. In fact, the cases where a thinker relies on some kind of mental model of the world in working out what is the case or what to do are rather varied and heterogeneous. There are a whole range of ways of planning and deciding that go beyond the decision-making of model-free reinforcement learning, as we shall see.

To encompass the range of what could be considered to be a cognitive or psychological model in this inclusive sense I will use the term 'informational model'. An *informational model* is a collection of interconnected representations, of any kind, which can support planning, prospection, or other forms of inference that derive novel conclusions about what to do or what is the case. I am not committed to there being any clear boundary between what should and should not count as an informational model, nor for what should count as having a 'world model' or 'knowledge structure' about some aspect of the world. What the cases show is that there is a continuum between just storing some useful information and having a rich interconnected model. The cases vary along multiple dimensions in how information is encoded and computed with. Different forms of representing and storing information make different contents easy to update, easy to access, and easy to compute with for different purposes, with trade-offs between these advantages.

One kind of informational model is a database of stored semantic memories, for example the collection of explicit conceptual representations about Paris that are stored in my long-term memory. Conscious deliberation has paradigmatically been regarded as drawing on this kind of information. I have been emphasising the way deliberate thinking also brings information and conclusions drawn from informational models of other kinds into the playground of thought. Some of these models—like cognitive maps or semantic state spaces—support non-local inferences. The aim of this chapter is to give a sense of the variety of informational models that can be relied on when we are deliberating.

The aim of the next chapter (Chapter 5) is to paint a picture of how concepts held in working memory allow us to rely on informational models and integrate them with information accessed through other concepts. The challenge is that conceptual thought has to integrate informational models, representational structures, and computational processes of different kinds. We will see, in this chapter, that informational models use representations with various kinds of semantically-significant representational structure (from Chapter 2), characterised by different combinations of the six features highlighted at the end of §2.1. In terms of computational process (Chapter 3), many informational models work with content-specific transitions, contrasting with the content-general transitions by which we can reason with conceptual representations. Some informational models may also support less content-specific and more broadly-logical computations (§4.7). It is the variety in the characteristics of these informational models which sets up the problematic to which the picture in Chapter 5 is a response.

This chapter divides the cases into sections in somewhat arbitrary way. I start with simpler cases, but there is no straightforward hierarchy of complexity. Different informational models that are relied on by deliberation illustrate different combinations of the features discussed in Chapters 1, 2, and 3. Of these, in this chapter I focus on:-

(i) Representational structure: organized representation, structural representation, general-purpose compositionality.
(ii) Computational process: content-specific, content-general.

Where evidence is available, I also comment on:-

(iii) Whether the computations are local or non-local.
(iv) What is found in the cognitive playground: inferences over the model, or just its outputs; whether the thinker deliberates with the informational model itself, or just its outputs.

And whether inferences over the informational model are modular in the sense of:

(v) special-purpose or general-purpose;
(vi) informationally encapsulated or not;
(vii) fast and automatic, or working-memory-dependent and deliberative.

4.2 Information in Domain-Specific Transitions

As we saw in the last chapter (§3.2), one way that systems store information about the world is in their dispositions to make content-specific transitions between representations. On a common understanding of visual processing, the primate ventral stream contains dispositions to move from certain distributions of contrast to representations of edges and textures, from edges and textures to surfaces, and then to objects and so on. Another set of examples involve perceptual decision-making (Gold and Shadlen 2007). We might see an array of random dots in motion and have to decide if there is coherent motion to the right or to the left. In macaque monkeys trained to perform this task there is evidence of a neural circuit that integrates moment-by-moment probabilistic information, produces a maximum likelihood estimate, and programs a behavioural output in response (Gold and Shadlen 2001; Beck et al. 2008; Shea 2014b). In both cases, content-specific dispositions implicitly encode information about the statistics of the normal visual environment.

The human visual system is not often talked of as having a model of the world, but the deep neural networks which perform the same task (Krizhevsky, Sutskever, and Hinton 2012), in something like the way the visual system does (Yamins et al. 2014), are often described as statistical models of the problem space. Special-purpose perceptual processing systems should also count as containing informational models in my sense. To the extent that they can be reused in offline simulation for the purposes of planning and deciding, they can be thought to encode a model of the task space.

These are informational models in only a thin sense. They contain information about regularities in the world, but since that information is represented only implicitly, it cannot enter into computations in its own right. Deliberation can only make use of it by running simulations over the representations between which the content-specific dispositions subsist. It is little more sophisticated than another case that is always described as model-free, namely model-free reinforcement learning.

A classic example of behaviour that appears model-free is demonstrated in the trap tube task (Mulcahy and Call 2006). Subjects have to use a stick to get food out of a clear horizontal tube. The tube has a visible trap in the middle down which the food will fall and get stuck. Most of the non-human animals that have been tested fail to appreciate that, moved the wrong way, the food will fall into the trap. After many trials they can eventually learn to push from the correct end, but then they continue to do so even when the tube is inverted so the trap will no longer catch the food and it would be easier to get the food from the other end.

The model-free reinforcement learning system learns from experience what reward will be obtained on average from each action it performs in each situation it encounters. These 'cached' action values do not depend on representing which kinds of rewarding outcome are produced by the actions (food, water, etc.), nor on representing how they are obtained, nor anything about the structure of the environment. That is why it is called 'model free'. So if an outcome loses value for the agent, for example because the agent has been fed to satiety with a certain foodstuff, the model-free system will keep picking that option until it can laboriously re-learn new values.

The model-free reinforcement learning system is very well-characterised in humans, primates, rodents, and even insects. A huge number of experiments have converged on a consensus about the computations responsible for reinforcement-driven choice behaviour and the brain mechanisms in which they are realized (O'Doherty, Cockburn, and Pauli 2017). Suppose that, when image A is on the screen, if the left button is pressed there is an 80 per cent chance that a grape will be delivered and a 20 per cent chance that nothing will happen; whereas for a right button press the chances are 30 per cent and 70 per cent, respectively. Organisms can learn the average payoff for each available action in each context. The model-free reinforcement learning algorithm does that by representing, when it takes an action, the expected value of that action-in-context. Comparing the expected value with the outcome that is actually delivered produces a teaching signal, a reward prediction error. This is used to update the stored action value, increasing it when the reward was greater than expected, decreasing it when it was less. There is strong behavioural and neural evidence for reward prediction error signals and their role in updating value representations through an algorithm known as temporal difference learning (Schultz, Dayan, and Montague 1997; Sutton and Barto 1998; O'Doherty et al. 2003).

It takes considerable experience to learn the correct action values. When the environment changes, the values have to be laboriously re-learnt. Once learnt, action values tell the organism what to do in each context. The value of action 1 in context A is cached in memory; ditto for all other actions and contexts. Encountering context A again, the system represents that action 1 is on average worth 0.8 and action 2 worth 0.3. These values are input into a decision policy. A standard decision policy will mostly pick the most valuable option. (But not always—it makes sense to pick the lower-value option some of the time in case the environment has changed.[1]) When 'deciding' what to do, there is a direct

[1] For a system with limited memory in a probabilistic environment it will always make sense to randomise to some extent (Icard 2021). A system with limited memory can work out the best option given information about the last n time steps. Sometimes that will also be the best option had it stored a complete history; other times a complete history would have recommended a different choice.

throughput from input to behaviour, with computational steps involving expected value on the way. The system knows what to do and has information about its environment, but it doesn't really know anything about the structure of the environment or about how its actions produce rewards.

Another example concerns motor control. We can accurately guide our limbs to points around us in space. This ability depends on a circuit that takes visual information from the limb in motion and continuously adjusts the limb trajectory so that it smoothly reaches the target. A detailed account of this visuo-motor control system, well supported by the evidence, involves forward and inverse models of how motor programs issue in behaviour (Wolpert, Diedrichsen, and Flanagan 2011). Once a target for action has been selected this system acts in direct throughput mode. Nevertheless, it involves predictions derived from an informational model (the forward model). The system is continuously predicting where the limb is likely to go given the current motor command. It uses the difference between these predictions and the goal state to update the motor commands being issued. Further updates are made when information arrives about where the limb is in fact going, but they come in later. Using predictions from the forward model makes for smoother motor execution. This is characterised as a model-based account, contrasted in the literature with a rival 'model-free' hypothesis in which movements are based directly on a learned policy (Hadjiosif, Krakauer, and Haith 2021). However, even the directly learned policy implicitly encodes some information, so if it could be used in simulation mode, then it too would qualify minimally as an informational model in my inclusive sense.

These three systems—special-purpose perceptual processing, model-free reinforcement learning, and motor control—all store information about the environment which they use in online processing. All three systems can also be relied on offline, by running simulations driven by deliberate thought. That is, they contain informational models, of at least a minimal kind. What about the properties canvassed in the first three chapters? All three systems deal with tasks that are reasonably stable over time, allowing the system to learn, from a wealth of experience, dedicated transitions between representation types, transitions that implement computations suited to the problems on which they have been trained. Thus, they make use of organized representations, computed over in content-specific ways. They have many characteristics of modularity: they are special-purpose (and domain-specific), fast, and automatic. Their computational steps are driven by specific inputs and not sensitive to the agent's overall beliefs—that is, they exhibit informational encapsulation. Their online operation does not draw on concepts. They can be used in simulation mode deliberately, deploying working memory, but it is only their outputs that are integrated into the cognitive playground. The information they contain is not directly available to be reasoned with in

Randomising so as to pick a lower-value choice some of the time is a way of sampling to see whether you are in one of these other histories.

deliberation. The thinker deliberates with the outputs of the informational model rather than with the model itself. Standard accounts of all three systems involve local rather than non-local computations (although it remains open that some relatively constrained but nevertheless non-local parallel constraint satisfaction may be at work in perceptual processing). In short, they exemplify a collection of features that contrast with the properties of conscious deliberation with conceptual representations. These characteristics have to be accommodated in any account of how these systems are relied on and integrated in conceptual thinking.

4.3 'Model-Based' Tasks

The last section looked at the characteristics of three relatively simple systems. We move now to systems that encode richer information about the structure of the environment and perform somewhat more sophisticated inferences with it. Research on 'model-based' decision-making examines the way an agent uses some kind of representation of the causal structure of the environment—for example of which actions cause which outcomes—to work out the likely consequences of their actions (Butlin 2021). These predictions are used to decide how to act. Model-based planning is studied with tasks where the choice at one step affects the options that will be available at future steps.[2] Turning left at a junction in a maze will open up one set of future choices and close off those available down the other branch. Moving the queen out now in a game of chess affects what can happen next. Chess players typically think through a branching space of future possibilities—effortfully—before they move a piece.

One set-up that has been used experimentally is a task similar to choosing an apartment. It involves searching for rewards and deciding at each time step whether to accept the current offer or to keep searching (Kolling et al. 2018). The experiments show that people decide in a way that takes into account how their decisions affect which states they are likely to encounter in the future (if you sign up for this apartment now and a better one comes along next week, you will miss out). Another task is a game in which the agent's choices may or may not affect the behaviour of the group they are playing with (Na et al. 2021). Taking the highest reward at step 1 is suboptimal if that is going to reduce how generous the rest of the group will be in future interactions. The researchers find behavioural evidence that participants engage in forward thinking, taking into account the long-run effect of their choices. Computationally, in both studies regions of ventromedial prefrontal cortex were found to encode the prospective value of the

[2] Butlin (2022) argues that this kind of learning is necessary for agency.

choice arrived at by forward thinking. Thus both behavioural and neural evidence showed that people were engaging in prospection, calculating the long-run consequences of their choices before deciding what to do.

As these results show, researchers are beginning to understand how information about task structure is stored and used. The classic test of whether an animal understands the connection between its actions and outcomes is to devalue one of the outcomes. For example, an animal can be fed to satiety on a previously valued foodstuff. An animal shows 'goal-directed' behaviour if it then turns down the chance to pull a lever (say) to get that foodstuff (Balleine and Dickinson 1998). It shows 'habit based' behaviour if it continues to pull the lever until it gradually learns that the outcome (foodstuff) is no longer rewarding. Model-based behaviour depends on representing world states and how they are related in a way that is independent of which particular states currently attract reward (Akam et al. 2021).

Courtin et al. (2022) measured and perturbed neural activity in the amygdala of mice exhibiting goal-directed behaviour. They observed one pattern of activity that reflects the value of the experienced outcomes and remaps when an outcome (e.g. getting sucrose solution) has been devalued. They observed another pattern of activity reflecting the action-outcome contingency (e.g. that the right-hand lever usually delivers sucrose solution). This activity was remapped when (e.g.) the lever that previously usually delivered sucrose solution started delivering milk instead. The animal's choices rely on a model that encodes action-outcome contingencies (e.g. the right hand lever usually produces sucrose) and outcome values (the current value to the animal of getting sucrose). This information is encoded, at least in part, in the amygdala, and relied on by the animal in planning its actions.

The most-used task for studying model-based reinforcement learning is the so-called 2-step task (Daw et al. 2011). Here is an analogy. I race sailing dinghies and live mid-way between two lakes. Whatever the national weather, on Lake Placid the wind is usually light but local conditions sometimes whip up a stiff breeze. On Lake Windermere it's usually windy but occasionally calm. I'm not a particularly good sailor, sometimes winning and sometimes losing races whichever lake I choose. One week I choose Windermere but it turns out, unusually, to be calm. I discover a new trick for getting more speed out of the boat in light winds, winning the race hands down. Which lake do I choose next week? Lake Placid of course—so that I have the best chance of getting to use my newfound skill in light winds. That is the signature of model-based choice. I choose A at step 1 but am sent down the low-probability branch. At step 2 I make a choice that is rewarded. This *decreases* the chance that, next time around, I will choose A at step 1. For a purely model-free reinforcement-learning agent, the chance of choosing A at step 1 would increase, because choosing A last time around was part of a chain of actions that led to reward.

When people are given the 2-step task and told about the structure of the problem, they mostly behave in a model-based way (although the pattern of behaviour is also affected by a model-free effect whereby the fact that a chain of actions led to reward makes it somewhat more attractive). It is difficult to learn action-outcome contingencies when the connections are probabilistic. Without instruction, only a minority of people learn that the first step choice has a common outcome and a rare outcome (in our analogy: that Lake Windermere is usually but not always windy) (Castro-Rodrigues et al. 2022).[3]

Another way to test the models used in reinforcement learning is to give the agent separate pieces of information and see if they can perform a simple inference to put them together. Participants observe that an auditory tone is associated with a particular picture, and later that the picture is associated with reward. When they then hear the tone again, do they expect the reward? The answer is yes, both for humans and mice, with evidence that the hippocampus is involved in making the inference (Barron et al. 2020).

A final example of the way information is stored and computed with concerns value. The task concerned novel foodstuffs. The new foodstuffs were novel combinations of familiar foods: avocado and raspberry smoothie, tea jelly, and so on (Barron, Dolan, and Behrens 2013). Participants had to rate how much they wanted, in the future, to consume each of these new foods, without being given the chance to try them first. Medial prefrontal cortex was again found to encode the future value of the various options. The experiment examined how these new values were constructed out of the representations people already had about the value of the familiar components (tea, avocado, raspberries, jelly, etc.). Barron et al. found that representations of the components and their values were activated in forming a representation of the value of the compound. This effect gradually went away after repeatedly imagining the compound. It seems that, by then, participants could value the new compound directly, without going via the components, even though they had never tried it. Their neural activity was, at that stage, the same as in participants who had had the chance to taste the novel compounds before rating them. The hippocampus was involved in performing the inference from the familiar components to the novel compound. This is evidence of a computation that constructs a representation of a novel compound and infers its value from representations of the value of the components. The decision about what to choose is thus made on the basis of computations performed on a simple model of the task environment.

These examples are a small sample from a very large body of relevant research. What this work shows is that agents that learn about which actions will be rewarded in different contexts store a simple representational model. The model

[3] This despite the fact that the motor system appears to have learnt the connection, reflected in quicker reaction times for the second choice following the common transition.

represents connections between events or states in the world *inter se*, connections between their actions-in-context and outcomes, and representations of how valuable those outcomes are. This information is used to carry out forward planning in order make choices. These types of experiments suggest that people are doing deliberate reasoning with representations of the causal structure of the problem space they are interacting with (Gershman et al. 2014). Devaluation sensitivity shows that this form of reasoning is integrated with the agent's goals and values (not informationally encapsulated), which is in turn suggestive of representations in the global playground. If this is on the right track, then we have a kind of thought-driven planning that makes use of informational models in the cognitive playground, drawing on working memory, and performs deliberate inferences over them. That leaves open questions about representational structure and the nature of the computational process.

Methodologically, it is extremely challenging to find out what representational structures and computational processes are responsible for this kind of planning and inference. However, neural data is increasingly providing an indication of the way information is stored and computed with in these kinds of cases, as we will see in the next section.

4.4 Structural Representations

In the last section we saw a few examples of what is classically described as 'model-based' reward-guided decision-making. This section will look at a suggestion about what the informational models involved in this kind of behaviour may be like. This is almost certainly not the only system responsible for classically goal-directed or 'model-based' behaviour, but it is useful for our purposes because it offers a more detailed characterisation of the representational structures and computational processes involved in one kind of informational model.

The representational structure here is a structural representation, a form of cognitive map. Section 2.2 introduced the spatial cognitive map in the hippocampus and wider medial temporal lobe, arguing that it forms a structural representation of relations between locations in the organism's environment. In this section we will see that the same neural areas are involved in constructing non-spatial structural representations of more abstract relations. There is no indication that the pre-play and replay phenomena observed in hippocampal place cells are a form of working-memory-dependent deliberate inference. (The rapid timescale suggests that they are not: Liu et al. 2021.) Deliberation can nevertheless rely on the outputs of inferences performed in the cognitive map, for example when I ask myself how to get back to the library when I am in the pub. The same may apply to the more abstract cognitive maps discussed in this section.

We saw in the last section that the hippocampus is involved in inferring a novel sound-reward association and in inferring the likely reward value of a novel combination of known foodstuffs. These are inferences that rely on memories of (non-spatial) relations between objects or events. Whittington et al. (2020) have constructed a computational model that links these kinds of task. Their model explains many aspects of the neural activity recorded in the medial temporal lobe (in particular in the hippocampus and entorhinal cortex).[4] The model is trained on sequences of sensory inputs that have an underlying structure. For example, the inputs might be generated by moving around the nodes of a two-dimensional grid (Fig. 4.1). The critical feature of the computational model is that it learns the underlying structure independently of the particular objects that populate the structure. For example, when it sees the chair (see Fig. 4.1), it represents that the current situation lies at a particular node in the abstract structure and also that there is a chair at that location. The node structure can be used to represent another environment that is also arranged in a two-dimensional grid, but where a different set of objects are encountered in moving around the grid.

The structure learnt by the model can be re-deployed in different tasks in which the stimuli encountered have an underlying two-dimensional grid structure. The model predicts the activity of many kinds of cells in the medial temporal lobe, including place cells in the hippocampus and grid cells in the entorhinal cortex. It also explains the finding in rats that spatially-tuned place cells 'remap' when a rat is moved to a new spatial environment. Whittington et al. (2020) show that this is plausibly because a structural representation of the underlying spatial

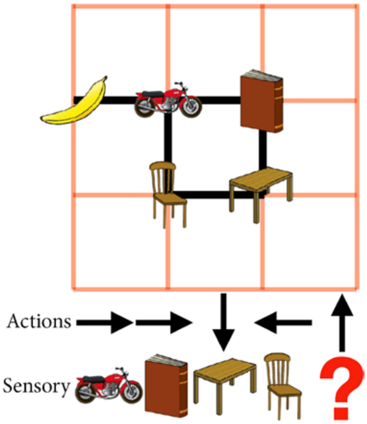


Fig. 4.1 In Whittington et al. (2020), transitions between objects occur as if the objects lie on a two-dimensional grid. For example, having seen a chair on the screen, pressing the up key will produce a motorbike next.

[4] See Mok and Love (2019) for a different account of grid cell and place cell activity.

structure is carried across between environments, being instantiated in different features (landmarks, geometry, smells) in different environments. They also show how the structure learnt by the model can generalise to a new task that is not entirely spatial. In this environment a reward occurs periodically every fourth lap round a circuit. The relation of the animal's current activity to when a reward will be delivered is a non-spatial relation—it depends on how many laps they have completed. The model ends up representing the number of cycles between the current lap and reward, and does so in a way that is consistent with neural recordings from mice performing the task.

A related team found evidence in humans for neural representation of the abstract structure of the way different stimuli-reward contingencies were related (Baram et al. 2021). Participants learn that, in one environment, although rewards are changing, two different stimuli always share the same reward probability; in another environment, the rewards for two stimuli are anti-correlated. The identity of the visual stimuli varied independently of the relational structure (correlation vs. anti-correlation). Baram et al. (2021) found that the representations evoked in entorhinal cortex were preserved when tasks shared the same abstract structure but not when the underlying structure changed. This is consistent with the Whittington et al. (2020) computational model of the way the hippocampus and entorhinal cortex form a structural representation of the environment. Schwartenbeck et al. (2023) found that the same brain area is involved in representing relations between simple geometric building blocks (on top of, beside) as they are used to construct 2D figures. It is not yet known if there are limitations on the kinds of abstract relations that this system can represent or whether it has to build on pre-existing conceptual knowledge of a domain.

Using a different task, Park, Miller, and Boorman (2021) also found evidence that the abstract structure of a non-spatial problem space is encoded in a grid-like code in entorhinal cortex. In their task participants had to learn the way sixteen individuals were organised into a social hierarchy (Fig. 4.2). The sixteen individuals differed in both competence and popularity. On day one participants learnt the relationships between people who differed by one rank in competence, on day two, by popularity. Participants were tested on longer-range relationships which they had not seen in training, but that could be inferred. When participants took these decisions, a grid-like representation of the relational structure was activated in entorhinal cortex. This is plausibly a structural representation of the network of social relations, allowing participants to infer relations of competence or popularity that they did not observe in training. When participants were asked to choose a partner to collaborate with, medial prefrontal cortex activity reflected the 'value' of the partner they selected, a combination of competence and popularity.

Structural representations of space in the medial temporal lobe have been studied extensively for fifty years, delivering a reasonably good understanding of

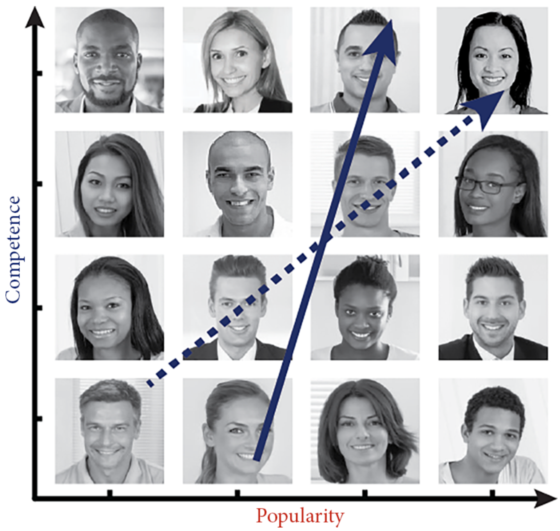


Fig. 4.2 In the task used by Park et al. (2021), participants had to learn how the people pictured were organised along two separate social dimensions, competence and popularity.

how information about spatial locations is stored and computed with (§2.2). The just-mentioned results suggest that information about other kinds of relational structure is likely to be stored and computed with in the same way. The medial temporal lobe is involved in representing the relational structure that obtains across an array of entities (locations, objects, people). It is computed with to infer new relationships. For our purposes, these results illustrate a family of informational models in which relations are represented by means of a structural representation. Computations performed over the structure allow the agent to infer new relationships, estimate the value of options, and take decisions. Unlike a circuit of organized representations linked by learned content-specific transitions, a structural representation enjoys the benefits of compositionality, albeit of a special-purpose kind. A relation between vehicles has specific representational significance. It can be combined and recombined with representations of any of the available relata (as in Schwartenbeck et al. 2023). Content-specific computations allow inferences to be made about these relata, independently of which particular relata are involved.

The cognitive spatial map is usually described as domain-specific (specific to locomotion in space). Here we see that the same kind of structural representation can be deployed in other domains too. But it is still special-purpose in the sense that it turns on making use of a structural correspondence between neural representations and relations of interest in the world. The experiments of Baram et al. (2021) and Park et al. (2021) suggest that the model is connected to participants' conceptual understanding of the problem. It could even turn out that concepts are what labels nodes of the structure. The evidence discussed does not show whether computations over the map are encapsulated from the thinker's wider beliefs; nor whether they occur automatically or require deliberation/working memory. The experiments show that at least the outputs of inferences over the model can be represented in the cognitive playground and relied on by deliberation. Inferences may take place locally, over parts of the model piecemeal, although computational models of the cognitive spatial map show that this kind of structural representation could also support non-local computations.

Finally in this section, I will mention evidence of a simpler form of structural representation which shares these features. The examples so far involve structures with more than one relation: two different spatial dimensions (north/south and east/west, or up/down and left/right), two different social relations (competence and popularity), or two different dimensions on a grid. In a simpler form of informational model items are represented as related by a single relation. For example, in an experiment performed by Nelli et al. (2023) participants learnt a novel relation between a random collection of objects: 'brispiness'. The objects are represented by patterns of neural activation in posterior parietal cortex and dorsomedial prefrontal cortex. We can consider the high dimensional neural activation space defined by plotting the activity of each neuron along a different axis. Regions in this space are vehicles of content and directions in this space can have semantic significance. Semantically-relevant dimensions will not usually align with any axis (with the activity of a single neuron). In this experiment, as a result of learning, objects become arranged in neural activation space along a dimension that reflects their relative brispiness. This arrangement is reflected in choice behaviour (accuracy and reaction times). To look at inferences involving this relation, participants are initially taught the relation only within two disjoint sets of objects: the relative brispiness of objects in set A and, separately, the relative brispiness of objects in set B. When they are later given comparisons at the boundary which show that the objects in set B are all brispier than those in set A, neural activation space rapidly reconfigures so as to arrange all the objects along a single elongated dimension. Subjects can then make correct relational judgements when one object is from set A and the other is from set B.

Many experiments look at a single relation of similarity-dissimilarity. For example, the scenes people see in movies elicit neural activations that form a similarity space reflecting the similarities between the scenes (Huth et al. 2012). The same is true for the words people hear when they are listening to stories (Huth et al. 2016). In some cases, at least, these are not just reflections of similarity but appear to be the basis on which people judge similarities between objects. For example, Charest et al. (2014) asked people to judge the similarities and differences between an array of different objects. Participants' individual, idiosyncratic similarity judgements were predicted by similarities and differences in the particular patterns of neural activity in inferior temporal cortex elicited by seeing the objects. Experience of the objects was arranged into a similarity space that formed the basis of their judgements.

This too is a simple form of structural representation. Objects are represented in a high-dimensional neural activation space (Churchland 1998; Shea 2007; Kriegeskorte and Kievit 2013). When two objects are represented in the same space, the distance between them is taken to be a measure of similarity (or some other relation). The distance measure feeds into downstream processes that depend on similarity and dissimilarity—processes like sorting an array of pictures by similarity (Charest et al. 2014). This is a very simple way in which a relation between represented objects (similarity) is represented by a relation between vehicles (distance in activation space) and then used in downstream computations. As such, these cases too exemplify a simple form of informational model (again one that would not typically be called 'model-based' according to the standard 'model-based'/'model-free' dichotomy). They have broadly the same kind of features as the examples earlier in the section and can be relied on by deliberate thinking in broadly the same way.

4.5 Relational Inference Tasks

The last section looked at some informational models that rely on structural representation. The results of computations over structural representations are available to deliberation. It was not clear whether deliberation is needed to use the model or whether computations over the model can take place without drawing on working memory. This section turns to examples where deliberation is central. There is a set of experiments in cognitive psychology that task participants with engaging in relational reasoning in order to answer explicitly relational questions. These are also cases where it is plausible that relations are represented by concepts, as part of conceptual representations exhibiting general-purpose compositionality, rather than by a structural representation (e.g. a cognitive map).

One experimental tactic is to examine transitive inference. People are given two premises, 'the car is slower than the train' and 'the tractor is faster than the train' (counterintuitively), and then asked whether the tractor is faster than the car (Andrews 2010). Participants who know that 'faster than' is transitive, and discount their background knowledge, will say that the tractor is indeed faster than the car. This kind of experiment probes people's explicit grasp of relations (and examines whether, in reasoning, they rely on logical inference or background knowledge). A relation can be mentally represented and computed with in these ways without being represented structurally. It may instead be represented by a concept in a thought with language-like structure.

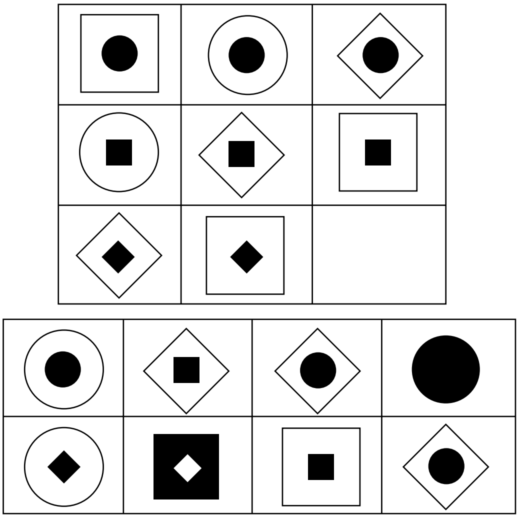


Fig. 4.3 An example of Raven's progressive matrices. Which picture from the lower box completes the empty square in the top grid? From Lovett and Forbus (2017, p. 62).

An extensively-used set of problems that in fact probe relational inference rely on Raven's progressive matrices. These are a series of pictures that form a pattern. The task is to use the pattern to infer the missing picture (Lovett and Forbus 2017) (see Fig. 4.3). The problems are designed so that there is no simple visual pattern. Neighbouring pictures are connected by two relations. To solve for the missing picture one has to identify the two relations and apply them to the missing box. Raven's progressive matrices are widely used as a test of fluid intelligence. They are of interest to us because they rely on the capacity to represent relations and compute with relations. Performing the task depends on working memory and it seems to draw on representing the relations explicitly (Krawczyk 2012; Lovett and Forbus 2017).

An elaboration of the working-memory-dependent ability to represent and reason with relations is the capacity to do so-called analogical inference (Krawczyk 2012). One measure may be familiar from school tests: questions like, 'cat is to kitten as dog is to…?' (puppy) (Goswami 2001). The same capacity can be tested using pictures rather than verbal materials (Markman and Gentner 1993; Gentner and Maravilla 2018). For example, participants have to match the boy reaching for food in one scene with the dog reaching for food in another scene (Krawczyk 2012). A boy in the second scene (perceptual match) acts as a distractor (see Fig. 4.4). Representing the relation depends on understanding the whole scene rather than tracking any simple perceptual cue. The ability to represent relations in a way that enables such inferences is a cognitive achievement in

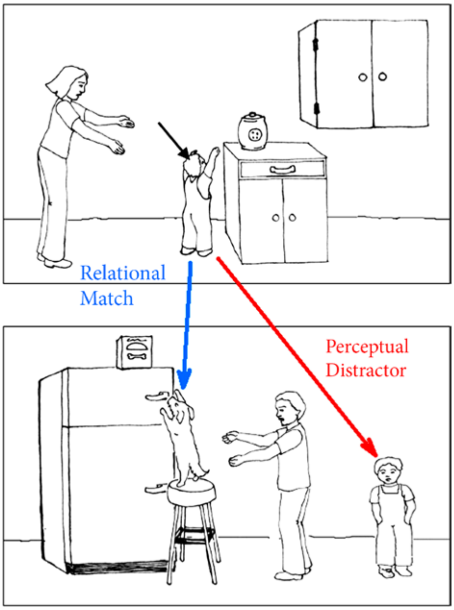


Fig. 4.4 To successfully perform the analogy task, participants had to match the boy in the top picture with the dog in the bottom picture, and not with the same boy in the bottom picture. From Krawczyk (2012, p. 16).

childhood (Goswami 2001) and is much more developed in humans than in other animals (Penn et al. 2008).

This large body of work shows that people are able to explicitly represent relations like X is the *offspring of* Y and W is *reaching for* Z, and to reason with these relations. In what sense is this analogical if it is the same relation in each pair (cat is to kitten as dog is to...)? The analogy exists at the level of the relata. *Kitten* is analogous to *puppy*, in that they are both offspring. In other cases, however, it is the relations themselves that are analogous. For example, there is an analogy between the way too many people streaming movies slows down the internet, and the way too many people coming home from work slows down the traffic (Silliman and Kurtz 2019). By using a subtle change detection task when people read these kinds of narratives researchers show that some people do, and some people do not, represent the analogy between the two situations.

What kind of informational model are people using to perform these kinds of relational and analogical inferences? The experimental data do not offer us a definitive answer, so I will focus on a plausible account that offers a contrast with the last section. This is LISA ('Learning and Inference Schemas with Analogies'). LISA combines distributed connectionist representations of the relata, learnt from perceptual features, with explicit representations of relations, implemented by dedicated units in the model architecture (Hummel and Holyoak 2003). To represent the relation of loving, there is a unit dedicated to representing the subject and another for the object. To represent that Jay loves Alex, the unit for being the lover is connected to and activated with the distributed representation of the features of Jay, and the unit for being the beloved is connected to and activated with a distributed representation of the features of Alex. If this representation is activated and the relata are mapped to a second pair of individuals, Chris and Sam, respectively, the model will perform an analogical inference: it will infer that Chris loves Sam. LISA is a model of how people can use relational knowledge to solve analogical inference problems. It does not rely on structural representations (where a relation on representational vehicles represents a relation). It has dedicated non-relational vehicles for each relation represented, together with a way of binding them together to form complete contents. So it exhibits a limited form of the kind of compositionality found in conceptual thought.

Since LISA has dedicated units for each relation (and each role in each relation), its performance does not scale well to representing a realistic number of relations. The semantic pointer architecture (SPA) of Eliasmith (2013) avoids this problem by compressing the distributed representations that result from combining relata and relation (see also Blouw et al. 2016). Relata (objects, people, etc.) are represented by distributed patterns of activation across units registering perceptual features. Higher levels in the processing hierarchy compress these patterns into simpler patterns that capture statistical regularities. The input representations can be reconstructed from the deeper representations, but not wholly reliably. These compressed representations are combined using vector convolution techniques (inspired by Smolensky and Legendre 2006).[5] Each relation (e.g. loving) is represented by a dedicated vector and is combined with two other vectors (e.g. for Jay and Alex) in a way that distinguishes between subject and object. The relata can be 'extracted' from the compressed combined vector: there is one vector operation that outputs the lover and another that outputs the beloved. Compression means that these extracted vectors are only imperfect reconstructions of the inputs. It is this compression that allows the model to scale effectively to a large combinatorial space of possibilities. But to the extent that it does succeed in capturing and inferring with relational knowledge, that is because it works with a dedicated non-structural representation (distributed activation pattern) of each relation, combined in a way that approximates general-purpose compositionality. This form of compositionality allows the system to engage in multi-step inferences to solve problems like the tower of Hanoi puzzle (Eliasmith 2013, pp. 191–8).

These computational models offer detailed proposals for the kinds of informational models that people are using in cognitive psychological experiments which

[5] Such compositions may be induced by learning in an appropriate deep neural network, without being built-in in advance (McCoy et al. 2019).

ask participants explicitly to draw conclusions about relations. Relational inferences are performed with informational models in which relations are represented by constituents that combine in language-like ways. Plausibly, then, relations are represented by concepts. These support computational processes that are less content-specific, for example turning only on the transitivity of the relation, or on an analogy between relations, transitions that are more towards the content-general end of the spectrum. This is modelled in LISA by a way of performing analogical inference that does not depend on the specific content of the relation or relata involved. In both LISA and SPA, computations involving relations and relata are local, taking place piecemeal over activation vectors; but these models are compatible with a non-local (and content-specific) process of parallel constraint satisfaction being responsible for recognising instances of the relata and the relations. LISA and SPA do not speak to our other distinctions. In the psychological experiments, relational inferences seem to depend on deliberation. It is not clear whether they are informationally unencapsulated, nor whether the inferential steps, or just the conclusions, are found in the cognitive playground. However, the experiments and computational models together do point to a family of informational models with a different profile, across our features of interest, than those discussed in previous sections.

4.6 Using One Relation to Stand for Another

In this section we look at cases where the thinker uses one relation to represent another. This is not simply a matter of structural representation. In a structural representation, a relation on representational vehicles (e.g. co-activation; distance in activation space) represents a relation on the entities represented by those vehicles (e.g. spatial proximity; similarity). In the cases in this section the thinker is representing one relation (e.g. ordering on the page) and using that represented relation to represent something else (e.g. relative academic achievement). The initially-represented relation could be represented structurally or by means of a freely-recombinable concept.

Consider an example. I give you some facts about the heights of four people, telling you that I will ask you some further questions about them: Leyla is taller than Rishi, Rishi is taller than Alex, Jane is taller than Rishi, and Leyla is taller than Jane. A natural way to tackle the problem is to record the facts I've given thus:

Leyla		Leyla		Leyla
				Jane
Rishi	→	Rishi	→	Rishi
		Alex		Alex

If I then ask, 'Is Jane taller than Alex?', the answer is easy to read off. You can rely on the correspondence between the *above* relation between names on the page and the *taller than* relation between people represented. The arrangement of names on the page is a structural representation of the height relations between the people named. Now suppose I do the same thing in my head using visual imagery. Then I am using a representation of the names being above and below one another in space as a representation of relative height. I am using one represented relation (represented visually) to represent another.

In this example the way the relation *taller than* is represented is independent of the way the individuals are represented (by names). I could instead have represented the individuals with scale drawings, aligned at the feet. The *taller than* relation would then be represented by spatial relations between the tops of their heads on the page. This diagram would also make it easy to read off facts about who is taller than whom. But the way the relation is represented would be tied to the way the individuals are represented. The example I gave using names exhibits role-filler independence. We saw the same in the example of analogical inference in the last section.

When I make inferences about the heights of Leyla, Jane, Rishi, and Alex using my simple diagram, the transitions do not depend on the identity of the people but do depend on the content of the relation (*taller than*). They are non-content-specific as to relata and content-specific as to relation. In fact, the inferences are only making use of the transitive nature of the relation. It is because *above* on the page is transitive that it can be used to stand for *taller than*. So this way of using one relation to stand for another also lies towards the content-general end of the spectrum. Penn et al. (2008) argue that non-human animals can only represent relations in ways that are tied to the ways they represent the relata perceptually, not exhibiting role-filler independence. By contrast, our example above demonstrates the flexibility that comes with the human capacity to represent relations in ways that exhibit role-filler independence.

A familiar example of using one relation to stand for another is reasoning with Venn diagrams. Spatial relations of containment and overlap are used to stand for set-theoretic relations of inclusion. Inferences about membership of the sets can be performed by seeing how areas on the page relate to one another in space. We represent $A \subset B$ with a circle for A wholly contained within a circle for B, and $B \subset C$ with a circle for B wholly contained within a circle for C. It is then easy to read off $A \subset C$ from the fact that the circle for A lies wholly inside the circle for C. In a similar way, if areas of a Venn diagram are taken to represent probability density, then probabilistic relationships can be inferred from the diagram, for example $P(B|A)$ is the ratio of the B area within A to the overall size of A, so is equal to $P(A\&B)/P(A)$. In reasoning with a Venn diagram, the thinker takes spatial relations, which the thinker represents visually, to stand for mathematical relations. The Venn diagram then becomes, derivatively, a form of structural representation. It is used in deliberation in a way that depends on the structure of space corresponding to the mathematical structure of interest.

In these examples the thinker relies on a structural representation outside the mind. This does not settle the question of what kinds of representations and processes are at work in the mind. That depends on how spatial properties of the diagram are mentally represented—organized, structural, and/or language-like—and whether computations on them are local or non-local. So the issue is rather subtle. But we can say that the way the external representation is reasoned with is content-specific as to the relation (but not strongly so). They exemplify of an external version of what we saw in sections 4.4 and 4.5 above: reuse of a relational structure to represent a number of different relations. Inferences made using the external structure here are carried out deliberately, relying on working memory, not automatically. It seems plausible that inferences take place within the cognitive playground and are relatively unencapsulated, but further evidence would be needed to settle that definitively.

4.7 Models Built out of Concepts

Finally, I want to turn to a range of cases where the informational model is built out of conceptual representations. While concepts showed up as potentially involved in some of the informational models in earlier sections, for example as labels for the relata in a structural representation, the informational models in this section are cases where researchers take the relevant knowledge structure to be a conceptual representation.

The field of core cognition offers one set of examples. These are domain-specific knowledge systems for thinking about physical objects, space, number, and agents. Each contains conceptual representations, which are thought to be innate (*sensu* unlearnt), together with a set of basic principles for identifying entities in the domain and reasoning about them (Carey and Spelke 1996). It is a contested empirical question whether the mind contains systems of core cognition of this kind, and especially whether they are innate. The behavioural evidence used to argue for core cognition does derive from a very substantial body of careful research in developmental psychology. These results call for explanation, and the hypothesis advanced by advocates of core cognition is a good candidate. For our purposes the important issue is not innateness, but which features (structure, process, etc.) core cognition exemplifies, if the hypothesis is correct.

In the domain of number, Susan Carey argues that, in addition to an analogue magnitude system for keeping track of the approximate number items in large arrays of objects, the child's system for tracking individual objects, the object file system, affords exact representations of the cardinality of small sets of objects (Carey 2009). The 'parallel individuation system' can keep track of one, two, three, or four objects by mapping objects one-to-one to sets of object files. Further, it can perform exact addition and subtraction amongst these sets. These innate conceptual roles give the child a set of built-in ways of reasoning about number. Furthermore, reasoning about the analogy between the sequence of counting words ('one', 'two', 'three',...) and the process of adding one to a set of object files allows the child to acquire the richer set of conceptual roles constitutive of grasp of the concept of natural number—in particular the principle that the last word in any count sequence represents the number of objects in the counted set.

In our terms, both the parallel individuation system and the natural number system are informational models—models of the domain of quantities. Dispositions to make inferences between concepts encode information about the domain, for example that adding one object to a set of n objects produces a set whose cardinality is the natural number which is the successor of n. These are local, somewhat content-specific computations taking place over representations constructed out of concepts (presumably exhibiting the general-purpose compositionality of conceptual representations). They are applied to the world in a way that is automatic and relatively informationally encapsulated (Carey and Spelke 1996), although their outputs must be available to deliberation (including in particular to the inference by which children notice an analogy and thereby acquire the concept of natural number).

Other examples include the concept of a physical object, with principles for tracking physical objects and their interactions (cohesiveness, contact, occlusion), and the concept of an agent. These concepts are applied to the world in a way that is fast and automatic, based on spatiotemporal analysis of perceptual input. To this extent, they are perception-like, and unlike concepts that are applied based on intuitive theories (or scientific theories). But they are supposed to be fully concepts, representations that figure in cognition, freely recombine with other concepts, and come with a set of in-built conceptual roles.

Moving beyond core cognition, the way people represent natural kinds is usually taken to be based on an informational model consisting of conceptual representations. People store in long-term memory a collection of facts about instances of a kind (e.g. dogs), information which they rely on to categorise objects under the kind concept and to make inferences about them. Much of the information may be represented explicitly, in the form of semantic memories, although some could be implicit in content-specific dispositions to move between representations, as we have seen (§3.4). What is characteristic of natural kind concepts is that they come with the assumption that members of the kind share an underlying property that is responsible for the surface features of the kind. Much studied developmentally, there is a stage where children are shown to prioritise internal properties or essence over surface features in categorising members of the kind

(Keil 1992). Although not innate, this is an informational model which, like examples from core cognition, is supposed to include patterns of inference involving a concept that are specific to its domain of application (here, to natural kinds).

A more contested example concerns cheater detection (§3.2). Experiments with the Wason selection task famously showed that people are better at testing the truth of a conditional in some domains than in others. An initial hypothesis was that we have a cheater detection module or a domain-specific set of assumptions concerning norm compliance. If that were correct, then people would have a domain-specific informational model involving normative concepts, equipped with dispositions to make certain content-specific inferences. If instead the advantage is just due to familiarity with the subject matter, then the effect would be evidence of a more diffuse family of informational models that encode expectations about the evidence for and against conditionals in each familiar domain (Cox and Griggs 1982). A third possibility is that the choice pattern displayed by participants in Wason-type scenarios is the result of a domain-general way of gathering data for testing an inductive hypothesis (Oaksford and Chater 1994). In that case, people have a set of dispositions for reasoning about if-then generalisations but these are not particularly content-specific.

There are other cases where people are plausibly working with informational models that have a wide sphere of application. One example is the 'mental models' long championed by Philip Johnson-Laird to account for deductive reasoning with conditionals (Johnson-Laird and Byrne 2002). These are something like collections of mental sentences recording a list of possibilities compatible with a linguistically-described scenario. This fits within the philosophically familiar paradigm of an informational model consisting of a set of explicit conceptual representations entertained in conscious deliberate thought. The system has distinctive features that are supposed to account for patterns in the way people perform inferences, for instance that false clauses are not represented in a mental model. To the extent that the operation of mental models has built-in assumptions—processing dispositions—these are applicable whatever the subject matter being reasoned about, and hence are content-general and broadly logical.

Two other cases of mental models enjoying a wide sphere of application are those for causation, and for probabilistic reasoning. In the case of causation, work on core cognition suggests that there is a domain-specific appreciation of some causal relations as part of our automatic perceptual or perception-like processing of physical objects and their interactions (Carey 2009). Michotte-style experiments show that interactions between simple object-like images can be perceived very differently—as a causal 'launching' or a non-causal 'passing', for instance—based on quite subtle spatiotemporal characteristics of the display. Those effects are based on an informational model that appears to be specific to the domain of object causation. It is also likely that humans have a more domain-general means of identifying causes and effects and learning about them. This is an area where there are many different theories, from nativist to empiricist to constructivist.

To take just one prominent example, Gopnik et al. (2004) argue that children have a specialised cognitive system that allows them to recover an accurate causal map of any aspect of the world. The causal learning system makes substantive assumptions about how patterns of correlation between variables reflect causal relations; and in particular about how to infer causation from interventions on a variable. The output of the causal learning mechanism is a representation of the network of causal relations between a set of events. Gopnik et al. suggest that what the child learns is a graphical causal model (Pearl 2000). There are thus two kinds of informational model on display here. First, there is the informational model of causation embedded in the learning mechanism. Second, there is the graphical causal model of those so-learned causal relations, for example how the various wheels and levers on a toy work. The former is relatively domain-general, applying to causal relations in any of the spheres which people interact with or communicate about. Nevertheless, its substantive assumptions are content-specific, in that they are suited to the realm of causation. (That is to assume that the learning system's assumptions are implicit in its inferential dispositions, rather than based on explicit representations of causal principles.) Being applicable so widely, they lie some way towards the less content-specific end of the spectrum. The Gopnik et al. model does not make commitments about our other features of interest, although the experiments make it plausible that the process is deliberative, not wholly informationally encapsulated, and involves concepts. It also seems plausible that the substantive causal models that are learnt can then be represented in the cognitive playground.

Goodman, Tenenbaum, and Gerstenberg (2015) advance a different account of how people infer causal relationships. Their account has even more general application. It accounts for probabilistic reasoning in general. Of most interest, for our purposes, is their hypothesis about representational structure. They argue that humans perform inferences in a probabilistic language of thought. This allows us to form informational models that consist of representations structured out of concepts using general-purpose principles of composition. These conceptual representations are used to represent probabilistic information in the form of a probability distribution over world states. Goodman et al. prescind from making claims about psychological processing, but it seems that their concept-based informational models are used in deliberation, in patterns of inference that are broadly logical, applying general principles for reasoning with probabilities.

Taking stock, the cases in this section illustrate a variety of ways that informational models can be formed out of conceptual representations. They cover a broad range of cases, from agents, through natural kinds, causation and number, to logical reasoning and probabilistic inference. Being applied to such broad domains, the representational scheme takes advantage of the flexibility of general-purpose compositionality. In most cases, working-memory-dependent deliberation is required. To the extent that computational details are offered, the computations have been local. But they also exemplify both sides of many of our other features of interest: content-specific and content-general transitions; informationally encapsulated vs. not; automatic and deliberate modes of operation. An account of thinking with concepts needs to be able to accommodate this diverse range of features.

4.8 Conclusion

Chapters 2 and 3 used various empirically-supported cases to ground distinctions between different aspects of semantically-significant representational structure, and between content-specific and content-general computational processes. The examples canvassed in this chapter illustrate that these features are distributed in diverse ways across informational models of different kinds. They also show that the idea of having an informational model comes in degrees, with no bright-line distinction between representing the world in model-based and model-free ways. Many of the informational models we discussed used structural representations, but others were formulated conceptually, and some just deploy organized representations processed in content-specific computations suited to a particular domain.

Deliberation can make use of all these different models, either by running simulations and relying on their outputs, or by making inferences with aspects of the model directly. This raises the question of how deliberate conceptual thought can rely on and integrate with informational models of these diverse kinds. That is perhaps straightforward for informational models built out of conceptual representations, but less obvious when other kinds of representational structures and computational processes are involved. Thus, the diversity of these informational models presents a puzzle—a puzzle which the next chapter aims to address. There, I develop a framework that, while being flexible enough to encompass all these cases, aims also to be detailed enough to be genuinely illuminating.

Chapter Summary

4.1 Ways of Storing Information

Prospection or forward planning relies, not just on semantic memories, but on diverse informational models of other kinds—illustrated by examples presented in this chapter. The models that are involved in model-based decision-making—if that is simply the converse of model-free decision-making—are in fact rather heterogeneous. An *informational model* is a collection of interconnected representations, of any kind, that can support planning, prospection, or other forms of inference which derive novel conclusions about what to do or what is the case.

(p. 88)[6] As well as semantic memories, deliberation brings information and conclusions drawn from diverse informational models of other kinds into the playground of thought. This variety sets up the problematic to which the framework in the next chapter (Chapter 5) is a response. There is no simple hierarchy of complexity; the cases illustrate different combinations of representational structure, computational process, and the other properties discussed in the first three chapters (listed).

4.2 Information in Domain-Specific Transitions

(p. 89) One way of storing information is in dispositions to make content-specific transitions between representations (examples from visual processing). Although not standardly described as constituting a model, it should be, and to the extent that visual processing can be used in offline simulation, we can consider it to be a generative model. These are informational models only in a thin sense, since the information is encoded merely implicitly, relied on by running simulations.

Apparently model-free behaviour is illustrated by performance on the trap tube task, where animals can learn to avoid a trap and extract food from the tube, but seemingly fail to appreciate the causal structure. Model-free reinforcement learning laboriously learns how to act to obtain reward, without learning what rewards are obtained, nor how. (p. 90) Model-free decision-making has been well characterised, including the computations involved in learning action values (based on reward prediction error). The system calculates what to do based on representing the expected value of performing available actions, but does not really know anything about the structure of the environment. (p. 91) Motor control uses an informational model, even in direct-throughput mode. Outputs of simulations with these systems can be relied on in deliberation, but their properties (listed) contrast with the characteristics of deliberation with conceptual representations.

4.3 'Model-Based' Tasks

(p. 92) This section looks at 'model-based' decision-making—where the agent uses some kind of representation of the causal structure of the environment and of how

[6] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

their actions will affect what happens next. Examples: two tasks in which participants were found to encode and decide on the basis of the long-run prospective value of their choice.

(p. 93) The classic test of whether a subject represents the way actions produce outcomes, independent of the varying value of these outcomes, is to look for one-shot learning when rewards are devalued. One study in mice found patterns of neural activity representing outcome values and, separately, action-outcome contingencies. A much-used experimental test of model-based decision-making is the 2-step task (described). (p. 94) People who are told the structure of the problem (which is hard to infer) behave in a primarily model-based way. Another test is to see whether subjects can chain together a stimulus-stimulus association with a stimulus-response association. A human study found evidence that a representation of the value of a novel foodstuff is constructed, before the compound is tasted, by means of an inference from the values of the ingredients.

These examples are evidence of a kind of working-memory-driven planning that makes use of informational models in the cognitive playground and deliberates with them. (p. 95) In the next section we will see that neural data can tell us about the representational structures and computational processes responsible for this kind of model-based planning and inference.

4.4 Structural Representations

This section looks at a case where we have a more in-depth understanding of the informational model responsible for one kind of model-based behaviour. The cognitive map in the medial temporal lobe can represent more abstract relations, making inferences whose outputs can form the basis of conscious judgements. (p. 96) Whittington et al. (2020) propose a computational model of how inferences are made about problems that conform to a two-dimensional relational structure. The model accounts for neural activity, including the remapping of place cells when moving to a different spatial arena and, in mice, neural activity which represents a non-spatial relation. (p. 97) Converging evidence was found in humans for representation of abstract relational task structure in the same brain area. A grid-like code in the same brain area was found to represent learnt social relations between people—who is more popular, or more competent, than whom.

These cases illustrate a family of informational models in which relations between locations, objects, or people are represented by means of structural representations, inferences over which allow the agent to infer new relationships, estimate the value of options, and thereby take decisions. (p. 98) Computations over these models are content-specific; how they exemplify our other properties of interest is unclear.

(p. 99) A structural representation may involve only one relation; this may be represented by a dimension in neural activation space (empirical example). A much-studied one-dimensional relation is similarity/dissimilarity; the similarity relations between neural patterns of activation are found to form the basis of people's judgements of similarity. (p. 100) Using distance in activation space to represent similarity is a structural representation, and counts as a simple form of informational model (although the behaviour would not typically be called 'model-based').

4.5 Relational Inference Tasks

This section looks at cases where people are explicitly tasked with deliberating with relations. An example is transitive inference, which can be performed using concepts of relations instead of with a structural representation. (p. 101) Working with Raven's progressive matrices, a widely-used test of fluid intelligence, relies on the capacity to represent relations and reason with them. An elaboration of the working-memory-dependent ability to represent and reason with relations is the capacity to do analogical inference (explained). (p. 102) These inferences involve drawing an analogy, either at the level of the relata, or between relations themselves.

A possible model of the way people represent relations in these tasks is LISA—which does not use structural representations, but non-relational vehicles representing relations, bound together with general-purpose compositionality. (p. 103) A more sophisticated model, the semantic pointer architecture, can represent more relations, using vector convolution to combine representational constituents in a way that approximates general-purpose compositionality. In terms of the other properties of interest (end of §4.1), some of the computations here are somewhat content-general; also working-memory-dependent and local (albeit compatible with recognition of the relations and relata involving parallel constraint satisfaction, as in an artificial neural network).

4.6 Using One Relation to Stand for Another

(p. 104) This section looks at cases where the thinker uses one relation to represent another. If I visualise a list of names so as to order people by height, I am using one *represented* relation (of spatial properties) to represent another relation (taller than). (p. 105) This way of representing heights exhibits role-filler independence (unlike other examples). Using one relation to stand for another, relying only on the fact that both are transitive, is a move towards the content-general end of the spectrum. In a Venn diagram, spatial relations on the page are taken to stand for set theoretic inclusion; it thereby becomes a structural representation. (p. 106) These examples involve reasoning with external relations, deliberately, in the cognitive playground.

4.7 Models Built out of Concepts

A final set of cases are where the informational model consists of a collection of conceptual representations. Systems of core cognition offer good examples (irrespective of whether they are innate). The parallel individuation system gives the child a way of reasoning about small numbers of objects, which is then enlarged into a system for representing all the natural numbers. (p. 107) These are informational models, constituted out of concepts and reliant on content-specific computational transitions. The representations of core cognition are applied rapidly and automatically, with built-in content-specific transitions; but they are supposed to be fully concepts, displaying general-purpose compositionality and figuring in conscious deliberation.

Natural kind concepts are part of an informational model, one which includes patterns of inference (about underlying properties) that are specific to its domain of application. (p. 108) Cheater detection is contested: it may involve an informational model specific to the domain of norm compliance; or it may be one of a series of models encoding conditionals learnt in familiar domains; or it may reflect a domain-general way of reasoning about if-then generalisations. Johnson-Laird's mental models are language like, displaying general-purpose compositionality and using content-general computations.

Michotte-style perception of causation is based on a somewhat encapsulated informational model that is domain-specific. (p. 109) Proposals about how we infer causal relations (e.g. about how a toy works) rely on principles which, while specific to the domain of causation, have wide application (in a process that appears to be deliberative and to involve concepts). Goodman et al. (2015) offer an even more general informational model: a probabilistic language of thought, used for learning and reasoning about probabilistic relationships in general.

These concept-based informational models exemplify both sides of many of our other features of interest: content-specific and content-general transitions; domain-specific and domain-general spheres of application; automatic and deliberate modes of operation; informational encapsulation and informational promiscuity.

4.8 Conclusion

(p. 110) The idea of having an informational model of the environment comes in degrees and is implemented in a variety of ways, including using conceptual representations, structural representations, and mere organized representations processed content-specifically. This raises a puzzle, which the framework in the next chapter addresses: how can deliberative thought make inferences with, and rely on the outputs of, such diverse models?

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0004

5
Concepts as an Interface

5.1 Reaching Conclusions via Simulations

The book aims to offer a more full-bodied account of conceptual thought than those that just focus on categorisation. A central phenomenon is concept-driven thinking: cognitive processes that reach new conclusions having started with a conceptual thought. Reasoning is one way to do that. In reasoning we move from some conceptually-structured thoughts to others using a domain-general process rather like theorem proving in logic. But much of what we do with concepts draws on informational models of other kinds. Thoughts prompt mental images, simulations of potential actions, feelings and evaluative responses. Those processes in turn cause further thoughts: judgements, plans, decisions, and actions. This chapter is about the way we arrive at concept-involving conclusions via these other types of informational models.

For example, thinkers can arrive at new beliefs via sensorimotor emulation or imagination (Grush 2004). The thinker simulates an action, observes the likely outcome, evaluates it, and thereby decides what to do (Carruthers 2015, pp. 152–60). I am in a furniture shop thinking about whether to buy a new chair. One consideration is whether I can take it straight home in the car. I rotate the chair in my mind's eye to see if it will fit. Probably not. What about with the back seat down? Then I can put the legs in first. If so, it probably will fit. My eventual conclusion—an intention to buy the chair—relies in part on my capacity for sensorimotor simulation and on the knowledge encoded in my sensorimotor systems about what happens to objects when they are rotated. A second example is spatial planning, where the thinker imagines different routes through the environment in order to decide which one to follow. Another is prospection, where the thinker uses the resources of episodic memory to imagine what a certain situation would be like. Simulation, prospection, and imagination are all ways of drawing on special-purpose systems in the service of forming new beliefs—that is, of reaching conclusions expressed in the general-purpose system of conceptual thought.[1]

To illustrate the contrast with reasoning, suppose in an idle moment I start thinking about Paris. That could start a chain of reasoning. I find myself thinking *Paris is the capital of France*. I then recall my 'bucket list' goal of visiting all of Europe's capitals. Knowing that France is in Europe, I reason my way to forming an intention to visit Paris, to go online at the weekend to look for tickets, and so on. I could instead reach the same conclusion by a completely different route. I start by thinking about Paris as before. I contemplate the city and various situations pass before my mind: elegant streets, lively cafés, shady spots in a formal garden on a sunny day—constructed from episodic memories and culturally-transmitted stereotypes. The scenes prompt various emotions and a positive evaluative response. As a result, I form the intention to visit Paris, and so to look for tickets, and so on. Both trains of thought arrive at the same conclusion; the second relies largely on information and processing dispositions encoded in special-purpose systems. Our focus is on the second process, and on the role that concepts play in it.

Often special-purpose systems know things it seems the thinker doesn't. More carefully, a simulation can be relied on to reach an accurate conclusion where the relevant information cannot simply be retrieved from long-term memory or inferred by reasoning alone. One classic set of experiments probed people's expectations about what would happen when a glass of water is tipped sideways. The task was to predict, for various different glasses, when the water would spill out. People who simply reason about the problem usually get the wrong answer. Those who imagine performing the action get it right (Schwartz and Black 1999). Sensorimotor models of the scenario are accurate, doubtless because of a wealth of practical experience. The thinker can probe that knowledge in imagination and arrive at the right answer in conceptual thought. Bascandziev and Carey (2022) showed that children too can learn-by-simulation. Many six-year-olds claim that a single grain of rice weighs nothing at all. They change their mind when they see that a single grain of rice will topple a small see-saw made out of a piece of card. Amazingly, children simply asked what would happen in this

[1] There are different brain mechanisms for imagining what one might have done and for imagining what one might do (Miyamoto, Rushworth, and Shea 2023). Many psychological processes form representations of more than one possibility, for example representing a range of hypotheses and their probabilities given current sensory evidence, or representing unchosen options when making choices for reward. Here we are concerned specifically with simulating or imagining potential future scenarios (in the service of working out what is the case or what to do).

scenario make an accurate prediction and change their judgement too—just as much as with a real demonstration. Even in children, then, simulation may be an important way that we learn explicit knowledge about the world.

An earlier line of work on 'mental rotation' also provides evidence of reliance on sensory representations to solve a problem (Shepard and Metzler 1971). The task is to report whether one image is a rotated version of another. The simulated images are prompted by visual perception, not concepts, but the process is carried out deliberately and is relevant to our question. These experiments were an early demonstration that representations not directly produced by sensory input, but with the same structure and content as perceptual representations, are involved in performing some tasks. Aphantasics, who lack conscious mental imagery (Keogh and Pearson 2018), seem to perform the task in a similar way, but with non-conscious simulations (Pounder et al. 2022). Lawrence Barsalou and his colleagues have collected a large body of data showing effects of sensorimotor processing on the way conceptual thinking unfolds (Barsalou 1999, 2003). Much of this imagery is unconscious (Barsalou 2009, pp. 1281–2, 1286). Barsalou makes the strong claim that a concept consists in a simulation of a perceptual state. Whether or not that is right, these results show that concept-driven thinking does engage a range of special-purpose systems. As in everyday life, in science too people solve problems by simulating situations and events (Nersessian 2018). They construct and manipulate a model of a scenario using sensory, motoric, and affective representations.

We can ask how it could be that a simulation allows the thinker to discover something new (Gendler 2004). The thinker is not going out into the world to make an observation. They are interrogating their own psychological systems. How does this allow them to uncover new facts? One answer points to encapsulation. Special-purpose systems may contain representations that are encapsulated, so not directly available for use in conceptual thought. Using a system in simulation mode is a way of bringing the information into conceptual thought (Aronowitz and Lombrozo 2020). A complementary answer is that special-purpose systems operate with trained dispositions which implicitly encode assumptions about the world. We saw examples of these content-specific dispositions in Chapter 3. Where information is implicit in a disposition to move between representations it can only be made use of by tokening the representations between which the disposition subsists (Shea 2015). By running a simulation the thinker can effectively rely on the implicitly-encoded information.

This chapter focuses on the role of concepts in this process. We start with a concept-involving thought. This could be a question to consider (whether to marry), a goal to be achieved, a memory to reflect on, a belief whose consequences are of interest, or just a concept (e.g. PARIS). We then deliberate, relying on working memory to hold and manipulate representations in the cognitive playground. Steps in this thought process can go via special-purpose informational models. Through simulation, prospection, or imagination, a representation of a situation or scenario is built up using special-purpose systems, often relying on sensory, motoric, affective, and evaluative representations. Equally, information encoded conceptually (§4.7) can be represented in the cognitive playground, or used by relying on content-specific dispositions between concepts (§3.4). I will use the term 'suppositional scenario' for the collection of representations that, by drawing on informational models, is built up in the cognitive playground. 'Simulation' here is used broadly to include prospection and concept-driven imagination—any process by which special-purpose systems are engaged to work out what is the case or would be the case. Inferences take place within special-purpose systems as we fill out a suppositional scenario and see what follows from it. A concept-driven simulation takes us via suppositional scenarios to new conclusions (e.g. a judgement or an intention), proceeding from conceptual representations via special-purpose representations to further conceptual thoughts.

Section 5.2 makes the case that conceptual thought can drive simulations in the kinds of special-purpose informational models described in the last chapter. Section 5.3 advances a hypothesis about how concepts are able to interface between special-purpose informational models and general-purpose conceptual thought. The slogan is that a concept is a 'plug-and-play' device. Section 5.4 discusses some evidence that concepts do indeed mediate between type 1 processes running in special-purpose informational models and type 2 processes taking place over conceptual representations. Section 5.5 argues that conditionals—explicit if-then beliefs—serve to shift information coded in content-specific dispositions into a format that can act as a premise in reasoning (and the converse). Section 5.6 delves into illustrative models of the way a concept held in working memory operates computationally, considering some analogies from computer science. I use the neutral term 'label' to describe how concepts operate. Section 5.7 explains how concepts should be individuated in my framework. Section 5.8 raises a puzzle for the plug-and-play hypothesis: how does the suppositional scenario constructed in a simulation come to reflect the compositional structure of the conceptual thought that drives it? I sketch a tentative solution. Finally, section 5.9 argues that the ability of concepts to act as this kind of interface makes human thinking an especially powerful way of working out what is the case or what to do.

5.2 Simulations Use Special-Purpose Informational Models

Our focus is the way conceptual thinking can reach new conclusions in reliance on simulations involving special-purpose representations. Concepts act as a crucial interface in this process. While many rightly emphasise the domain generality of conceptual thought (Fodor 1998), this has under-estimated the role of concepts in organizing domain-specific and special-purpose resources. Others have emphasised the domain-specific aspects of conceptual thinking (Prinz 2002; Barsalou 2009) but overlooked the importance of being able to marshal these resources in domain-general ways. Between the two, what has been underplayed is the way concepts provide an interface between special-purpose representation models, on the one hand, and general-purpose compositional reasoning, on the other.

A concept provides access to a rich body of information about its subject matter. Some information is stored in the form of explicit conceptual representations—what psychologists call semantic memories. My PARIS concept gives me access to the belief that *Paris is the capital of France*. That provides material to reason with. Content-general computations in reasoning will take us to new conclusions. While doubtless important, this is far from the whole story. Concepts of concrete objects connect with sensory and motoric features: DOG with how the animal looks, feels, and smells; PENKNIFE with how to open the blade and what kind of actions it affords. Concepts also lead into affective responses, for example the feeling of moral disapprobation that comes with categorising someone under RACIST, which is often also tied up with valence, for example when you value person Y more than person X. These are amongst what Liz Camp calls the *characterizations* connected with a concept (Camp 2015).[2] When a concept-involving thought drives a simulation, a suppositional scenario is constructed out of various characterizations.

I have been emphasising modality-specific information, but representations driven by a preferred modality (e.g. vision) can also be driven by other inputs (e.g. touch) and so should perhaps be considered supra-modal (Calzavarini 2022). Special-purpose systems also encode information in amodal structures (Frankland and Greene 2020; Calzavarini 2022). The cognitive map of the spatial environment is one example. It may be domain-specific, representing the domain of spatial locations, but we have seen that the same or similar relational structures are used to represent other domains (e.g. social hierarchy; Park et al. 2021). For this reason, I consider these systems *special-purpose* rather than domain-specific. They use a representational system or range of representational models suited to some purposes and not others. Another example is the high dimensional activation space in the brain which encodes objects of different types (Huth et al. 2016; Frankland and Greene 2020; Tang et al. 2023). Concepts afford access to the information in these special-purpose systems, often represented structurally, employing compositional principles that are different to the general-purpose concatenation of conceptual thought (as we saw in Chapter 2). These

[2] Camp requires characterizations to exhibit a high degree of interpersonal similarity, which I do not require here.

representations are processed in content-specific ways (Chapter 3), and form informational models of various kinds (Chapter 4).

The capacity for simulation does not seem, in itself, to require the exercise of concepts in conscious deliberate thought. The simulations in medial temporal lobe in the service of route planning are one example (Dragoi and Tonegawa 2011). Non-human animals engage in various other forms of prospection and simulation (Clayton and Dickinson 1998; Passingham 2021; Tomasello 2022, pp. 48–52), seemingly without drawing on a human-like capacity for deliberate concept-involving thought. Our question is how simulation works when it *is* driven by deliberate concept-involving thoughts. More specifically, what role do concepts play? We will see that concepts perform a special job. They interface between general-purpose thinking and special-purpose informational models. Combining the two makes concept-driven simulation a particularly powerful way of drawing conclusions about the world.

5.3 Concepts as Plug-and-Play Devices

In this section I offer an account of the role of concepts in simulation-mediated inference. I argue that a concept is a 'plug-and-play' device (Shea 2022b). The idea of a plug-and-play device is modelled on the way representations are involved in offline computations in special-purpose systems like the spatial cognitive map. That insight can be carried over and applied to concepts; but with one crucial modification, as we shall see.

The mechanism by which offline simulations in a cognitive map are used in route calculation does not involve concepts in our sense. Place cells are active online as an animal moves through its spatial environment. As we have seen (Chapter 2), as a result of experience, a co-activation structure is formed over the place cells which corresponds to the spatial structure of the locations to which they correspond. The location-specific receptivity of a place cell is a considerable achievement, using a range of different cues to register where the animal is located (visual input from any direction, proprioceptive updating, etc.). Even more useful, however, is the capacity to take the array of place cells offline and make use of the co-activation structure over them. This is what happens when, in offline mode, the hippocampus runs through chains of place cells in order to try out different possible routes. The place cells are severed from their input-output connections and 'played with' offline in order to make use of the information encoded in the co-activation structure.

This is a widely-applicable trick, a computational device that can be played out in many places. Representations that are useful because they correlate in a sensitive and specific way with task-relevant features of the environment are also often interconnected in memory. They may form structural representations or representational models of other kinds. The principle we see at work when running a simulation is that representations are severed from their worldly connections, 'played with' in constructing suppositional scenarios and running simulations that make use of their interconnections, and then restored to their online input-output profile so that the conclusions that have been worked out offline can be used to guide behaviour.

The literature on simulation and imagination shows that people are able to engage in this process deliberately, driven by conceptual thought. I start with a thought like, *Will the chair I saw earlier fit in the car?* The concepts involved in the thought are connected to various kinds of information. The concept MY CAR calls up an imagistic representation of the spatial dimensions of the car. I imagine different ways of manoeuvring the chair into it. I then categorise the result: IT FITS or IT DOES NOT FIT. We have thereby arrived back in the general-purpose system of conceptual thought. A few steps of reasoning then lead to the intention to buy the chair. I have arrived there via 'playing with' representations in special-purpose systems.

Concepts act as an interface in this process. They plug into special-purpose systems, driving simulations. But they also plug in at the other 'end'—they plug into the general-purpose compositional structures of conceptual thought. A concept is a plug-and-play device with plugs at both ends. It provides an interface between the informational models and content-specific computations of special-purpose systems, at one end, and the general-purpose compositionality and content-general reasoning of deliberate thought, at the other.

Conceptual thought gives us a general-purpose capacity to use representations offline—both concepts themselves, and the special-purpose sensorimotor, affective, evaluative, and amodal representations which are accessed through concepts. A concept, it seems, gives us the capacity for offline use of any of the special-purpose representations to which it is connected, and a way to manipulate many of the representations of objects, properties, and relations in those systems.

We saw in Chapter 2 that the general-purpose combinatorial system of conceptual thought—whether it is language-based or a separate competence—contrasts with the special-purpose modes of combination at work in structural representations. The representational competence of concept combination is not restricted by subject matter. A thinker who can think *Layla loves Rishi* has the capacity to represent aRb using any of the singular and relational concepts that they possess (meeting the generality constraint). Less recognised is the fact that this normally underpins a second capacity, the capacity to construct a suppositional scenario corresponding to the thought. When thinking recombines existing concepts to formulate a novel thought, that can often prompt a simulation of a novel scenario. Whether the thinker will succeed in simulating a scenario equal to the thought is another matter. They may face 'imaginative resistance' (Gendler 2000; Yablo 2002), or formulate a thought for which no scenario can be imagined (e.g. green ideas sleep furiously), but the literature on metaphor shows that people succeed in making sense of a surprisingly wide range of novel combinations (Camp 2006). Where people do succeed in simulating a novel scenario based on the thought, it is the combinatorial power of concepts that allows them to imagine the new scenario. Conceptual recombination drives a simulation that puts together special-purpose representations in novel configurations. When some kind of sense can be made of the combination, a novel suppositional scenario is the result.

A simulated scenario usually integrates information from many different special-purpose informational models. Suppose I'm at my desk thinking about what to make for dinner (working hard as ever). I remember that there are plenty of tomatoes in the garden. Also mint. A tomato salad with mint? This simulation prompts an evaluative response (not good). Maybe with green beans instead, from that stall on the way home? And some squash. Roast that first? Also fresh peas? No, too much preparation. Perhaps a few cherries from the garden instead, even though they're probably not quite ripe. The picture I'm building integrates information from lots of different special-purpose systems: sensorimotor knowledge of ways of preparing ingredients, olfactory and gustatory simulation of potential dishes, affective-evaluative responses to the imagined results; a spatial map of where things are in my environment; motoric knowledge of the effort cost of various potential actions; and semantic knowledge of the monetary cost of potential ingredients. The suppositional scenario is multi-system and holistic. Concepts allow the thinker to manipulate and recombine specific elements of these scenarios.

* * *

How do concepts give us this capacity? There is much evidence in psychology and cognitive neuroscience for a family of representations that have the adaptability and flexibility required. These are the temporary task-dependent or working memory representations that are usually localised to prefrontal cortex. These representations are variously theorised as representing task-relevant goals and means (Miller and Cohen 2001), adaptively coding for information that is specifically relevant to current concerns (Duncan 2001), or flexibly holding items in working memory (Bouchacourt and Buschman 2019). These accounts share a commitment to a flexible capacity for tokening representations that are adapted to the current context or task, and which serve to give access to and manage information in the mind's special-purpose systems.

I argue that these representations are concepts. Concepts, recall, were introduced as a type of mental representation. I pointed to canonical instances: freely-recombinable constituents of the thoughts that occur during deliberation (§1.4). Temporary representations in working memory which have these characteristics are concepts. To token a concept is to token a representation that can be combined and manipulated in working memory, and that is connected to a wider store of information in long-term memory. What is it to be 'connected'? I will have much more to say about that in due course, but the basic idea is that the label in working memory can access and hold online representations in other systems so that computations are performed with them. A slightly stronger idea is that working memory is part of an executive system that can manipulate representations accessed via working memory labels and exercise control over how they are processed. I can endorse this stronger claim, where it is empirically well-supported, but the slightly weaker thesis is sufficient.

From one perspective, a working memory label looks like the wrong kind of entity to be a concept. A concept is often supposed to be a store of information that is used in categorisation and inference (Machery 2015). At the very least, it was supposed to be a mental word for X that shows up every time one thinks about Xs. Working memory labels appear to be too temporary and task-dependent to *be* concepts. Here it is crucial to distinguish between two ways theorists have talked about concepts. On the one hand, they are a representation type, a representation of X, that is tokened in occurrent thinking. On the other hand, they are a store of information about Xs that is used in categorisation, inference, and other cognitive processes. Many theorists have assumed that the same representation plays both roles. If so, it would be innocuous to talk about concepts both ways—to assume that concepts have both these properties. We have seen, however, that to think clearly about concepts, it is crucial to separate these two roles. Furthermore, the data suggests that the two roles dissociate in practice: the representations in working memory displaying general-purpose compositionality that we reason with in deliberation are not themselves bodies of information that are used in cognition. That may seem like a fussy distinction, since working memory labels provide access to bodies of stored information. But it turns out to be important that the collection of stored information is separate from the entities over which online deliberation occurs.[3] Later I will argue that, although they do not correspond to a single mental word (vehicle type), temporary working memory labels deployed on different occasions can count as tokens of the same concept in an individual thinker (§5.7).

My claim is supported by evidence that the capacity liberally to recombine concepts (meeting the generality constraint) depends on working memory representations in prefrontal cortex (Halford, Wilson, and Phillips 2010; Krawczyk 2012; Frankland and Greene 2020). These representations are connected to representations of other kinds, for example in anterior temporal cortex

[3] Bodies of stored information can also exist without the capacity for controlled semantic cognition. So in non-human animals, who have less capacity for flexible cognition or free recombination in working memory, 'concept' is commonly used in the second sense—a stored body of information that is used in categorisation.

(Frankland and Greene 2020). The working memory representations enter into general-purpose combinatorial structures. When running a simulation, they serve to activate, sustain, and manipulate representations in special-purpose systems: structural representations, for instance, or points in a high dimensional state space. Special-purpose representations are operated on in content-specific ways. At the other 'end', a working memory representation interacts and combines with other working memory representations in ways that break free of the state space to which each is connected. Section 5.6 below discusses some illustrative models of how prefrontal cortical systems facilitate the capacity for conceptual recombination.

The interaction exemplifies both role-filler binding and role-filler independence (Penn et al. 2008). Temporary representations in working memory are 'role' representations, undergoing operations (combination, broadly-logical reasoning) that are independent of the particular contents to which they provide access. The role representations are then 'bound' to contents in special-purpose systems ('fillers'). Inferences take place there, when constructing a scenario and running a simulation, that depend on having bound the working memory role representations to these fillers.

Further support comes from neuropsychological evidence. Patients with deficits in semantic cognition have selective difficulty in understanding the meaning of words, and of pictures, objects, sounds, faces, and events in both verbal and non-verbal tasks. The deficits are selective in that other capacities are preserved: visual processing, phonology, decision-making, and so on. Within semantic cognition a distinction is standardly made between semantic dementia, on the one hand, and semantic aphasia or failure of semantic control, on the other (Corbett et al. 2009; Lambon Ralph et al. 2017). Patients with semantic dementia lack underlying knowledge about most categories. For example, they can't understand the word 'cup' or, shown a picture of a cup, tell you what it's called, what to do with it, or very much else about it. (This is different from anomia, a deficit specifically in naming items, with conceptual knowledge preserved.) Semantic dementia is particularly associated with bilateral damage to the anterior temporal lobe. By contrast, patients with a deficit in semantic control exhibit superficially similar difficulties with word and picture tests, but on closer examination they show some preserved conceptual knowledge but a specific impairment with linking and processing conceptual knowledge. How well they perform a task depends strongly on how complex the task is in terms of sorting and making use of conceptual knowledge—in simple tasks they demonstrate that they still have considerable knowledge about the category. Their performance is aided by cues but easily interfered with by distracters, and they are prone to associative errors (between 'squirrel' and 'nuts', say). This deficit in semantic control is associated with damage to the left prefrontal cortex and/or left temporoparietal cortex.

Semantic dementia is a failure in the storage of information, semantic aphasia is an impairment in its use, part of a wider executive deficit. This broadly supports our distinction between the way information about a category is stored and the way it is used online in episodes of thinking. Damage to the areas of prefrontal cortex that instantiate temporary working memory labels, or to their connectivity to other areas of the cortex, produces deficits in semantic control. Damage to areas where information about a category is stored produces semantic dementia. Lambon Ralph et al. (2017) advocate a model in which all the stored information about a category (sensory, motoric, functional, valanced, etc.) is connected to a single hub in the anterior temporal lobe. On this view, all the information about a category is connected to a single index. On other views, information about a category is stored in a more distributed manner, with no single index. My framework is compatible with either view. Online, thinking depends on working memory labels and executive functions. If Lambon Ralph et al. (2017) are right, a working memory label accesses information via a semantic index in anterior temporal lobe. On a distributed view of long-term memory, information about a category is connected together in memory in a more distributed fashion. A working memory label can access this information via many parts of the distributed network. There is functional reality to the fact that information about the category is stored together, but this need not be achieved by there being an index—a single representation (in anterior temporal lobe) by which that information is connected and accessed.

I have said that one ambition for my framework is to show how the different representational structures, computational processes, and informational models that are involved in concept-driven thinking can work together. In this chapter I am arguing that concepts act as an interface: between the general-purpose compositionality and content-general transitions of deliberate reasoning, and the special-purpose representational structures and content-specific computational processes found in many of the other informational models in the mind. But don't concepts then face an interface problem? How do they provide the capacity to interface between all these other things?

I hope it is becoming clear that there is no interface problem, at the level of concepts, if my framework is on the right track. There might be an interface problem if we had to make inferences between very different representations, perhaps from a structural representation providing a cognitive map of space directly to the organized representations of a perceptual quality space. (Even there it is not obvious that there is a deep problem, since connections between representations of different types can be learned from experience, e.g. the perceptual signs to expect when located at a particular spatial location.) But recall the separation between storage and processing. Since general-purpose composition and content-general reasoning take place over working memory labels, rather than over bodies of stored information, there is no problem of putting together concepts whose stored information is different and heterogeneous. The content and significance of the different working memory labels that are put together in forming a thought depends very much on the different bodies of information to which they are connected. But those bodies of information do not need to interface (though they may) in order for the thinker to combine and reason with the concepts.

When conceptual thinking relies on a simulation in a special-purpose informational model to infer something new (e.g. that the sofa will fit in the car), the conclusion becomes expressed as a conceptually-structured thought, for example by categorising the output of an episode of visuomotor imagery. In conceptual thought that conclusion can be considered alongside conclusions based on other special-purpose systems (e.g. evaluating the sofa as aesthetically pleasing). Context strongly influences concept-driven thinking (Spiro et al. 1987; Barsalou 1983, 2016; Medin and Shoben 1988). So the way a suppositional scenario is constructed in a special-purpose system will be affected by the other contents being represented on that occasion. That is the topic of section 5.8 below. Within a special-purpose system, what counts as relevant and similar will also depend on context. For example, two people may seem similar in a context where we are thinking about hierarchy and dominance. The same two individuals may seem quite dissimilar in a context when we are thinking about competence. The ability of special-purpose systems to make non-local relevance-dependent inferences is explored further in the next chapter (§6.3). There are also likely to be effects that stem from the integration of the outputs of different special-purpose systems within the global playground. This is the question of how different aspects of our conscious experience are unified with one another so as to form part of a single coherent whole. That poses an interface problem, not for concepts and offline thinking, but for the way non-concept-involving representations are integrated in experience. This is a phenomenon which it is important to highlight, as I have argued, but it is not something that this book attempts to explain.

Having laid out the basic framework, the next section discusses some further evidence that serves to flesh out the metaphor of a concept as a two-ended plug-and-play device.

5.4 Mediating between Type 1 and Type 2 Processes

The claim that a concept is a two-ended plug-and-play device is supported by research on concept learning. Experiments show that people can learn a new category in one of two different ways (Ashby and Valentin 2017; Gabay, Roark, and Holt 2023). One way is multi-dimensional, implicit, and automatic. The other way is low-dimensional, rule-based, and deliberate.

Deliberate category learning is a matter of inferring a rule for categorising items based on one or at most two features of the stimuli. I look for a rule distinguishing Xs from Ys and infer that Xs have long necks and no spots. Feedback about how to categorise is not essential, and when it is provided, the category label may be given before or even long after the samples to be categorised. Learning is impaired by cognitive load, that is, by having to perform a concurrent task that draws on working memory. As we saw in section 1.2, this is the signature of the type 2 (or 'system 2') style of cognitive processing (Evans and Stanovich 2013).

By contrast, multi-dimensional category learning exhibits the signature of type 1 processing. It occurs automatically, without deliberate reasoning, showing little impairment under cognitive load. The category can be demarcated by a large number of features in a high dimensional state space. Learning works better if the category label comes after the stimulus. Feedback on performance is essential, and the response must be made within a short time after seeing the stimulus. The system learns response-relevant categories, illustrated by the fact that learning is impaired if the location of the response key is switched (Ashby and Valentin 2017). Categorisation is a matter of a content-specific transition from a number of represented features (e.g. perceptual features) to application of the concept. In principle, the category might itself come to be represented in high dimensional state space, in which case concept-application would involve a content-specific transition from a non-conceptually-compositional representation of X to a concept of X.

These two kinds of learning take place at opposite 'ends' of the plug-and-play device that is a concept. Category inclusion can be defined explicitly using conceptual representations of what it takes for a sample to fall under a concept. Deliberation uses reasoning to test hypotheses about what it takes to be an X. By contrast, categories can be carved out within the representational space of a special-purpose system without relying on deliberate reasoning. Learning processes which are independent of cognitive load take place within sensory, affective, motoric, and evaluative systems. Salient distinctions between stimuli can be carved out as regions in the high dimensional state spaces used by special-purpose informational models. The neat thing about concepts is that they are keyed into both systems. So categorisation judgements can rely on either kind of information, depending on the nature of the task.

Conceptual representations in working memory are propositional in the strong sense identified in section 2.4. They can be operated on by content-general computational processes, processes which depend only on the content of broadly-logical concepts like AND, OR, and NOT. Using these concepts in thought correspondingly extends the range of thoughts which can drive suppositional scenarios. A concept's having two ends is, however, a double-edged sword when it comes to solving logic problems. Content effects are always on hand to interfere with doing purely logical reasoning so as to obtain the answer dictated by the norms of deductive logic (De Neys 2012). That may explain why people are generally so bad at logic, especially when problems are presented verbally.

Conceptual representations in working memory are also hierarchical. We can combine the concepts RED, CAR, and TYRE in different ways so as to think about the *tyre on the red car* or about the *red tyre on the car*. That may be important when we deliberately and explicitly use one relation to stand for another in inference (§4.6). Hierarchical combination and logical concepts also come into play in heterogeneous inferences, which involve external representations in different formats (Barwise and Etchemendy 1996; Aguilera 2021) (perhaps also when a map is supplemented to represent negation or disjunction; Camp 2007).

Here is a highly metaphorical picture of this interaction. A large gathering of musicians has formed to make improvised music. There are many different instruments. A central conductor sends out instructions for different groups to produce different sounds: the cello to produce one effect, the flute another; then an oboe, a zither, a harp, and so on. The instrumentalists react to one another to build up a coherent sound. The conductor recombines the elements, subtracts and adds new elements to change the sound. What she does is affected by the unfolding sound picture. What the instrumentalists do is driven by the conductor as well as by one another. A group that has not received the conductor's recent attention fades away and stops playing. The conductor can try recombining elements in any way she likes, but it is only in the unfolding sound picture, where each player is affected by every other, that the elements are integrated together and the overall sound picture is worked out. The conductor is a limited-capacity component driving the process, able to recombine elements freely and to try novel configurations. The product is an interconnected resonating sound picture, integrating different special-purpose components, components that adjust to one another to achieve some coherence, and change over time with their reactions to one another and the changing instructions of the conductor.

I have been fleshing out a picture where the constraints on a simulation come from informational models in special-purpose systems, as well as from conceptual thought. This means that a suppositional scenario can go beyond simply combining information retrieved via a collection of concepts. Constructing the scenario may involve filling in elements, including effects that are configural, depending on overall properties of the constructed scenario (e.g. Gestalt effects). In my chair example, I work out how the chair will fit in the car, but I go on to realise that the car would then be too full to fit the kids in. Spatial, temporal, and perhaps causal constraints within the suppositional scenario can lead to new features being filled in. A special case of this is the process of 'analogical completion', when a thinker is working with an explicit analogy or trying to understand a metaphor, where a mapping between two relational domains suggests new features that can be filled in (Gentner and Jeziorski 1993; Camp 2006, 2019). The interconnectedness of the suppositional scenario has an impact on what is represented in the simulation.

5.5 Shifting Information between Systems

So far I have argued that concepts act as an interface between general-purpose conceptual thought and special-purpose informational models. In this section I will show that they also allow us to transfer information between systems. Recall the different ways that information is encoded and computed with. Many special-purpose systems use organized representations and structural representations. The computations performed over them are content-specific, operating with specific assumptions about their domain of application. Conceptual representations held in working memory use a general-purpose mode of combination. Reasoning processes take us from some conceptual representations to others. Some of these steps are non-content-specific (broadly-logical). Other steps in reasoning are content-specific. They rely on assumptions 'built into' a concept. We saw plausible examples of these direct-CS transitions in section 3.4: from *x is a whale* to *x is a mammal*; from *y is a dog* to *y barks*. Conceptual thought allows us to shift information back and forth between these different forms of information storage.

We have seen that categorisation often depends on a content-specific disposition to move from perceptual representations to a concept. When I see a thing of a certain shape and size, with the texture of feathers and a particular way of moving, I categorise it under BIRD. My BIRD concept is applied to stimuli falling in a region of perceptual feature space. Of the two kinds of categorisation discussed in the last section, this is the automatic mode. (Birdwatchers speak of getting the 'jizz' when automatic categorisation kicks in.) We can sometimes, however, make explicit the information encoded in automatic categorisation dispositions. I can see how I apply a concept, both to actual stimuli and to suppositional scenarios. By looking at how I am disposed to apply BIRD I could conclude that *birds have bills*. That might be something I had not before formulated as a conceptual thought. Having done so, I can store it as a semantic memory. The belief that *birds have bills* is now available to form the basis of reasoning. I can perform broadly-logical inferences on it, for example. Learning that the Gentoo penguin is actually a bird, I recall that *birds have bills* and infer that the Gentoo has a bill of some kind.

I can do the same with information discovered through simulations over suppositional scenarios. To return to my chair example, having learned-through-simulation that the Acme chair will fit in the car, I can store that conclusion as a conceptual belief. Next time I think about the matter, I don't need to run the simulation again. I can simply recall that *the Acme chair will fit in the car* and reason from there.

Philosophical investigation by the method of cases is a way of shifting information between systems. We ask ourselves how we would categorise actual and hypothetical cases. This might be a matter of applying concepts on the basis of perceptual representations, as just discussed. Often, however, what we are interrogating are direct-CS transitions: the information that is implicit in our dispositions to make transitions between conceptual representations in thought (Strevens 2019). This is especially the case when we are asking ourselves what inferences we would draw about actual and hypothetical cases. In practice, this kind of investigation may well be mixed, drawing on simulation as well as direct-CS transitions. But an important aspect is the ability to take information that is stored in direct-CS transitions and make it explicit. It can then be used as a self-standing premise in reasoning, including in broadly-logical reasoning.

Learning by rote moves us in the opposite direction. Learning multiplication tables, the student goes from an explicit representation that 7 × 8 = 56 to an automatic disposition to respond with '56' when queried with '7 × 8?'. Once learnt, we don't answer the question by carrying out repeated addition, nor by recalling as an explicit premise that 7 × 8 = 56. Laborious training has given rise to a direct connection between conceptual representations. Susan Carey's influential account of how children acquire concepts of natural number crucially depends on their acquiring a disposition to make direct-CS transitions between number concepts (Carey 2009). This example also serves to highlight the fact that very many of our direct-CS transitions are acquired socially—in the children's case, it is by learning the counting numbers by rote. We pick up the assumptions and habits of mind of our social group. Our dispositions to make direct-CS inferences with our concepts are shaped accordingly.

We saw in Chapter 2 that replay of sequences in the hippocampal cognitive map can be used to choose actions. More generally, simulations driven by an informational model can be used to train a new model-free action policy (Kurth-Nelson et al. 2023, p. 454). Similarly, an explicit conceptual representation can also be used to train up new connections to special-purpose systems, connections which can then underpin new mediated-CS transitions. The experiment of Barron et al. (2013), involving novel foodstuffs, showed the process at work (§4.3). Initially participants could only imagine a novel foodstuff like tea jelly by activating a sensory representation of tea and a sensory representation of jelly. However, after imagining a compound repeatedly during the experiment they acquired the ability to imagine the compound directly, without going via imagining the components. They acquired a new content-specific disposition for moving back and forth between their TEA JELLY concept and a sensory-affective representation of the compound. They could then perform mediated-CS inferences about tea jelly, for example when asked whether they would prefer tea jelly to an avocado-raspberry smoothie.

Logic has given us a tool that is well suited to explicitly representing the input-output dependencies that content-specific dispositions execute, namely the if-then conditional. Conditionals are a means for shifting information between systems: by being used to train up a new content-specific disposition, as we just saw; or by making information explicit that is implicit in direct- or mediated-CS dispositions. To return to my first example: I notice that the things I am disposed to categorise as birds all have bills. My categorisation disposition, moving from perceptual representations to application a concept, is not a conditional, but it implicitly encodes the information, *if it is a bird, then it has a bill.* In the conceptual system I can make that explicit using the concept IF…THEN. That information is then available to be used as a premise in reasoning. Before being made explicit, although it was informing the way I applied the concept, it was not information I could use in reasoning. The same goes for direct-CS transitions between concepts. Suppose I am disposed to move from *x is a whale* to *x is a mammal.* I can notice the disposition and make it explicit in the representation *if it is a whale then it is a mammal.* This allows me to reason with this information, ask whether it is true, and perhaps reject it. Rejecting problematic direct-CS dispositions is a particular aim of conceptual engineering, when we find that our existing concepts build in sexist, racist, or other objectionable assumptions (Haslanger 2000, pp. 230–1; Machery 2017).

Conditionals can also encode information that we have extracted from a simulation. I work with a suppositional scenario and see that, if I turn the chair on its side, it will fit in the car. That conclusion is something I can represent explicitly in conceptual thought. I have arrived at it by mediated-CS inference, but I don't have to do that again. Next time I can use an explicit premise in my reasoning: *if I rotate the Acme chair, then it will fit in my car.* This also works in the other direction. Having learned that *if p then q*, I can use that to train up various content-specific dispositions. Conclusions that were reached through deliberate type 2 reasoning come to be automatic. The whale → mammal transition is probably like that. Our dispositions to move back and forth between conceptual thoughts and special-purpose representations can also be trained in this way, as we saw in the tea jelly experiment.

There is a huge philosophical literature on the semantics of conditionals. The point I want to make here is compatible with many different views. It just depends on the broad observation that conditionals are a way of encoding suppositional inferences. Thus, conditionals are one of the tools we have for shifting information back and forth between special-purpose informational models and general-purpose conceptual thought.

5.6 Models of Working Memory Labels

I have been talking loosely about a concept being 'connected' to characterizations in special-purpose representational systems. In this section I discuss computational models that offer some options for making this more precise. The models also serve to illustrate how it is that working memory representations facilitate the capacity for recombination which I appealed to in laying out the plug-and-play framework (§5.3).

A very general starting point is that there are ‘files’ in long-term memory which hold a collection of information about a given category or subject matter. This was the way of characterising concepts—in terms of storage—that I was at pains to separate from concepts as representational vehicles tokened in occurrent thinking. As mentioned above (§5.3), a stronger idea is that there is a dedicated mental representation—an index—which serves to connect together the information in the file and provide access to it (Fodor 2008, pp. 94–6). François Recanati has developed a detailed philosophical account of mental files, linked to a psychological account of how they work in cognition (Recanati 2012). In his theory the mental representation is an address for a store of linked information, a store from which representations can be retrieved to use in inference (Recanati 2012, p. 37). Lambon Ralph et al. (2017) argue that amodal semantic hubs in anterior temporal cortex play this role. My framework can embrace indexes, where they exist, but it is also compatible with their absence—with information in long-term memory being stored in a more distributed fashion. What my framework depends on is that some items of information are indeed stored together in long-term memory so that when a thinker accesses some information about a category they are thereby able to access further information about it. That is the functional reality to the claim that items of information are stored together in long-term memory.

In either case what is stored is supposed to be something richer than bare associations. It is information that is about or characterizes a referent. SALT-SAVOURY might be in, but SALT-PEPPER is out. The mental file can comprise information of various different kinds, both encyclopaedic knowledge (semantic memories like *Paris is the capital of France*) and Camp’s broader characterizations (Camp 2015). There are many different proposals for the way information is stored: prototypes, exemplars, mini-theories, etc. This has been a major focus of experimental work on categorisation. For the issue we’re discussing here I can be inclusive about the kinds of information stored and remain neutral between various more specific theories. I am, however, committed to there being a form of deliberate thinking that involves concepts being tokened in working memory; and to a concept, so-activated, providing access to a collection of information stored in memory—perhaps in various different kinds of memory—with different items from the collection being retrieved on different occasions.

How is it that concepts can play this role? Here it is useful to look at specific models of working memory—without committing to their being the final true theory—to illustrate how working memory representations could succeed in playing the role required. I start with the model of declarative working memory in Oberauer (2009). A small number of representations, A, B, and C, are tokened and integrated in a structure, like the structure of a proposition. Each is temporarily bound to an item of information activated from long-term memory: contents like objects, events, and words. The temporary bindings are arbitrary: any of the A, B, and C can be bound to any item of information from long-term memory. The temporary representations are freely recombinable, enabling the thinker to formulate a very large number of different thoughts.

To flesh out this verbal model, we can turn to models of how the architecture could be implemented computationally. These notions of files, indexes, addresses, and binding derive from computer science, so computational models offer useful ways of understanding them. Gallistel and King (2009) emphasise a more specific device from computer science: pointers. A pointer is a representation of an address where some relevant information is stored. A university timetabling program might contain a pointer to the number of students registered for the philosophy of biology MA seminar, call it PG-PHIL-BIO. Operations can be defined and carried out on PG-PHIL-BIO without calling up the numerical value to which it points: divide by 15 and round up to get the number of seminar groups, add to all the other PG-PHIL-... pointers to get the total amount of philosophy postgraduate teaching, etc. When a concrete output is needed, the value pointed to (17, say) is retrieved from memory and entered into the computation.

Eliasmith (2013) has developed the pointer idea in a particular way in his 'semantic pointer architecture'. This is a biologically-inspired model of various aspects of cognition. Processing takes place between interconnected layers of an artificial neural network. Stimuli are initially represented in a high dimensional perceptual state space. These representations are compressed into more abstract representations in subsequent layers of the network. These more abstract representations can be operated on in their own right, for example by using various vector operations to bind them into relational structures. They are pointers because the abstract representations can be 'de-referenced' so as to reconstruct the perceptual representations from which they were compressed. Since compression is involved, the reconstructed representations are only an approximation of the original representations. The pointers are not simply addresses or neutral labels for the more specific perceptual information. They retain some of the structure of the perceptual space.

Kriete et al. (2013) put forward another biologically inspired model of pointing, differing from Eliasmith (2013) in one key respect. Their model has variables in prefrontal cortex that combine in compositional structures (as with our working memory representations). These variables are more like the classical pointers of Gallistel and King (2009) in that the connection to stored information is arbitrary (also as in Oberauer 2009). The pointers do not inherit anything of the nature of the material to which they point, in the way Eliasmith's semantic pointers do.

That is also a feature of the best-developed philosophical application of the pointer idea to concepts (Quilty-Dunn 2021). Quilty-Dunn takes concepts to be unstructured symbols that do not in themselves encode any information, other than pointing to an address where information is stored. Concept combination and logical inference take place over these unstructured 'atoms' independently of the information they point to. For other tasks, information at the pointed-to address is called up and used. Quilty-Dunn has a nice account of how this works in the context of understanding sentences containing polysemous words (see also Liu 2024; cf. Brody and Feiman 2023). In understanding a sentence like 'the school burnt down', the concept SCHOOL retrieves information from the store to which it points. Interaction with the other words in the sentence ensures that it is building-related information that is retrieved (school as building, rather than school as organisation or institution, say).

For my account, I want to embrace one aspect of the pointer idea but resist others. It is crucial that there are representations that we use in reasoning that exhibit role-filler independence with respect to the special-purpose representations to which they are connected. This allows for the free recombinability that is characteristic of conceptual compositionality and facilitates computations that are non-content-specific. But we don't need to assume that all information is so-coded. I want to allow that concepts might encode some information in the form of direct-CS transitions: dispositions to move between conceptual representations within reasoning directly (§3.4). But unlike Eliasmith's theory, these direct-CS transitions need not be an abstract version or compression of the special-purpose information pointed to by a concept. I also want a relation that is reciprocal: from special-purpose representations to a concept (as in categorisation) as well as from a concept to stored information. That is a feature of Eliasmith's model, but not of the classical computer science idea nor of Quilty-Dunn's model. Furthermore, with a pointer in a computer, the information pointed to cannot be used where it is stored. It must be retrieved and brought to a central processor to be computed with. By contrast, representations in an informational model that are activated by a concept in working memory can be processed in the special-purpose systems where they are stored, without being copied elsewhere.

To highlight the contrast, I have adopted the term 'label' instead of 'pointer'. The concept-sized representations that recombine in conceptual thought are neutral with respect to the special-purpose representations to which they are connected. The label on a bottle stands in an arbitrary relation to the substance it labels. But, crucially, labels can also enter into structures that carry information in their own right. Memories stored in the form of explicit conceptual representations—semantic memories, like *Paris is the capital of France*, and concept-based informational models (§4.7)—act differently from other kinds of information connected to a concept (e.g. sensorimotor characterizations). Reasoning takes place over labels-in-compositional-structures. This can be performed without drawing on characterizations in special-purpose systems. The label is tokened when a concept is held in working memory. It serves as a gateway to a collection of stored information. Connected information may be indexed by structures in long-term memory, like semantic hubs in anterior temporal cortex (Lambon Ralph et al. 2017) or indexes in the hippocampus (Teyler and Rudy 2007; Tanaka and McHugh 2018), or it may be stored in a more distributed fashion in terms of patterns of connection between items in long-term memory.

In short, the hypothesis that concepts are plug-and-play devices fits nicely with the functional role of working memory labels: representations in working memory that are temporarily connected or bound to representations retrieved from long-term memory, that can be combined in general-purpose compositional structures, and that can be processed in content-general computations that are independent of the special-purpose representations to which they are connected. This broad framework is supported by a range of neural and psychological evidence and, while there are important differences between them, is consistent with a variety of plausible models of how the computations could be implemented.

5.7 Concept Individuation

The plug-and-play picture invites a clutch of questions about the individuation and metaphysics of concepts. If the concepts used in occurrent thinking are temporary labels in working memory, how are concepts to be type-identified across different episodes of thinking? And across different thinkers? What determines their content? And does the reference of a concept change as the store of information in long-term memory changes? I am not, in the book, attempting to tackle the question of content-determination. Getting a clearer picture of the way conceptual thinking works is hard enough. That picture will be important for an empirically well-founded theory of the metaphysics of conceptual content, but that is not our task here. I do, however, need to say how concepts are individuated, that is, type-identified across occasions. I also want to argue that my picture does not imply that concept reference needs to be holistic or fine-grained.

First off, how am I individuating concepts? What makes different instances of a mental representation count as tokens of the same concept-type? I have been operating with a vehicle-based way of individuating representations. A working memory label is a vehicle type, a temporary representation tokened in working memory. If the same working memory label is reactivated again later during the same episode of thinking, that counts as tokening the same concept again because it is a tokening of the same vehicle type. Here (and only here) the metaphor of a mental word is apt. The vehicle type is picked out in terms of non-semantic properties. What makes some non-semantic properties (like: which letters make up a word) an appropriate way of picking out a vehicle type is partly functional: they are non-semantic properties in virtue of which the vehicle will be processed the same way by the wider system (just as '**dog**' and 'dog' are processed the same way by the language system). (See Shea 2018, pp. 38–40, for detailed discussion.)

Different instances of the same working memory label tokened during an episode of thinking count as instances of the same content-bearing type because they share non-semantic properties in virtue of which computational processing treats them the same. For example, they may be distributed patterns of neural activation that are the same or sufficiently similar that they are processed in the same way.

That does not tell us how to identify concepts across different episodes of thinking. The labels deployed in working memory could be arbitrary as to their subject matter. The same label-vehicle might, on different occasions, be used to label different stores of information in long-term memory. The crucial thing is the way a working memory vehicle is connected to a body of stored information. Tokens deployed on different occasions that are used to label the same body of information thereby count as tokens of the same concept for that thinker.

This delivers a mental type that is a candidate for carrying content: a vehicle type that represents Xs, say. Notice that I am not individuating 'the' concept of Xs. It is an account of that individual thinker's concept of Xs. (More carefully, of one of their concepts of Xs, since they may have redundant concepts: Millikan 2000.) This account depends on the view, which is shared by almost all theories of concepts, that there is a functional reality underpinning the idea that items of information are stored together in memory. It is the persistence over time of a store of information that allows us to type-identify different working memory labels tokened on different occasions as being instantiations, in that individual, of the same concept. We have seen that some theorists make a stronger claim: that information is stored together in long-term memory in virtue of an index—an index being a representation that serves to connect items of information together, and through which they are retrieved. If and to the extent that such indexes exist, same-index will also serve to type-identify labels across episodes. But indexes are not necessary. Storage in long-term memory may be distributed, with no single index. Accessing some parts of the network gives access to other parts. Even so, the functional fact that items of information are stored together is sufficient to allow us to type-identify labels. A given store of information will change as the thinker learns new facts and forgets others. Like Theseus's ship, we can trace a given store of information over time even as it loses and gains individual pieces of information. Labels deployed on different occasions are instances of the same concept-type, in that thinker, just in case they are connected to the same store of information.

The same concept will be used to retrieve different facts and characterizations on different occasions. The plug-and-play framework is designed to accommodate the widespread empirical findings to that effect. Does this mean that it has different contents on different occasions? In one sense, yes: the thinker will be thinking different things about Xs on different occasions. In another sense, no, because the thinker is thinking about Xs each time they deploy tokens of the concept—provided the reference of the content does not shift between occasions. It is tempting to think that it will, if we are in the grip of a picture where the referent is picked out by the way we think about it (e.g. it is the kind that is medium-sized, hairy, and barks). But the idea that characterizations determine reference in this straightforward way is implausible. They are part of what gives the thinker a connection to the referent, so they could underpin an asymmetric dependence relation, for example, or be the causal basis of a representation–world connection figuring in a teleosemantic theory. Most externalist theories of content allow that reference can remain the same as stored information changes. For example, Millikan's theory of concepts holds that the referent of a concept is the substance it is designed to track (Millikan 2000, 2017). This remains the same across radical changes in the conceptions the thinker has about Xs, many of which can be mistaken (and which can be completely different in different thinkers). So the referent of the concept need not change even as the body of stored information changes, and even though the information retrieved on different occasions may be very different.

Nor does the picture imply holism. If the content of a concept consisted of stored information, and the contents of the stored representations were individuated in the same way, then the content of every representation would depend on its connections to every other. However, individuating concepts in terms of vehicle types and reference has no such consequence. A concept is picked out in terms of a vehicle and its referent; items of stored information are representations, individuated in the same way. What the thinker thinks about Xs does indeed depend on what information they store, but the content of their concept is not individuated in terms of that stored information; hence no holism.

What does follow is that the thinker has no guarantee that they have the same beliefs about Xs now as they had when they previously thought about Xs (ditto for other characterizations). The stored information changes over time as the thinker learns, forgets, and changes their mind. Since the information stored together could contain inconsistencies or even contradictions, the thinker can end up thinking contradictory things when thinking about Xs on different occasions. Furthermore, with a large enough change—in the thinker or in the world—there is a possibility that the referent of a concept could shift (on certain theories of content). So there is a potential for equivocation over time. Furthermore, a vehicle-based way of individuating concepts implies that the same thinker may have two concepts of the same referent without realising that they co-refer. That is a merit of the theory, allowing us to explain Frege-type cases (Millikan 2000; Sainsbury and Tye 2007; Fodor 2008; Recanati 2012). But it is a problem for the thinker. These are all real phenomena, so need to be accommodated by a theory of concepts. They are all things the rigorous thinker needs to guard against.

What about different thinkers? My vehicle-based way of individuating concepts does not give us a way of type-identifying concepts between different people. Concepts in different thinkers may share the same referent. That does not make them the same concept in any strong sense. They are no more the same concept as when two concepts within the same thinker share the same referent. Same reference across thinkers does not imply that the concepts will have the same informational significance: that the thinkers will have similar beliefs about, conceptions of, or characterizations of Xs. To the extent that thinkers live in shared environments and interact with Xs in similar ways, we would expect them to share many characterizations; even more so if they live in the same culture, since so much is acquired from people around us. But same concept in this referential sense does not imply shared conceptions.

There may, however, exist commonalities at the neo-Fregean level of sense: collections of interconnected characterizations that are shared by different thinkers. We might have one body of information connected to the public language term 'Superman' and another to 'Clark Kent' (albeit everyone knows that they are co-referential in the fiction). For a more realistic example, Susan Carey argues that, in relation to the mathematical concept of division, a small set of interconnected beliefs come along together at the point when the child first understands that repeated division never reaches zero (Carey 2009). That body of beliefs, shared between children when they reach the stage of properly understanding division, is a good candidate for a sense. My project does not depend on taking a settled view on how concepts should be type-identified at the level of sense, if at all.

Another kind of structure is, however, relevant. The body of stored information connected to a working memory label may contain further structure. For a friend who is also a well-known personality, I have one collection of information about their public profile and another about what they are like in person. I don't need a positive theory of these within-collection structures, but I do need my picture to be compatible with their existence. They constrain how a simulation unfolds when we token a concept and draw on stored information to construct a suppositional scenario. For example, the characterizations we have for concepts of a particular subject matter—people, say—may be arranged around a specific *perspective*, that is a disposition to characterize people in certain ways (Camp 2019). These additional structures are important in determining, amongst all the characterizations connected to a concept, which get taken up and integrated into the suppositional scenario currently being constructed in the global playground.

A final question: how does my proposal differ from Fodor's atomism? I take from Fodor (and others) the insight that concepts display general-purpose compositionality and enter into content-general inferences. That is crucial to the special power of conceptual thought. However, my picture is more driven by neuroscience, a more expansive palette of computational models, and insights about how computations are implemented in the brain (Frankland and Greene 2020, make a nice case for this agenda). Couldn't Fodor just adopt these insights?

Concepts are combinatorial atoms and the rest is a matter of the information stored with the atom. After all, he already has a distinction between the atom that enters into computations and the store of information to which it is connected. However, there are still many differences between these pictures, especially in the broad way I think about characterizations and in my more expansive conception of computation and inference. There is also an important difference in terms of individuation. For Fodor a concept is a mental word that serves to give access to stored information. The same word, *qua* vehicle type, shows up on each occasion. This is to assume, I think, that the representation type which figures in occurrent thinking is also the representation that indexes the store of information. My framework accommodates the fact that these things can be and often are separate. Indeed, my framework is compatible with there being no index—with information storage being distributed. So although the commonalities with Fodor's atoms are playing a crucial role in my picture, the differences are also significant.

To sum up, I am taking a token concept to be a label in working memory, connected to a rich, diverse, and potentially heterogeneous body of information, stored in both general-purpose and special-purpose informational models. Tokening the label provides access to the connected information, a very small subset of which ends up being processed on an occasion, some of it getting tokened in the global playground as part of an interconnected informational model.

5.8 The Simulation Combination Puzzle

Many theorists have emphasised the need for concept-driven thinking to achieve the generalised recombinability and role-filler independence of conceptual reasoning. At the same time, concepts need to be capable of the role-filler binding involved in formulating a thought on a particular occasion, tailored to a specific place, time, and situation. The problem is to achieve both independence and binding (Penn et al. 2008, p. 125; Kriete et al. 2013; Frankland and Greene 2020; Jackson et al. 2021). I follow the general form of the solution proposed by these theorists. Concepts are working memory labels that are connected to rich bodies of stored information, both further conceptually-structured representations (e.g. semantic memories) and representations in special-purpose informational models. Different information is called up on different occasions. Although a large body of information may be potentiated when a concept is activated, only a small subset is used. Each constituent concept contributes a subset of its connected information, a part which coheres to some extent with the information contributed by the other concepts. For example, in thinking about a school fire, the interaction between FIRE and SCHOOL calls up information from SCHOOL

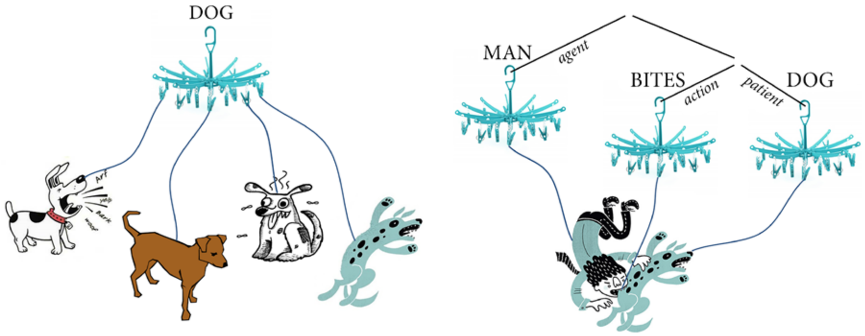


Fig. 5.1 (Left) A label in working memory is connected to many different representations in special-purpose systems. (Right) Working memory labels combine to form a thought. The thought drives a simulation of an integrated scenario that reflects the other labels and their mode of combination. Images from http://clipart-library.com. Graphic copyright the author.

concerning typical school buildings, and information from FIRE concerning large-scale building fires (rather than a homely domestic fire, say).

We can think of each concept as a hook with a collection of odd-shaped objects dangling off it (Fig. 5.1, left panel). When it is hung up together with other hooks, the dangling objects selectively arrange themselves into a coherent picture (Fig. 5.1, right panel). As Frankland and Greene observe, 'A full understanding of a particular combination must flexibly estimate the interactions between the component parts' (2020, p. 281). The representational models built in special-purpose systems will carry much of the load here, selecting components that fit together coherently in the light of the world-knowledge encoded by the model. For example, simulating GREEN LEAF will select for shades of green consistent with natural foliage, simulating GREEN SEA will select for different shades of green.

There is a deeper problem, however. A concept is a plug-and-play device that plugs into structures at both ends. As well as the structure of special-purpose representational models, there is the combinatorial structure by which labels are combined in thought. That has a crucial effect on the simulation. We simulate something quite different for *man bites dog* than for the more mundane *dog bites man* (Fig. 5.1). We don't simply call up the information labelled by DOG, MAN, and BITING and arrange it together into a coherent scenario. The permissible arrangements are constrained by the mode of combination over concepts. What we simulate for DOG depends on its role in the thought. Similarly with language: although hearers can often work out what is meant simply from a list of words, without grammatical cues, in a significant minority of cases grammatical cues are essential to work out what is meant (Mahowald et al. 2022). We understand the nominal compound 'avocado green' quite differently from 'green avocado'. The compositional structure of thought—which may be hierarchical (§5.4)—constrains what is simulated.

As well as seeming obvious from experience, supporting evidence for this comes from experiments on sentence comprehension. There is considerable evidence that understanding a sentence produces automatic sensorimotor activation (Bergen 2012). When a participant is required physically to move their finger away from their body to press a button in response to a stimulus, the response is slower if the just-read sentence implicates movement towards rather than away from the participant (Glenberg and Kaschak 2002). So, 'Liz told you a story' slows down the action compared with, 'You told Liz a story'. The hearer is characterizing themselves somewhat differently in the two simulations. The simulation is involved in comprehension of the sentence (Liu 2024). How then do the details of the simulation come to reflect the structure of the thought? For another example, consider the wide collection of characterizations connected to my YACHT concept. When I think, *I am sailing a yacht*, that calls up a perspective from the helm, producing a very different set of characterizations from those triggered by the thought, *yachts racing across the bay*.

Similarly, in the literature on simulation and learning-by-thinking it is assumed on all sides that the content of a simulated scenario reflects the mode of combination of the concepts which drive it. That is not an objection to those accounts, but it does leave us with a puzzle about how it is that a suppositional scenario comes to reflect the compositional structure of the thought that drives it. The solution is not automatic or straightforward. Just calling up the relevant information is not enough to ensure it will arrange itself in the appropriate way. For example, the text-to-image neural network model Dall-E, powerful as it is, mostly fails to generate images that appropriately reflect the relations in the text prompts, e.g. 'the teacup is under the cylinder' (Conwell and Ullman 2022).

Existing models capture the combinatorial structure over labels but do not illuminate how the material connected to a label is selected in a composition-sensitive way. As Jackson et al. (2021) argue, concept-driven thinking (their 'controlled semantic cognition') has to do two seemingly conflicting things at once. It needs to be able to combine concepts in a way that is independent of the information they carry. As Frankland and Greene say, it needs to have 'combinatorial procedures that are distinct from the contents over which they operate' (2020, p. 295). But it also needs to connect up conceptual thought to special-purpose representations in a way that is suited to a particular context and task. Using neutral labels enables context-invariant composition. But that leaves the other half of the problem unresolved.

For example, Blouw et al. (2016) define a convolution operation which combines semantic pointers (the device we saw in Eliasmith 2013). A vector for *dog* can be convolved with a vector for *agent* to produce a vector for *dog-as-agent*. The operation is reversible, so that the agent can be extracted from any convolved vector by a deconvolution operation. This follows the tensor product architecture suggested by Smolensky (1995), with the difference that, since Blouw et al. use circular convolution (which compresses information), the components are only imperfectly recoverable. Blouw et al.'s model does not give us an account of how the content pointed to (the special-purpose representations pointed to by the DOG pointer) is modulated to take account of the other pointers with which it is combined and the mode of combination. It does not tell us why a simulation of *dog bites man* will select a different subset of the information labelled by DOG than that selected in a simulation of *man bites dog*.

Other models that are built to ensure 'combinatorial procedures are distinct from the contents over which they operate' have the same problem. Halford et al. (2010) use tensor products so that, in a representation of *Sally loves John*, the vectors for Sally, John, and the relation Loves(x,y) are combined in such a way that none is modified by the others. Other models keep special-purpose representations separate. Graves et al. (2016) have a 'controller' that performs operations that are separate from the contents being processed. Knowlton et al. (2012) have a hybrid symbolic-connectionist architecture in which, in the symbolic system, there are dedicated individual units for each relational role and for each object being combined. Kriete et al. (2013) have a pointer architecture in which pointers for agent, verb, and patient are each stored separately (in different 'stripes' in prefrontal cortex). The model can be interrogated flexibly to look up the content pointed to by agent, verb, or patient, but that information is not modified by the mode of combination or by the other contents with which it is combined.

A different solution is to build into the model that the content called up by DOG in the context of *dog-as-agent* differs from that for *dog-as-patient*. Frankland and Greene (2015) found in a human neuroimaging study that there was one region of left mid superior temporal cortex for dog-as-patient and another for dog-as-agent.[4] The region dealing with X-as-agent encodes different information about Xs than the region dealing with X-as-patient. That would explain why different contents are activated in the two cases, but it does not explain how these different contents derive from some invariant content, the shared content encoded by the DOG concept. Frankland and Greene (2015, p. 11737) suggest that this shared content is housed somewhere else in the neocortex, but they do not say how it is modulated into role-specific contents. Across all of these proposals, we are left without a model that conjoins a content-invariant mode of composition with a context- and composition-sensitive selection of information in the simulated scenario.

[4] Schwartenbeck et al. (2023) found a similar result in a quite different paradigm. Participants learnt how to combine geometric visual building blocks into figures using one of two relations (*on top of* and *beside*). They found neural representations of a block-in-relation, for example of Block1 when and only when it was on top of another block.

Here is a tentative solution, suggested as much to highlight the nature of the problem as to resolve it definitively. My solution takes up a suggestion in Frankland and Greene (2020) and extends it to the mode of combination. Return to our green leaf/green sea example. GREEN is a label for a region in a high dimensional perceptual feature space. LEAF is a label for a different region. Simulating GREEN in the context of LEAF activates a part of the GREEN region that coheres with the LEAF region. GREEN SEA would activate a different part of the region labelled by GREEN. Each concept acts as a context for the other concepts, cutting down the regions of state space activated by each label. That gives us a picture of modulation-by-context.

To extend this to the dog-bites-man/man-bites-dog puzzle, first think about what happens during categorisation (rather than simulation). When we see a situation unfolding, there is something about it that makes us think, *the man is biting the dog!* rather than, *the dog is biting the man*. The same is true for language. There is something about the way we perceive the situation that makes us say 'man bites dog' rather than 'dog bites man'.[5] Verbs come with an agent/patient distinction (Pinker 2007, pp. 31–3). Our knowledge of how to use the verb comes with knowledge of what makes one thing its agent and another thing its patient. Similarly, when categorising a situation under the concept BITE, our dispositions for applying the concept must register what is the agent of the action and what is being bitten.

To transfer this to our simple feature space model, consider the relatively large region of perceptual feature space labelled by an individual's concept DOG. Only a subregion is compatible with *dog-as-agent*: regions where the dog is doing something. In the categorisation direction, we are disposed to think *dog bites…*only when the input is in the subregion of the DOG feature space where the dog is doing something. The disposition to move from special-purpose representation to concept-label in an agent role only applies to a subregion of feature space. The subregion lies in a high dimensional feature space. It is not here picked out by some simple perceptual distinction (as it plausibly is for *green leaf* vs. *green sea*). But there is something about ways of representing dogs that makes just some of them candidates for being dogs playing an agent role.

Now recall my earlier argument that, unlike pointers, labels have a bidirectional connection with special-purpose representations. We can, then, rely on the same move as we did in explicating *green leaf* vs. *green sea*. When simulating MAN in the context of MAN BITES…, the label MAN indexes a region of state space which is cut down by its context. That context is BITES and, in particular, MAN being the agent of BITES. So the portion of state space that MAN contributes to the constructed scenario is the subregion in which the person is in an agent role, i.e.

[5] Thanks to Daniel Rothschild for this suggestion.

is doing something. Sun and Manohar (2023) show that this could be achieved in the brain through transient strengthening of synapses. Furthermore, the thing the man is doing has to overlap with the state space of the action BITES; for example, he could be doing it with his teeth. Further interactions with the region connected with DOG results in an integrated scenario where the regions of state space activated by each label are adjusted to fit with one another and to respect their mode of combination. Each concept provides a context for retrieving information from the others. What counts as similar within colour space may be different when the space is projected along the landscape dimension from when it is projected along the botanical dimension (Grand et al. 2022). The concepts themselves may be polyvalent: the information they encode may be suited to characterising predicates of different adicities or singular terms, depending on the combinatorial structure in which they are deployed.

My suggestion is that the compositional structure of a conceptual thought constrains the subregions of high dimensional state space that are activated by each constituent concept. Just as, during categorisation, relational concepts are applied in a way that is sensitive to which object is the agent and which is the patient of the relation; so also, moving in the other direction, when a relational concept drives a simulation, whether a concept is in the syntactic role of agent or patient provides a corresponding constraint on the scenarios that will be simulated. The syntactic structure into which working memory labels are combined furnishes content constraints on the type of scenarios that will be simulated. In Figure 5.1, the fact that DOG is hooked into the compositional structure using an 'agent' hook cuts down the bits of dangling information that are candidates to feed into the simulation. Similarly, in understanding a metaphor, functional structure constrains the interpretation, whereas only open class terms are subject to metaphorical elaboration (Glanzberg 2008).

This hypothesis applies, to conceptual thought, the theory from linguistics that the syntactic structure of a sentence contributes to its meaning. Syntactic structure specifies a semantic content, but only of a very general kind (Harris 2020). Vyvyan Evans calls it a 'parametric content', as opposed to the detailed analogue content of the cognitive model that is eventually constructed (Evans 2015; see also Paivio 1986). In the sentence, 'Those boys are painting my railings', the syntactic structure alone specifies that the meaning has to be a scenario in which, *those somethings are somethinging my somethings*. The syntactic structure constrains the state space of possibilities within which a scenario can be constructed. Paul Pietroski has put forward a detailed model of sentence meaning that has the same effect (Pietroski 2018). In his account word meanings are coarse-grained and concepts are fine-grained (like the particular regions of state space activated in a particular simulation). Word meanings are instructions for how to access a concept. Which concept is fetched on an occasion depends on the syntactic structure and function words in the sentence. Syntactic structure is telling the comprehension system what kinds of information to look up, for each word, in order to understand the sentence (cp. which kinds of information to retrieve, for each label, in order to simulate the situation).

In sum, although it is uncontroversial that a suppositional scenario constructed in a concept-driven simulation reflects the compositional structure of conceptual thought, there is a puzzle about how that works in cognitive or computational terms. Accounts that separate combinatorial procedures from stored contents are generally silent about the way that, when simulating aRb, the information labelled by each component (a, b, R) has an effect on which information labelled by the other components enters into the simulated scenario. Frankland and Greene (2020) is an important exception. Their empirically-based model does accommodate these contextual effects. It does not, however, explain how the mode of combination affects what is simulated (the dog-bites-man/man-bites-dog problem). I have argued that their account can be extended in a natural way to provide a plausible answer. My suggestion is that compositional structure constrains which subset of the information labelled by each concept is a candidate to feed into the simulation, constraints that mirror the way that people are disposed to deploy conceptually structured thoughts in categorisation.

5.9 Bringing It All Together

Concept-driven thinking has given humans an especially powerful way of inferring novel conclusions about the world. How so? The generality with which concepts can be combined, irrespective of their subject matter, must be part of the answer. So too is the ability to carry out reasoning over syntactically structured conceptual thoughts. But we do much more with concepts than reasoning. Fodor (2000) famously argued that theorem-proving-type computations in the language of thought cannot account for abductive inference and relevance-based search (declaring this to be the great mystery of cognitive science). Part of the answer is that we can carry out inferences that depend on informational models in special-purpose representational systems.

What makes concepts crucial in this is process is that they act as an interface. They serve to connect together, combine and rearrange information from special-purpose informational models into a coherent suppositional scenario—an interconnected representation of a situation. A concept is a working memory label that can be combined and manipulated independently of the body of information to which it is connected. But a concept also acts as an interface to that information. In a process akin to role-filler binding, a collection of labels held together in working memory activates a subset of the special-purpose information to which each is connected. It is this that allows us to build a rich interconnected representation of a situation.

Much of the information that is activated may initially be incompatible and inconsistent. But when it is held together in working memory, in the service of working out what to do or what is the case (rather than free imagination), inconsistencies can be resolved into a more coherent picture.[6] That is to say, concept-driven thinking brings together, in the cognitive playground, information from different special-purpose systems, forming a rich informational model—a suppositional scenario in the cognitive playground. Running simulations with a suppositional scenario is a form of inference, drawing conclusions that can then be expressed in conceptual thought. Inference within this kind of coherent representational model works quite differently from the theorem-proving that is characteristic of reasoning. It relies on the informational models and content-specific computational dispositions encoded in special-purpose systems.

The ability to perform these kind of inferences with integrated suppositional scenarios depends on a general-purpose capacity for holding items in working memory, composing them, and computing with them. Concepts are the crucial interface between these general-purpose and special-purpose capacities. Concept-driven thinking is a combined operation that achieves more than the sum of the parts. Therein lies the special power of human conceptual thought.

Chapter Summary

5.1 Reaching Conclusions via Simulations

This chapter is about how conceptual thought arrives at concept-involving conclusions via informational models of other kinds. Simulation, prospection, and imagination are all ways of drawing on special-purpose systems in the service of forming new beliefs (examples). (p. 118)[7] A given intention could be formed as a result of simulation or through reasoning (example).

Simulation can be relied on to reach an accurate conclusion where the relevant information cannot simply be retrieved from long-term memory or inferred by reasoning alone. (p. 119) Evidence of this process is found in experiments on mental rotation, and in results showing effects of sensorimotor processing in conceptual tasks. Simulation can allow the thinker to discover something new, without getting new evidence from the world, by bringing into thought information that is only encoded implicitly in the operation of special-purpose systems, or is encapsulated.

[6] Cp. Carruthers (2011a, p. 439), who has a process of evaluation subsequent to an initial process of generation.

[7] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

This chapter is about the role of concepts when running simulations in special-purpose systems and constructing suppositional scenarios in the cognitive playground. (p. 120) Sections.

5.2 Simulations Use Special-Purpose Informational Models

In focusing on either the domain-general aspects (Fodor) or domain-specific aspects (Barsalou) of conceptual thinking, theorists have underplayed the way concepts provide an interface between the two. (p. 121) A concept offers access to a rich body of information, both explicit conceptual representations (semantic memories) and wider *characterizations*. Much information is encoded in special-purpose systems—which may or may not be domain-specific, and may employ modality-specific, supra-modal, or amodal representations. (p. 122) Simulation itself need not be driven by concept-involving deliberation—we are interested in the role of concepts in the process when it is.

5.3 Concepts as Plug-and-Play Devices

This section argues that the role of concepts in simulation-based inference is as 'plug-and-play' devices. The archetype, which is not conceptual, is the way place cells are severed from their input-output connections and played with offline in trying out different possible routes. This is a widely-applicable trick. (p. 123) Reaching a conclusion via simulation involves playing with representations in special-purpose systems. A concept is a plug-and-play device with plugs at both ends: at one end, the informational models and content-specific computations of special-purpose systems; at the other end, the general-purpose compositionality and content-general reasoning of conscious deliberation.

A concept, it seems, gives us the capacity to use, offline, any of the representations to which it is connected. Recombining concepts drives the construction of novel suppositional scenarios, putting together special-purpose representations in new configurations. (p. 124) For example, thinking about what to prepare for dinner involves visuo-motor, olfactory, and gustatory simulation, affective-evaluative responses, a spatial map of locations in the environment, motoric knowledge of effort cost, and semantic knowledge of monetary costs.

There is much evidence in psychology and cognitive neuroscience for temporary task-dependent or working memory representations (often in prefrontal cortex). To token a concept is to token one of these working memory representations. (p. 125) We need to differentiate the representation that is combined and reasoned with in online processing from the body of information, stored in long-term memory, to which it gives access. There is evidence that the capacity liberally to recombine concepts does indeed depend on working memory representations in prefrontal cortex, separate from the representations that are activated in simulations and language comprehension. (p. 126) Temporary representations in working memory are bound to 'filler' representations in special-purpose systems; operations of combination and broadly-logical reasoning exhibit role-filler independence. Further support derives from neuropsychology and the standard distinction between semantic dementia, which is an impairment in the storage of information about a category, and semantic aphasia, which is an impairment in the capacity to work with and integrate information about a category. (p. 127) This broadly supports our distinction between the way information about a category is stored and the way it is used in online thinking; storage may go via a single index or may be distributed.

Do concepts face an interface problem? Concepts interface with special-purpose systems, but because the representations that are combined and processed in reasoning are neutral labels, there is no problem with putting together concepts that concern different subject matter—no interface problem. (p. 128) A simulation in a special-purpose information model may be influenced by several of the concepts currently being entertained in working memory (§5.8), each operating in relevance-dependent ways as explored in section 6.3; and there are likely effects of coherence in the global playground, which is relied on but not explained here. The next section fleshes out the plug-and-play metaphor.

5.4 Mediating between Type 1 and Type 2 Processes

The metaphor of a concept as a two-ended plug-and-play device is supported by research on concept learning. Deliberate category learning involves inferring a rule for categorising items which is based on one or two distinguishing features; it is impaired by cognitive load. (p. 129) By contrast, multi-dimensional category learning has the signature of type 1 processing. These take place at opposite 'ends' of a concept: by reasoning amongst conceptual representations, or by carving out distinctions amongst representations in special-purpose systems. Conceptual combination can use logical concepts. (p. 130) Also hierarchical structure; both extend the range of thoughts that can drive suppositional scenarios.

A good metaphor here is a musical conductor: a limited-capacity component driving an integrated, coherent ensemble effect. Constructing the scenario involves filling in elements based on information and constraints found in special-purpose systems, and on the coherence of the overall scenario.

5.5 Shifting Information between Systems

(p. 131) Conceptual thinking allows us to shift information back and forth between special- and general-purpose systems, and between representing implicitly in content-specific dispositions and representing explicitly. Information that is merely implicit in a disposition to apply a concept to a certain region of perceptual feature space can be made explicit, reasoned with, and stored as a semantic memory. The same applies to conclusions reached through simulations over suppositional scenarios. Philosophical investigation by the method of cases is a way of making explicit information that is implicit in dispositions to categorise or to make direct-CS transitions between concepts. (p. 132) Learning by rote moves in the opposite direction, giving rise to new dispositions to make direct-CS transitions. A conceptual representation can also serve to train up new connections to special-purpose systems, for example, the ability of participants in Barron et al.'s (2013) study to imagine the novel foodstuff *tea jelly*.

Conditionals offer us a way to make explicit, reason with, and endorse or reject information that is implicit in content-specific transitions. (p. 133) Conditionals can also be used to train up new automatic content-specific dispositions (e.g. with the whale → mammal transition, the explicit conditional probably comes first). This point about conditionals does not depend on a particular view about the semantics of conditionals, just on the more general claim that they serve to encode suppositional inferences.

5.6 Models of Working Memory Labels

In what way is a concept 'connected' to stored information? (p. 134) A general starting point is that a concept is connected to a file of information stored in long-term memory (the existence of an index for the information is optional). The stored information includes both semantic memories and wider characterizations. I will look at some illustrative models to show how concepts can play this role; Oberauer (2009) has a small number of active representations integrated in a structure, each temporarily 'bound' to items of information activated from long-term memory.

(p. 135) A more specific device from computer science is a pointer. Eliasmith's 'semantic pointer architecture' has developed a particular version of pointers: representations on which operations can be performed in their own right, but which retain some of the structure of the representations to which they point. Kriete et al. (2013) have pointers—working memory representations that combine into compositional structures—whose connection to stored information is arbitrary. Quilty-Dunn (2021) applies the pointer idea to concepts, in particular to explain comprehension of polysemous words. (p. 136) I want to embrace part of the pointer idea, namely free-recombinability and operations that are independent of the special-purpose information pointed to, but to allow: direct-CS transitions as well recall from the pointed-to information store; reciprocal connections; and operations on special-purpose representations in the systems where they are stored, not requiring retrieval into central processing. I adopt the term 'label' for the representations that combine in working memory: a label can label something further but also encode information in its own right. (p. 137) In short, the hypothesis that concepts are plug-and-play devices fits well with the functional role of working memory labels.

In short, the hypothesis that concepts are plug-and-play devices fits nicely with the functional role of working memory labels, of which there are several plausible models.

5.7 Concept Individuation

This section says how different token representations should be type-identified as instances of the same concept; content-determination, however, is a question for another day. Different instances of a given working memory label tokened during an episode of thinking are tokens of the same concept; they are treated the same way in processing. (p. 138) Labels are type-identified across different episodes of thinking in terms of the store of information to which they are connected. This tells us how to pick out same-concept tokens, not across individuals, but in a given individual; and an individual may have more than one concept of a given referent X.

Although the thinker deploys different characterizations on different occasions, and the store of information changes, the referent will usually remain the same (as many theories of content imply). (p. 139) Nor does the picture imply holism, since stored information is individuated in terms of referential content. None of this rules out failings—problems that occur in practice: thinking contradictory things about Xs on different occasions, storing contradictory information, or unwittingly having two separate concepts of X.

This does not give us a way of type-individuating concepts between different thinkers, except at the level of reference, which does not imply shared conceptions. (p. 140) There may, however, be cases where concepts come along with stable packages of shared information. Whether or not that is true, the way an individual deploys stored information is often structured: it is organized around a perspective, namely a set of dispositions to characterize a particular subject matter in a particular way (like various perspectives we can take on people) (Camp 2019). My picture differs from Fodor's atomism in that working memory labels are not names, i.e. a vehicle type which, across different episodes of thinking, stably provides access to the same body of information; my picture is compatible with there being no index for stored information, and if there is an index, it is not (or need not be) the index that figures in composition and reasoning in occurrent thought.

(p. 141) In short, a concept is a working memory label.

5.8 The Simulation Combination Puzzle

Formulating a thought on a particular occasion involves using a small subset of the information connected to each constituent concept, selected so as to cohere together. (p. 142) Informational models in special-purpose systems contribute world-knowledge to how the information is integrated. Figure 5.1: a label in working memory is connected to many different representations in long-term memory, a small subset of which are integrated into a simulation that reflects the other labels and their mode of combination. A deeper problem is to explain how the mode of combination over concepts constrains what is simulated. (p. 143) Experimental evidence shows that simulations reflect the mode of combination. This is also assumed in the literature on learning-by-simulation; but an account of this is not automatic or straightforward.

At the same time as working with concept labels in a way that is independent of their contents, concept-driven thinking has to connect up conceptual thought with special-purpose representations in a way that is suited to a particular context and task. Blouw et al. (2016) have a nice model, but their method for combining semantic pointers does not modulate information to take account of the other pointers or the mode of combination. (p. 144) Models that separate combination and reasoning procedures from stored information generally face the same problem (examples). Frankland and Greene (2020) do find that there are different ways of representing dog-as-agent and dog-as-patient, but do not explain how these are created out of some shared DOG content.

(p. 145) My tentative solution is based on Frankland and Greene's idea that each label is connected to a state space, and that the simulation has to fall within a subregion of the state space connected to both labels. To extend this to the mode of combination, first notice that, in categorising a situation, features of what is perceived make us think *man bites dog* rather than *dog bites man*. That is, only some regions of the perceptual feature space labelled by DOG are compatible with dog-as-agent. Since labels are bi-directional, there are corresponding constraints in the other direction: the way MAN is simulated is constrained to lie within the regions compatible with being an agent and, within these, with being an agent of *biting*. (p. 146) My suggestion, then, is that the compositional structure of a thought constrains the subregions of high dimensional state space that are activated by each constituent concept. This hypothesis applies, to conceptual thought, the theory from linguistics that the syntactic structure of a sentence contributes to its meaning. (p. 147) In sum, I have suggested that compositional structure constrains which subset of the information labelled by each concept is a candidate to feed into the simulation, constraints that mirror the way the same conceptually-structured thoughts are applied in categorisation.

5.9 Bringing It All Together

Concept-driven thinking is powerful, not just because of the generality with which concepts can be combined and processed, but also because they can be used to carry out inferences in special-purpose representational systems. What makes concepts crucial is that they act as an interface: a collection of labels held in working memory activates a subset of the special-purpose information to which each is connected, thereby building up a rich interconnected representation of a situation. (p. 148) New conclusions can then be reached, both by reasoning, and via inferences that take place either within special-purpose systems, or over suppositional scenarios in the cognitive playground. Therein lies the special power of human conceptual thought.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0005

6
The Frame Problem and the If-Then Problem

6.1 The Frame Problem

The last chapter painted a picture of the way concepts act as an interface between special-purpose informational models and conceptual reasoning. This chapter argues that the picture presented there shows how it is that human cognition manages to solve the notorious frame problem (to the extent that it does). Part of the solution is to avoid the problem, as I will explain, but that throws up the lesser-known 'if-then' problem. Human cognition has a way to navigate its way, well enough, between these two problems, by relying on the plug-and-play character of concepts.

The frame problem we are concerned with here is the problem of relevance-based search (Fodor 1987, 2000, 2008; Samuels 2010; Xu and Wang 2012; Antony and Rey 2016; Shanahan 2016). A person or computer carries out inferences in order to work out what is the case or what to do. How does the system select, in a way that is computationally tractable, which stored representations to perform inferences on? How do we take decisions on the basis of what is relevant without having to consider and reject all that is not relevant?

The problem arises because there is no simple rule to decide which information is relevant to a given question or task. Relevance is 'isotropic'—relevant considerations can come from any direction (Fodor 1985; Chow 2013). I'm thinking about what to have for breakfast. It turns out that deforestation in Borneo is a relevant consideration. (Does the margarine contain unsustainable palm oil?) Unless potential relevance is constrained to a specifiable and tractable subset of everything I know and believe, it seems that, in order to assess which stored representations are relevant, the system will have to check through all stored representations and assess each for relevance. But the task of checking every piece of stored information for potential relevance is computationally intractable.

The frame problem arose long ago in artificial intelligence research.[1] Cognitive and computational scientists building computer systems to perform complex tasks—tasks that display aspects of what in the human case we might call intelligence—found that the selection of relevant information from a large store of memories presented real practical problems. On the other hand, this seems to be something that humans do with some facility, perhaps giving an indication of a human cognitive competence that was not well modelled by classical computational systems.

A closely-related problem is the question of how to model abduction or inference to the best explanation. Here again the relevance of information is isotropic. Considerations that are potentially relevant to the goodness of an explanation can come from anywhere. In addition, inferring the best explanation seems to require an overall evaluation of a wide range of factors. It calls for a global assessment of a collection of beliefs. For example, the conclusion that the post-industrial increase in the earth's mean temperature is largely caused by human activity is well supported. It is the best explanation of a wide range of data and phenomena. However, reaching the conclusion that this is the best explanation is extremely complex, requiring a very wide range of information to be weighed and assessed (data, models, scientific arguments). One reaches a conclusion on this question by taking a global assessment of many different considerations, some central, other peripheral, not all pointing in the same direction. Even if we do not actually perform a genuinely global assessment of the import of everything we believe in order to answer this kind of question, the phenomenon suggests that some kind of non-local computational process may be involved (§1.3).

The frame problem is, in the first instance, a problem for those designing computational systems. It presented itself as a major obstacle to feasible artificial intelligence when classical computational systems were the central tool of AI research. (With the rise of deep neural networks (DNNs), the frame problem faded into the background, as we shall see.) It is also a problem for those seeking to understand the mind computationally. This is why it is so significant philosophically. Our most successful account of intelligent thought and action is the representational theory of mind (RTM). Representations are manipulated physically in ways that are faithful to their semantic content. What Fodor calls 'central cognition' appears to be able to retrieve information according to relevance, and to perform the non-local computations required for abduction. How is that achieved by the physical manipulation of representations, in a way that is

[1] The name comes from an earlier (related) computational problem about updating a scene without having to specify a large number of 'frame axioms' about facts that will not change as the result of a given event (Sprevak 2005) (a problem that has largely been solved: Shanahan 2016).

computationally feasible? Fodor declared that the failure to answer that question makes the workings of central cognition the great mystery of cognitive science (Fodor 2000, pp. 23, 99; Xu and Wang 2012).

In section 6.2, I show how DNNs succeed in avoiding the frame problem, to a large extent. However, in doing so, they end up facing a problem of their own, which I label the 'if-then' problem. The if-then problem and the frame problem are in some ways complementary, but when a task is not susceptible to the if-then solution, and calls for broadly-logical reasoning from explicit representations, the frame problem still arises (§6.3): which representations should reasoning be performed on? In section 6.4, I argue that the account of concept-driven thinking advanced in Chapter 5 offers a partial solution: special-purpose informational models can be used as a way of generating relevant information. With concepts acting as an interface between special-purpose informational models and general-purpose reasoning, cognition can partly avoid, and partially, imperfectly, solve the frame problem (§6.5).

6.2 Avoiding the Frame Problem Leads to the If-Then Problem

Computational modelling offers insights about how human cognition might solve the frame problem. In recent years concern with the frame problem has subsided in computer science. In AI research, classical computation has been eclipsed in most areas by DNNs. The interest of DNNs is not that they are realistic psychological models—they clearly differ in profound ways from human cognitive competences—but because they show how certain problems can be solved in principle, and potentially offer partial models of particular aspects of human cognition. DNNs do not seem to face the frame problem—at least, relevance-based search does not arise as a concrete issue that modellers are forced to grapple with. As it has receded as a practical concern, theoretical work on the frame problem has also subsided. Nevertheless, the problem has not gone away. As we will see in this section, DNNs do not so much solve the frame problem as avoid it. In the following section (§6.3) we will see where the bump in the carpet has popped up now.

DNNs in effect build in assumptions of relevance. By having a huge number of free parameters (e.g. 1.7 trillion in the GPT-4 large language model), and by being given enormous amounts of data, DNNs can be trained to produce appropriate outputs in response to a wide range of different inputs. They store what they have learnt, not in the form of discrete memories of the data they were trained on, but in the entire pattern of weights distributed across their interconnected layers. When the system encounters a new input, it has no need to retrieve information stored during training on which to perform inference. It just proceeds to produce the output that has been trained into it by experience. Each relevant past experience has left its trace on the system's processing dispositions through gradual adjustments made, across many successive cycles of training, to the whole pattern of weights.

It was long thought that this approach was inadequate to deal with real-world tasks. Fodor attributed the inability of computational models to perform these tasks to their failure to solve the frame problem, declaring:

> the failure of artificial intelligence to produce successful simulations of routine commonsense cognitive competences is notorious, not to say scandalous. We still don't have the fabled machine that can…translate everyday English into everyday Italian; or the one that can summarize texts; or even the one that can learn anything much except statistical generalizations. (Fodor 2000, p. 37)

But now, of course, DNNs are actually doing quite well at performing these tasks. The breakthrough came in 2012 when a convolutional DNN broke all records for categorising pictures from the ImageNet data set (Krizhevsky et al. 2012). (I vividly remember this result for the way it reinvigorated my undergraduate lectures on connectionism that autumn.) Since then, DNNs have demonstrated considerable facility at 'summarizing texts' (Yousefi-Azar and Hamey 2017; Bubeck et al. 2023) and 'translating everyday English into everyday Italian' (Bahdanau, Cho, and Bengio 2014; Stahlberg 2020; Bubeck et al. 2023) (Google translate has transformed off-the-beaten-track travel); also many other tasks.

What these results show is that statistical generalisation is much more powerful than previously thought. Yes, all kinds of background information is potentially relevant to translating a sentence. But it turns out that strikingly good results can be achieved by encoding input-output dispositions that implicitly encode particular assumptions of relevance. The pattern of weights reflects the way incoming information was relevant to producing the correct output for samples it was trained on. The perhaps surprising finding is that these assumptions allow the system to generalise effectively—to produce appropriate outputs in response to inputs that it has not previously encountered.

DNNs share with previous modular approaches to the frame problem an underlying limitation. The assumptions that are implicit in their operation only work well within a specific domain. Convolutional neural networks build in and learn assumptions about which features are important for categorising images. These assumptions are implicit in the content-specific dispositions they acquire as a means to solving the input-output problem on which they were trained (§3.2). Those dispositions are not suited to performing tasks in other domains.

This limitation may be circumvented by deploying multiple modules, each designed to deal with a different specific domain (Shanahan and Baars 2005). Image processing can be done by a trained ConvNet, language processing by a Transformer. Human cognition does involve multiple special-purpose informational models, as we have seen (Chapter 4). Many of these depend on implementing a suite of content-specific dispositions. Extensive experience in a domain, through the course of evolution and individual learning, endows the system with a set of if-then dispositions appropriate to its domain. There remains the problem of how to decide which inputs go to which systems. Perhaps a competitive process can help here, especially for inputs that are sufficiently distinct that they make no sense when presented to the 'wrong' module (Shanahan and Baars 2005). A module trained to process language will not settle on a specific output when presented with data from a visual image. A visual processing module presented with the same data will settle on the categorisation 'elephant' but would make no sense of linguistic data. On some views, the winning outputs are integrated together in a common working memory system (Shanahan and Baars 2005). On other views, all that's needed is a collection of competing special-purpose modules—so called 'massive' modularity (Carruthers 2003). We will return in the next section to the question of integration across different special-purpose systems.

A second criticism of the modular approach is that, even when we confine our attention to one special-purpose module, its behaviour will be insufficiently flexible to produce appropriate outputs. Often what is appropriate depends heavily on the context. Hearing 'fire!' shouted by the house manager in a theatre prompts quite different behaviour than when it is declaimed by an actor in the play; so too on a cold camping trip; or at a military training ground.

DNNs have shown that this problem is often surmountable. Context is something that the system can register as another aspect of the input. The appropriate output is not just a reaction to the currently-presented word or stimulus, but to a short history of information that the network has been fed. In Transformer-based large language models, the output is just the predicted next word, but the input is a long chain of text. What the system outputs next after the last word depends heavily on what came before. For example, given a joke, the PaLM model can output a string of words that explains the joke.[2] But the input here is not just the joke, but a whole mini-essay that also gives the system two prior examples of jokes with explanations. With all that text as context, the assumptions of statistical relevance that have been trained into PaLM's 540 billion parameters mean that the output that follows this long input is a text that amounts to the explanation of a joke. If DNNs are trained to deal with inputs that consist of such long streams of data, they can be highly sensitive to context—just by being sensitive to different features of the input.

If context counts as just another input, then the system has to go into a different state for each context it might encounter. Then, when the final element of the

[2] 'The joke is that the speaker's mother is trying to get them to go to their step dad's poetry reading, but the speaker doesn't want to go, so they are changing their flight to the day after the poetry reading' (Chowdhery et al. 2022).

input comes along (‘fire!’), it can output the response that is appropriate in that context. The system deals with the past by changing the state of the processor. Botvinick et al. (2019) give the example of a DNN that was trained on an array of different reinforcement learning problems. Different problems consist of different stimuli with different stimulus-response-outcome probabilities, but all the problems share the same structure. The network does ‘meta-learning’, acquiring a set of weights that allows it to learn about a particular problem—a particular set of stimulus-response-outcome probabilities—on the fly, in its network dynamics. The context is a string of past inputs, and the system deals with the past by changing the state of the processor.

This raises a problem that C. R. Gallistel has long pressed as an objection to artificial neural networks (Gallistel 2008; Gallistel and King 2009). Suppose that, on the way home one evening, an agent observes that a certain tree has come into fruit. When, the next morning, they decide whether to turn left or right on leaving their shelter, the observation the night before can act as part of the input on which their behaviour is conditioned. The observed state of the tree is an input, I_1. Seeing the fork in the path the next morning is another input, I_n. All the observations they make in between count as further inputs, I_2 to I_{n-1}. Their output O_L, turning left, is a response to the (complex) input I_1, I_2,…, I_n. Had they instead observed that the tree was not in fruit the night before, that is a different input, I_1'. The agent will behave adaptively if they are disposed to make a different output, O_R (i.e. turn right), when presented with the (complex) input I_1', I_2,…, I_n. To condition its behaviour appropriately on the distant past, the system would have to have appropriate input-output dispositions with respect to extremely long chains of input. The DNN solution is to deal with the past by changing the state of the processor. That means it has to allocate dedicated processing resources to each potential input it might encounter. It was a significant discovery that a neural network could be trained to respond appropriately to so many different inputs, in a way that generalises accurately to other inputs of the same type. But this depends on the network model having an enormous number of free parameters. In the largest Transformer-based language models, the input—the prompt—can now be very long indeed. However, increasing the length of the prompt has a dramatic effect on how much computing power it takes to train the system and how many parameters are needed. This is a symptom of the weakness Gallistel pointed to. A system that has to devote dedicated processing resources to each chain of input it might encounter will eventually run up against the ‘infinitude of the possible’ (Gallistel and King 2009, pp. xi, xvi, 51, 136–48). It is a practical impossibility to encode a separate processing disposition for every potential input string to arbitrary depth into the past.

Botvinick et al. (2019) raise the same problem as a practical obstacle for DNN modellers. The meta-learning solution they demonstrated for reinforcement learning will only extend so far. Even very large language models like GPT-4, which can take as their context an input thousands of words long, are poor at performing tasks that call for a longer term memory within the context of the task, for example to write a novel with a coherent overall narrative (Bubeck et al. 2023). Botvinick et al. suggest a solution to the if-then problem, namely to store explicit memories of circumstances encountered and outcomes received. This echoes Gallistel and King's argument that a practical computational system for solving real-world tasks must store and process explicit memories.

Taking stock, DNNs have shown that the if-then solution to the problem of relevance and context-sensitivity is much more effective than was ever imagined when modular architectures were originally touted as a solution to the frame problem. But the if-then way of taking account of context gives out eventually, as it encounters the 'if-then problem': the 'infinitude of the possible' and the need to devote dedicated processing resources to each long chain of input it might encounter. That obstacle can be overcome by remembering the past explicitly—that is, not by changing the system's input-output dispositions (e.g. weight matrix), but by storing explicit memories of circumstances it encounters.

6.3 A Compound Architecture Still Faces the Frame Problem

We have canvassed two different ways of dealing flexibly with variable context: learned if-then dispositions and inference from explicit memories. These approaches have complementary costs and benefits (Botvinick et al. 2019; Shea 2023b). The if-then solution is learning-heavy and computation-light. It calls for a large amount of experience to acquire a range of useful input-output dispositions; but then it can produce an output rapidly in response to the current input. A system that stores explicit memories can potentially learn what to do much faster, even after a single exposure, but calculating how to respond to the current input is typically more computationally demanding—it may involve a tree-search through a combinatorially large space of chains of possible states and outputs.

We have seen that human cognition deploys systems that work in each of these ways. Many special-purpose informational models rely on content-specific processing dispositions. On the other hand, reasoning over conceptual representations can deal flexibly with stored explicit memories. Penn et al. (2008) argues that this kind of compound system is a good way to model human cognition. Researchers working on artificial intelligence also construct compound systems. Botvinick et al. (2019) discuss a compound model that uses a DNN to learn the problem space, and combines that with a gradually expanding memory record of every situation it has encountered (world state, action, reward) (Graves et al. 2016). When encountering a new situation, the system works out what to do by comparing the new situation to the most similar situation stored in its episodic memory, picking the action that proved most rewarding in that situation in the past. This means that the system can do one-shot learning, repeating what worked on a single occasion in the past without needing to have each experience painstakingly re-presented multiple times. Having once discovered what to do in response to a new situation, retrieval from episodic memory will tell it what to do when it encounters that situation, or one sufficiently similar, again.

Human cognition can perform content-general inferences on explicitly represented information retrieved from memory—both semantic memories and suitably conceptualised episodic memories can enter into broadly-logical reasoning. But memories can also be subject to content-specific inferences, for example in special-purpose informational models. The same is true in AI architectures. Many teams are experimenting with using explicit memories to transcend the limitations of a purely if-then solution, sometimes by processing those memories simply as further inputs to a trained DNN. Ryoo et al.'s (2022) model stores as explicit memory a summary of its whole history of inputs. The memory is read, written, and processed using a Transformer-based language model at each step. Park et al. (2023) simulate a group of agents living in a simple artificial world. The agents act and interact by receiving text as input and producing text as output. Each also has a memory of its own individual characteristics, circumstances, and preferences. Adding these memories as part of the input to the Transformer model means that each 'agent' produces outputs that reflect that individual's character and situation.

When explicit memories act as inputs to a trained DNN, those inputs are just acting as further contextual cues to a trained if-then disposition. The system is still working within the range of behaviours it has been trained, end-to-end, to perform. It doesn't take us beyond the if-then way of avoiding the frame problem. However, the capacity for content-general reasoning does offer the chance to go further. Given an explicit representation of a situation completely outside the range of situations on which its if-then dispositions were trained, a system endowed with the capacity for content-general reasoning can still do something sensible. It can perform broadly-logical reasoning to combine the things it knows and reach new conclusions. This gives the system something worthwhile to do with memories that fall outside the range of things it has been trained to have specific dispositions to respond to.

The same is true for generating novel thoughts. General-purpose compositionality means that concepts can be combined in new ways to formulate completely novel thoughts, representing situations that fall far outside the system's experience. Einstein could formulate the idea of running at the speed of light to pursue a light beam (Einstein 1970, p. 53).[3] The capacity for broadly-logical reasoning means the thinker can perform inferences on novel thoughts, even when their

[3] Interestingly, Einstein describes this in terms of the 'free choice of such concepts', not obstructed by being 'immediately connected with the empirical material' (Einstein 1970, p. 49).

trained if-then dispositions are of no use. Einstein inferred that he should expect to see the beam of light as an electromagnetic field at rest and spatially oscillating. The capacity for content-general inference allows human thinkers to deal intelligently with novelty: novel explicit memories or novel combinatorially-generated thoughts.

To take stock, the frame problem is avoided by systems that are trained to have sufficiently rich input-output dispositions, but that eventually runs into the if-then problem. A solution to that is to store explicit memories and, in the human case at least, to compute with them in content-general ways. This is where the bump in the carpet re-emerges. Although the two solutions are in some ways complementary, a compound approach that relies partly on stored explicit memories will then face the problem of selecting which memories to compute with. It will still need to overcome the frame problem.

We see this in the AI systems that deploy a compound architecture that includes an explicit memory store. The system in Graves et al. (2016) stores explicit memories, as we just saw. It can work out what to do in the current situation by repeating the action that led to reward when the same or a similar situation was encountered in the past. But to retrieve memories of the same or similar situations, it has to perform an operation that takes all stored memories as input. Pritzel et al. (2017) build a reinforcement learning system that writes all experiences to memory. Although they have a more efficient mechanism for calculating which past experiences are most relevant to the current context, doing that depends on first using (actually approximating) a time-consuming 'k nearest neighbours' search across its whole memory store (Yang et al. 2020, p. 129276). Park et al. (2023) store, for each virtual agent, a comprehensive record of every event experienced by that agent. A small subset of these memories are retrieved to act as part of the input at a given time-step. Memories are selected based on their recency, importance, and relevance to the current situation. Calculating relevance, however, involves calculating the similarity between the current situation and every event stored in memory. In short, these compound models face the problem of relevance-based search and deal with it using operations that take account of the entire store of memories.

What these computational models suggest is that, although there are advantages to a compound architecture that overcomes the if-then problem by storing explicit memories, the frame problem—the problem of how tractably to search the store of memories for relevance—then re-emerges.

6.4 A (Partial) Solution

I am going to suggest that the solution deployed in human cognition is not just a compound, but a hybrid—a hybrid, in that it can take advantage of the if-then approach as a way of searching memory. The last chapter sketched the picture, with concepts mediating between general-purpose reasoning and special-purpose informational models. This is imperfect: it is not a complete solution to the frame problem, but a way of approximating a solution. I will argue that it offers a good picture of how human cognition manages partly to avoid the frame problem and, when it does arise, to deploy a partial solution.

A suggestion in the literature on how human cognition deals with the frame problem is to use content-addressable memory. The idea is that only a small subset of memories are retrieved as being potentially relevant in the current context. For example, the system could store information about a given individual or category in a mental file (Chow 2013). Relevant information can be retrieved from memory by searching all the information in the mental file. Which files should be accessed? Carruthers suggests that, in the context of considering a linguistic statement, one should perform a content-based search of all the concepts expressed by the statement (Carruthers 2003). That would certainly generate some relevant information, but for an even moderately complex real-world problem, it would still involve searching through an enormous number of representations to check each for relevance. In considering, 'should I have cereal, toast, or fruit for breakfast?', does my decision-making system really need to check for relevance everything I know or believe about breakfast cereals, toast, and fruit? Each of these concepts content-addresses a huge amount of information (not to mention my concept of *myself*). And even that wide-ranging approach would miss many relevant considerations, unless it were to expand outward and access information addressed by concepts used within the files (e.g. MARGARINE in the TOAST file). So while content-addressable memory is surely part of the solution, we still need an account of how it can be implemented in a way that is not, on the one hand, too myopic to be useful or, on the other hand, too demanding to be feasible.

My suggestion is that this is achieved in our case, at least in part, by leveraging the assumptions of relevance found in special-purpose informational models, especially in their content-specific processing dispositions. Here I will consider two kinds: the direct-CS transitions that take place between conceptual representations (§3.4); and non-local transitions that take place through a form of parallel constraint satisfaction in a representational state space (§4.4). Most of our focus will be on the latter, but I start briefly with the former.

Recall from section 3.4 that there is evidence that content-specific transitions take place between conceptual representations directly. Just tokening an occurrent belief may dispose the thinker to token a consequent thought. For example, someone who thinks *Moby is a whale* may thereby be disposed to think *Moby is a mammal* (similarly if MOBY is replaced with any other singular concept). The inferential disposition is 'built into' the concepts and does not require a general premise (i.e. *all whales are mammals*). This is quite unlike looking up all the information in a mental file, since only a small number of transitions are potentiated. These inferential dispositions in effect build in assumptions of relevance, not for a specific domain (as in the visual system), but for a specific concept. This is not a solution to the problem of searching a list of memories—there is no search and selection process operating—but these kinds of transitions are probably part of the way that cognition introduces relevant information into the stream of thought.

The second solution is to rely on similarity or proximity in a representational state space (Churchland 1998; Shea 2007; Kriegeskorte and Kievit 2013). As we saw in section 4.4, contents in special-purpose informational models are sometimes represented in a state space, on the basis of which people make judgements of similarity (Charest et al. 2014), or judgements about other relations (Nelli et al. 2023). This can be deployed as a way of looking up relevant information: of retrieving memories that are similar to the current context.

These similarity spaces are found, not only within domain-specific systems like visual face processing, but also more widely. Huth et al. (2016) recorded fMRI data while participants listened to hours of radio stories. They modelled the meaning of words in the audio stream using word embeddings (where the vector for a word characterises which other words it tends to co-occur with) and used a regression model to predict voxel-by-voxel brain activity from the word vectors. They found that, using their regression model, they could predict activity in many cortical areas based on which word was being presented in the auditory stream. The weights in the regression model revealed activity organised along a number of semantic dimensions, for example a dimension with perceptual and physical categories at one end and human-related categories (social, emotional) at the other. The axes of variation in the neural signal separate words into categories like: tactile ('fingers'), visual ('yellow'), numeric ('four'), locational ('stadium'), abstract ('natural'), temporal ('minute'), professional ('meetings'), violent ('lethal'), communal ('schools'), mental ('asleep'), emotional ('despised'), and social ('child'). Activity in diverse neural areas reflects variation along these semantic dimensions, particularly in superior temporal cortex (long associated with semantic processing), parietal cortex, and prefrontal cortex.

In short, there is now considerable neural as well as behavioural evidence for the kinds of representational spaces postulated by Churchland (1998, 2012). Both seeing images and understanding sentences generates representations that are organised into similarity spaces. These spaces need not be domain-specific. They encompass the kinds of abstract semantic dimensions found by Huth et al. (2016). What is interesting for our purposes is that making transitions within a semantic space offers a computationally tractable way to perform relevance-based search.

For example, when I am considering how to behave in relation to one person X, I can move to representations of similar individuals, Y and Z, and recall how I acted in relation to them in the same situation. Moving to nearby portions of semantic space is a way of prompting relevant information. But this is not like looking up and searching through all the information in a mental file, checking each piece of information for relevance. The shape of the semantic space effectively builds in certain assumptions of relevance. To this extent, it is a way of reusing the if-then way of avoiding the frame problem as a (partial) solution to the frame problem.

Furthermore, semantic spaces offer a ready way to deal with context-sensitivity. Representations organised in a semantic space are related along several different semantic dimensions at once. For example, face stimuli are automatically organised along dimensions of trustworthiness and dominance (Oosterhof and Todorov 2008). Relevance can be assessed along just one of these dimensions, or any combination of them. For example, in a dynamic state space in prefrontal cortex that registers both colour and direction of motion, activity can be projected along the dimension—colour or motion—that is relevant to the current task (Mante et al. 2013). The dynamics of a network can be changed by 'clamping' one or more dimensions and considering only relations in the remaining subspace. In the last chapter (§5.8) we saw how the content of one concept can act as a contextual cue which constrains the processing taking place in a special-purpose system activated by another concept with which it is combined. That offers a model of how different dimensions of similarity (trustworthiness, dominance, etc.) are selected in different situations.

Grand et al. (2022) compared the way human participants and a DNN arrange objects along different dimensions (Fig. 6.1). For example, *tiger* and *dolphin* are judged as similar in respect of *size* but very different in respect of *dangerousness*. This is predicted by activation patterns in the trained DNN. Representations are close together in state space when projected along the size dimension but far apart along the dangerousness dimension. Applying this insight to a space representing people, when I represent an individual X in that space it should be straightforward to retrieve individuals who are similar with respect to dominance. A contextual cue can thus act as a 'clamp' so that retrieval in a semantic space takes place along a contextually relevant dimension. The same space can be sampled along more than one dimension. This means that, when retrieving memories to use in inference on a given occasion, a single informational model can be sampled for relevance in more than one way.

Doesn't this just push back the problem? Retrieval can rely on assumptions of relevance implicit in semantic state spaces, but how does the system learn the state spaces over which this occurs? The answer is that these spaces are learnt, laboriously, from experience, as we have seen. We have good empirical evidence that this is the case and plausible computational models of how it occurs. The frame problem is not the problem of how semantic spaces or categorical dispositions could be acquired in the first place (important as that question is). State spaces may also figure in an account of how some concepts are acquired, namely

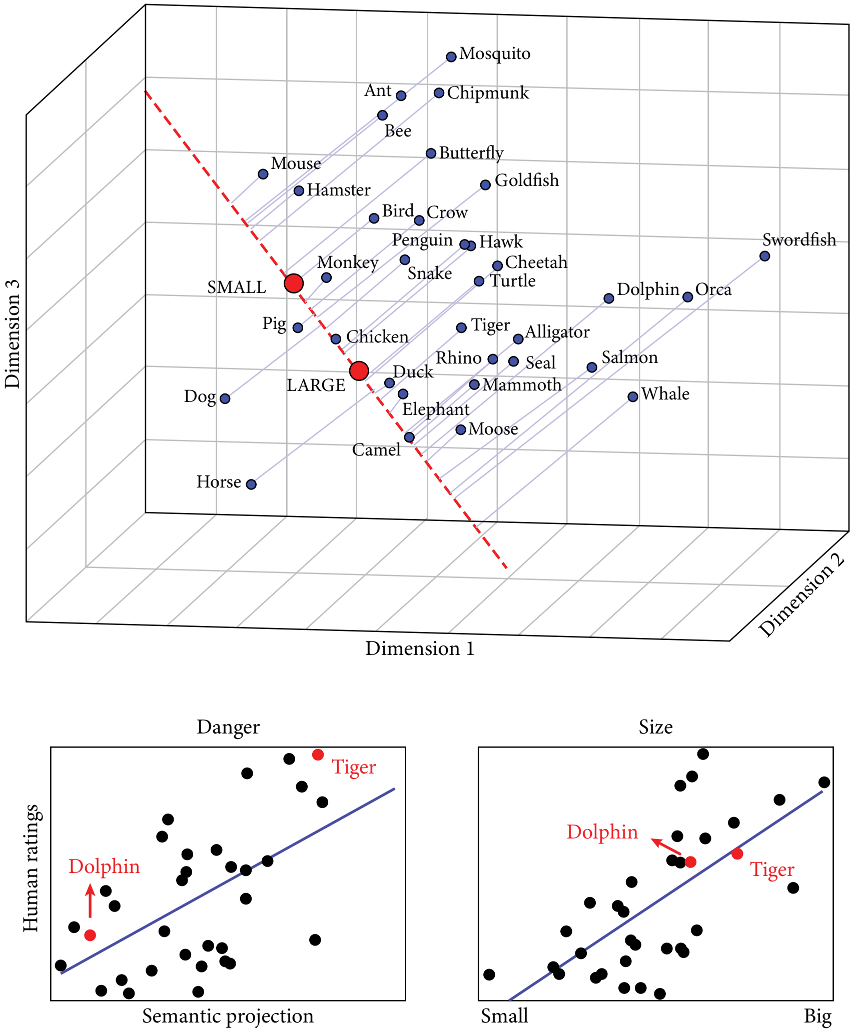


Fig. 6.1 (Top) Items arranged in a high-dimensional semantic space (illustrated here in three dimensions) project onto a semantically-significant dimension of variation (small to large). (Bottom) The same items can be projected onto different semantically-significant dimensions of the underlying space. Dolphin and Tiger are close along the *size* dimension but distant along the *dangerousness* dimension. From Grand et al. (2022). See the open access online edition of the book for the full colour figure.

through alignment between partially incomplete state spaces (Aho, Roads, and Love 2023; see also Søgaard 2023).

Relevance-based search was challenging for RTM because it seems that the search for relevant information is non-local—it somehow takes into account a whole collection of information. As we saw in section 1.3, we do in fact have computational models, consistent with RTM, where transitions effectively take account of a collection of information in parallel. Such computations are not mysterious if we don't limit ourselves to step-by-step classical computations (as considered by Fodor in declaring the mystery). When proximity in the state space of a trained neural network is used to retrieve relevant information, that is a non-local computation (of the kind highlighted in §1.3).

Assessing similarity is non-local in the sense that it involves weighing many different characteristics at once and calculating their resultant. How is that computationally tractable? Within a semantic space it works because the geometry of the space reflects all these different features at once. The geometry of the space is trained into the network by experience. Once trained, closeness in similarity space reflects an overall assessment that integrates lots of features at once. Many different samples have been encountered and had an effect on the local gradients at each point in the space, each experience having more impact in some areas than in others. The moves then made in the trained similarity space are computationally undemanding, but reflect that wealth of experience. This is captured in the model by a step that takes account of a whole matrix of values at once. It occurs in real neural systems by a process that takes place across a whole array of neurons in parallel.

Here is an analogy. Consider a comet moving through the solar system. When it is at a certain point, we might ask how the comet calculates where to go next. Its next step will depend on its interaction with a huge array of objects, some close by, others distant. Large numbers of nearby asteroids will each have an impact. A really close asteroid could have a big effect. Much further away, the sun will have a large effect; also to some extent each of the planets. To parody the frame problem, it looks like, in order to work out where to go next, the comet has to calculate the effect of each of these other celestial bodies on its future trajectory. How does it make so many calculations in real time? Why isn't it paralysed in one spot, working out where to go next?

The answer, of course, is that the comet does not need to interact separately and serially with all the other bodies. They all have an effect on the local gravitational field and the comet reacts to that. The local gravitational field is the resultant of the integration in parallel of a huge number of different forces. By reacting to the resultant force the comet's behaviour reflects the parallel effects of a huge array of interactions all at once. Moves in semantic state space are like that in that the relevant representations are accessed by making moves in a space that reflects parallel constraint satisfaction across a whole collection of information.

Another example is a computation that proceeds by exploring a whole state space in parallel. We see examples in some computational models of route calculation in the hippocampus (mentioned briefly in §1.3). The calculation is based on a process that takes place in parallel across the whole array of place cells. This effectively sweeps through many different routes that trace back from a given goal to the location of the agent (Samsonovich and Ascoli 2005; Khajeh-Alijani et al. 2015). The relevant computational property is unlikely to be simply a matter of activation, but instead a dynamical property like the phase offset between the activation at different locations during synchronous activity, activity like the sharp wave ripples observed electrophysiologically in the hippocampus. Recent work suggests that sharp wave ripples may be the basis for episodic memory recall in the hippocampus (Norman et al. 2019), in which case this would be clear-cut example of a global computation that performs relevance-based search and retrieval.

6.5 How Cognition Partly Avoids and Partially Solves the Frame Problem

I have sketched a way that a hybrid computational system can address the frame problem, partly by avoiding it with dedicated if-then computations, and partly by approximating a solution in areas where the limits of the if-then solution are reached. This picture offers us an account of how human cognition solves the frame problem, to the extent that it does. Content-general reasoning with conceptual representations allows us to consider novel scenarios and sensibly process representations that transcend the experience on which our special-purpose informational models have been trained. However, since concepts act as an interface to special-purpose informational models, those systems can be re-purposed offline, in simulation mode, to generate relevant considerations on which to perform inferences in thought. Doing this across multiple different built-in assumptions of relevance can approximate an isotropic search for relevance.

How does my suggestion differ from other proposed solutions to the frame problem? A first observation is that DNN models have shown that the if-then way of avoiding the frame problem is much more powerful than previously thought (for example when Fodor was writing about the frame problem in the 1980s and 1990s). It turns out that you can get a long way with systems that build in implicit assumptions of relevance—provided the systems have enough exposure to experience in their evolutionary history, and especially in their learning history, to be able to realize a suite of complex input-output dispositions.

A second element of my proposal is the feature of conceptual thought that formed the centrepiece of the previous chapter. We reach conclusions in conceptual thought not just by reasoning from explicit memories, but also by running simulations in special-purpose systems. This is a way that conceptual thought can take advantage of the domain-specific assumptions of relevance that are built into special-purpose informational models. Concepts can act as mediators between a range of different special-purpose systems. Suppose I'm thinking about my extended family sitting around the living room on a social occasion. A great aunt arrives. I can simulate what could happen next by relying on the implicit assumptions of my system of naïve physics; my system for tracking moving agents; and my system for tracking social hierarchy. I can see that the aunt is most likely to move towards the gap on the sofa, but I predict that this move will be disastrous because she will then be rude to the relative that she would sit next to. Trained if-then modules each have their own assumptions of relevance, but conceptual thought can in effect rely on lots of different assumptions of relevance, of diverse kinds, mediating between them to generate potentially relevant considerations and evaluate them.

The combinatorial power of conceptual thought is important here. Its relevance to the frame problem only becomes clear when we focus on the way concepts allow us to rely on the assumptions of relevance contained in special-purpose informational models. Concept compositionality is then seen as a way of relying on and juxtaposing different kinds of assumptions of relevance from different domains. That is quite different from looking up all the information connected to a concept (all the entries in a mental file) and assessing each for relevance. Each special-purpose system just throws up the one or two considerations it takes to be most relevant. (It is obvious, as soon as I simulate the social hierarchy, that the great aunt's sitting *there* would be disastrous.)

This way of retrieving relevant information on which to perform deliberate inference circumvents the need to do what the classic formulation of the frame problem asks us to do, namely to search through a large list of memories and select those that are potentially relevant on which to perform inference. Running a simulation in a special-purpose system need not involve searching a list of memories. The system has a disposition to token various representations in various circumstances (both online, in response to input, and offline, in simulations). Those representations need not be stored explicitly anywhere. We have effectively re-cast the problem. The benefit of the capacity for reasoning with explicit representations is the ability to deal with the past by reasoning with explicit memories, rather than having to treat the past as a further contextual cue, part of one long chain of inputs. One way to do that is store an explicit representation of each past situation (as in the Pritzel et al. (2017) model, say). Doing it that way throws up the problem of searching the list for relevance in a tractable way. But the benefits that accrue from reasoning with explicit representations of the past don't require the memories to be stored that way. Information can be stored in the form of trained dispositions to token an explicit representation given certain inputs. Models of episodic memory based on pattern completion work like that (Teyler and Rudy 2007). I have been arguing that deliberate thinking can rely on memories generated by special-purpose systems in that way. This does not displace the question of how it is that *relevant* memories are generated. But it does show that search through a list of explicit memories is not the only option.

We saw that representational state spaces provide a ready assumption of relevance in their similarity structure. Projecting along different dimensions (e.g. size and dangerousness) allows the same representational space to be sampled for relevance in a number of different ways. My suggestion is that, when we are engaging in deliberate conscious thinking to work out what is the case or what to do, we can effectively retrieve a range of different relevant considerations by running simulations in different special-purpose informational models and, within a model, by sampling for similarity along a number of different dimensions. Taken individually, none is a comprehensive search of all relevant information; taken together, they can go some way to approximating an isotropic search, one in which relevant information can come from many different directions.

A further refinement is that deliberate thinking has access to representations about how to think, including tips for searching for relevance. Often we learn these socially. One set of strategies involves randomising in some way, to put oneself in a new context: move to a new physical location, look up a random word in a dictionary, think of answers beginning with 'T', ask 'who, what, where, when, why, how?', etc. The new context provides a new way of probing special-purpose informational models for relevant information. Other socially-acquired strategies are more specific. For example, if you're planning a mountaineering trip to a remote location, don't forget to think about what type of cooking fuel you'll be able to get. In between randomising and very specific pieces of relevance-searching advice there is a whole suite of tools for recall: tools that we learn socially. That we create and share these tools in itself suggests that the problem of relevance-based recall is a real practical problem faced by human cognition.[4]

Taken together, these tools and techniques give deliberate thought ways of finding information that is potentially relevant in diverse and heterogeneous respects. Once we generate a limited set of relevant information, conceptual thought adds the capacity to reason step-by-step with this information. That is how I get from contemplating breakfast to thoughts about the rainforest. Thoughts of foods and flavours generate some options. But I can locate those concepts in a semantic space that has quite abstract dimensions. I may, for instance, organise consumer goods by their environmental impact. (Not very accurately, to be sure, but perhaps with some crude evaluative feel.) That throws up a dimension of contrast between margarine on toast and sliced apple, say, and a dimension of relevance that brings to mind the palm oil plantations of Borneo. Some recent AI models have this hybrid character. Although large language

[4] Although outside the scope of the book, it is worth noting that social processes are also important for generating knowledge in their own right. For example, scientific discovery is a deeply collective process, based on the culture, norms, and institutions of science. That is another way of achieving relevance search. Even if no individual solves the frame problem, if they sample in different ways and transmit information culturally, the social process may approximate a collective isotropic search for relevance.

models may not use broadly-logical reasoning in their internal processes (Traylor, Feiman, and Pavlick 2021), they can approximate or display the capacity for broadly-logical reasoning in their outputs, especially when appropriately prompted (Bubeck et al. 2023). This means the same underlying LLM can be alternately prompted in a hybrid way, first in a way that encourages it to rely on its learned content biases (i.e. assumptions of relevance), and then re-prompted to encourage it to perform logical inference on these representations (Creswell, Shanahan, and Higgins 2023). Moving back and forth between these two kinds of prompting improves the system's performance. This exemplifies the kind of divide-and-conquer strategy I have been advocating.

When we are constructing a suppositional scenario in the cognitive playground, that in itself may act as a prompt for relevance. As I build up a picture of my ideal breakfast-in-bed, I fill in bits that are obviously missing, like a teaspoon to go with the teacup, and also configural properties of the scenario, like the fact that I have imagined too many different items so that they won't fit together on the tray. Representations filled in as a result of constructing the suppositional scenario can act as further contextual cues for retrieving relevant information from memory.

These ingredients do not amount to an exhaustive way of searching for relevance. Relevant information can still be overlooked. Reasoning is a powerful domain-general way of reaching new conclusions, but it can lead the agent in quite the wrong direction if relevant and important information is not fed into the decision-making process. So the approach I have sketched is an imperfect solution. However, human decision-making is imperfect. We are sometimes myopic and overlook considerations whose relevance would be obvious, if only we had considered it. We are famously biased in the factors we take into consideration; and the information that comes to mind can be powerfully primed by context (Tversky and Marsh 2000; Azzopardi 2021). (That is the downside of relying on built-in dispositions about how context implies relevance.) We can be very effective in situations we have encountered many times before, but if we want to have a good chance of recalling information relevant to a novel situation, we often have to rely on explicit strategies and mnemonics.

Most of these elements have been discussed before in relation to the frame problem, in some guise. The role of concepts as mediators has not been emphasised in previous approaches. Carruthers suggests a similar role for sub-vocalised language (Carruthers 2003). But he thinks of language as a way of accessing a collection of content-addressed beliefs, not as a way of driving simulations in special-purpose systems. Nor does the complementarity between the frame problem and the lesser-known 'if-then' problem feature much in the previous literature. It is in the context of the recently-discovered power of DNN-based if-then approaches, and the fact that they nevertheless still face limits that call for the storage of explicit memories, that the trade-off between these two different styles of computation becomes clear (Botvinick et al. 2019; Shea 2023b). I also make explicit the way that built-in assumptions of relevance in special-purpose systems can be relied on in memory retrieval, and add a model of how they can be polled for relevance in more than one way.

In short, the model of concept-driven thinking developed over the preceding chapters offers a way of avoiding and solving the frame problem. It is a realistic computational proposal for how representational processing can be configured to sail a middle course between the if-then problem and the frame problem, taking advantage of the complementary costs and benefits of each. Most importantly for our purposes, it is an empirically plausible hypothesis as to how human cognition manages to avoid and solve the frame problem, to the extent that we do.

Chapter Summary

6.1 The Frame Problem

This chapter is about how human cognition manages to solve the frame problem and the lesser-known 'if-then' problem. The frame problem is the problem of how cognition manages to select relevant information on which to perform inferences. Relevant considerations can come from anywhere (isotropy), but checking every piece of stored information for relevance is computationally intractable. (p. 156)[5] This was a practical problem for good old-fashioned artificial intelligence, but it seems that humans solve it with some ease. Closely related is the problem of abduction, which furthermore seems to involve the non-local weighing of a range of different considerations at the same time. The frame problem is also a problem for theorists trying to understand the mind, Fodor's great mystery of central cognition. (p. 157) Sections.

6.2 Avoiding the Frame Problem Leads to the If-Then Problem

With the rise of DNNs, the frame problem has receded as a practical issue for AI researchers, but DNNs do not so much solve the frame problem as avoid it. DNNs do not store explicit memories, but effectively build in assumptions of relevance in their learned weights. (p. 158) It was long thought that this approach was inadequate to deal with real-world problems, a failing that Fodor attributed to the failure to solve the frame problem. But now they can. Surprisingly, given enough training, DNNs can learn processing dispositions that build in appropriate assumptions of relevance, and which generalise effectively.

[5] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

DNNs share a limitation with earlier modular approaches: their built-in assumptions only work within a specific domain. That can be circumvented to some extent by having a range of different domain-specific modules, but this calls for some way of deploying them selectively and integrating their outputs. (p. 159) And even a special-purpose system needs to be sufficiently flexible to produce different outputs in different contexts. DNNs treat context as just another aspect of the (very long) input. That is to deal with the past as a contextual input that changes the state of the processor. (p. 160) This raises a problem pressed by Gallistel: the system has to allocate dedicated processing resources to each potential input it might encounter. Botvinick et al. (2019) suggest a solution, which is to store explicit memories of the circumstances encountered and the outcomes received.

(p. 161) In short, although DNNs have shown that if-then dispositions are much more effective in avoiding the frame problem than previously thought, this solution gives out eventually; a suggested solution to this 'if-then' problem is to store explicit memories.

6.3 A Compound Architecture Still Faces the Frame Problem

Learned if-then dispositions and inference from explicit memories have complementary costs and benefits. Human cognition deploys both approaches; some AI models do the same. (p. 162) Stored memories can enter into content-general and content-specific inferences. If retrieved memories just act as further inputs to trained if-then dispositions, then the system has to have been trained on and dedicate resources to responding to each such input; by contrast, a capacity for content-general reasoning can be applied to a representation of a situation wholly outside the system's training experience. Content-general reasoning can also be applied to novel thoughts; general-purpose compositionality can generate such thoughts.

(p. 163) A compound architecture, while helpfully taking advantage of the complementary profiles of the two approaches, still faces the frame problem—the problem of selecting which memories to compute with. Compound AI models do face this problem (examples), and deal with it by performing operations on the entire store of memories. The frame problem has re-emerged, and exhaustive search, while feasible in the models, does not amount to a solution.

6.4 A (Partial) Solution

I argue here that the solution deployed in human cognition is a hybrid, with plug-and-play concepts taking advantage both of if-then dispositions and general-purpose reasoning, and reusing the if-then approach as a way of retrieving relevant memories.

(p. 164) Content-addressable memory may be part of the solution but, without further constraints, looks to be either myopic or too wide-ranging. I suggest retrieval can rely on content-specific dispositions. Direct-CS transitions effectively assume that certain contents are relevant, which they introduce into thought directly. (p. 165) A second way of introducing relevant information is to retrieve representations that are nearby in a representational state space. These similarity spaces are found widely, with representations organized along a number of semantically-relevant dimensions. Making transitions within a semantic state space offers a computationally tractable way to perform relevance-based search. The state space effectively builds in certain assumptions of relevance.

(p. 166) Semantic spaces offer a ready way to achieve context sensitivity. Relevance can be assessed along any one of several different dimensions, for example: colour or motion of a stimulus, dominance or trustworthiness of individual people. A contextual cue can act as a 'clamp' so that retrieval takes place along a relevant dimension, as with the contextually-relative judgements of similarity in the experiment by Grand et al. (2022). (p. 167) Acquisition of these state spaces is a different problem—learning laboriously from experience—of which we have plausible accounts. Figure 6.1: representations in a high-dimensional state space are arrayed differently when projected onto different semantically-significant dimensions of the underlying space (e.g. size vs. dangerousness).

When proximity in the state space of a trained neural network is used to retrieve relevant information, that is a non-local computation, of a kind that would appear mysterious if we were limited to step-by-step classical computations. (p. 168) Similarity in state space is the resultant of taking account, in parallel, of a large number of parameters at once. An analogy is the way a comet moves in the solar system. At any point it moves based on the resultant of the forces generated in parallel by a very large number of other celestial bodies at once. Another way non-local inferences could occur is illustrated by a computational model of route calculation in the hippocampus that involves a process that propagates in parallel across the whole array of place cells at once.

6.5 How Cognition Partly Avoids and Partially Solves the Frame Problem

(p. 169) I have suggested that human concept-driven thinking relies on special-purpose informational models to generate relevant considerations, using multiple contextual cues to retrieve information according to multiple built-in assumptions of relevance in order to approximate an isotropic search.

This differs from previous theories, first, in placing greater reliance on built-in assumptions of relevance, motivated by new DNN models demonstrating that the if-then way of avoiding the frame problem is more powerful than previously thought. Second, the account of concepts in Chapter 5 shows how conceptual thought can rely on different assumptions of relevance in diverse if-then systems, and integrate their results. (p. 170) The combinatorial power of concepts is important here, each connecting into different assumptions of relevance in special-purpose informational models, in a way not previously emphasised. The picture of retrieval shows that, when memories are not stored as a list of explicit representations, there is no need to search a list to retrieve relevant information (retrieval works by pattern completion or some other dispositional process). (p. 171) Semantic state spaces show how it is possible to perform such retrieval in more than one way, by sampling along multiple different semantically-relevant dimensions.

A further refinement is to use deliberate strategies for searching for relevance, usually acquired socially. Human thinking can move back and forth between contextually-cued recall and step-by-step inference, so as to hit on relevant considerations and reason with them; some hybrid LLMs do the same. (p. 172) Filling in a coherent scenario in the cognitive playground may itself suggest relevant considerations. My picture presents a solution which is partial and imperfect; but so is human cognition. Most of these elements have been discussed before in some guise, but my picture: emphasises the role of concepts as mediators, points to the power of the if-then way of avoiding the problem (while agreeing that this is not on its own a solution), and shows how this tactic can be re-purposed as a way of recalling information to use in reasoning, polling memory in multiple ways so as to approximate an isotropic search, albeit imperfectly.

(p. 173) In this way, the account of concept-driven thinking developed in the first half of the book has an important explanatory payoff: it shows how human cognition can dance around the frame problem, partly avoiding it and partially solving it.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0006

7
Drawing on Meaning

7.1 Introduction

This chapter is about the way in which inferences depend on content. More carefully, it is about the way transitions between representations relate to the content of the representations involved. Meaning is an absolutely central aspect of our mental lives. As we engage in thinking, the contents of the representations involved seem crucial to the way our thinking unfolds. RTM puts that under pressure with its commitment to capturing thought processes in terms of causal transitions between representational vehicles. This generates several philosophical puzzles about the role of content in cognition. I want to suggest that these debates, while being different, share an underlying assumption. The assumption itself is hard to state, and will take some work to uncover but, roughly, it is the idea that transitions between representations draw on their content. I will argue that, consistently with RTM, we can vindicate the idea that transitions between representations draw on their content in a substantive way.

I start off by mentioning these different philosophical debates in order to bring the underlying assumption about content into sharper focus (§7.2). Then I present the standard account of the role of content in transitions between representations (§7.3). But that account is incomplete. Section 7.4 recalls the distinction between content-specific and content-general transitions introduced in Chapter 3. The standard account covers only content-general transitions. With content-specific transitions, content is involved in a different way. I argue that content-specific transitions depend on the content of non-logical concepts in a way that content-general transitions do not. It may seem implausible that the thinker is somehow drawing on meaning less when they infer with an explicit representation than when the same information figures in thought only implicitly. Section 7.5 defends that conclusion.

7.2 The Phenomenon: Drawing on Meaning

The precise nature of the phenomenon I am targeting is hard to pin down. It is itself a matter of philosophical debate. At issue is the way semantic contents are involved in transitions between representations: that they are responsible for or explain a transition, or that there is some other kind of close dependence between a transition that takes place and the content of the representations involved. I am not here aiming to give an account of the metaphysics of that relation (whether it is metaphysical dependence, or some kind of explanatory or in-virtue-of connection). My concern is to argue that the evidence about how conceptual representations are involved in cognitive processing, laid out over the foregoing chapters, implies that there are two rather different kinds of involvement. (The various metaphysical options can be applied to both.) This section brings the phenomenon into slightly sharper focus by laying out some of the philosophical debates in which it operates as an underlying assumption.

I start with an analogy. Linguistic processing exhibits a contrast between cases that involve processing meaning (the majority) and those that do not. We sometimes hear speech merely as a string of sounds. In most cases we also apprehend the meaning of the words and sentences (Moore 1953, p. 59, quoted by Bayne and Montague 2011, p. 6; Drożdżowicz 2019). In the latter case, the content of what is said comes to figure in our thought (Fricker 2003; Longworth 2016). Psychology and psycholinguistics make use of a distinction between semantic and nonsemantic processing (Kroll et al. 2010). For example, it has been argued that dyslexic children show a specific deficit in phonological processing whereas autistic children are impaired in reading for meaning (Frith and Snowling 1983). We can explain this difference in terms of whether or not mental representations corresponding to the meaning of the sentences are accessed. (Linguists would say: whether semantic representations are accessed.) A string of speech sounds may be processed phonetically but not semantically, so no mental representation of the meaning of the sentence is produced (be these conceptual representations, or representations of other kinds, like sensory images).

We cannot make the same move at the level of thought. Thinking with a mental representation cannot be a matter of comprehending or interpreting the representation, as it is with language, on pain of launching a vicious regress. Nor can it be a matter of looking up the referent (Quine 1968). So the linguistic contrast presupposes that processing mental representations involves drawing on their semantic content, in a way that manipulating a public language sentence phonetically or syntactically need not.

The idea that thinking draws on thought content figures prominently in theorising about concepts. The meaning or content of a concept is supposed to explain the way the concept is used in categorisation and inference (Machery 2009, pp. 9–12; Weiskopf 2009a; Hampton 2015; Vicente and Martínez Manrique 2016), sometimes with different kinds of content (e.g. prototypes, theories, sensory images) underpinning different kinds of cognitive role (Millikan 2000; Camp 2015; Strevens 2019, pp. 64–5; Margolis and Laurence 2019).

The debate about the causal efficacy or causal-explanatoriness of content is committed to the same underlying assumption. The objection is that the classical computational theory of mind undermines the role of semantic content (Block 1990; Rescorla 2012, 2014), especially if contents are externalist (Fodor 1991; Figdor 2009). The assumption that semantic contents are closely involved in representational transitions is shared by both sides in this debate. Even those who allow that contents are causally inefficacious argue that contents play an important explanatory role (Egan 1992; Peacocke 1993; Shagrir 2001; but cf. Stich 1983). The shared assumption is that there is some close dependence between the content of a representation and the transitions in which it figures. I am not here aiming to takes sides in the debate about whether content is causally efficacious or causally explanatory. My focus is on elaborating the underlying assumption, namely that thinking draws on thought content in some way (I will say: in more than one way).

Ruth Millikan considers and rejects the argument that meaning externalism prevents thinkers from 'knowing what they are thinking about' (Millikan 2000, p. 95). She does think there is a challenge: to show how thinkers can know that their thoughts are not empty or equivocal. In Millikan (2000) she puts forward a comprehensive answer (see esp. pp. 95–108, building on Millikan 1984). The underlying assumption is that there is a commonsense phenomenon of 'knowing *what* one means in making a judgement' (Millikan 1984, p. 322) or 'knowing what one is thinking of' (Millikan 2000, p. 95).

The more recent debate about the nature of understanding, and what it is to grasp a subject matter, make a similar assumption. In that debate, grasp of content is involved in the way thinkers make transitions between representations (Strevens 2008, 2010; Grimm 2012). Those in favour of cognitive phenomenology are also interested in thought content. They are usually concerned to show that differences in thought content show up as differences in phenomenal character (Bayne and Montague 2011, pp. 15–17, citing a suggestion in Kripke 1982, p. 41, cf. p. 43; Jorba and Vicente 2014). Thus they too presuppose that thought content plays a critical role in the way our thoughts unfold.

These are all cases involving thoughts at the object level, representations that are about the world outside the thinker. As soon as we reflect on the role of thought content in our mental life, however, it is natural to start thinking in terms of self-attribution (of beliefs, or of other contentful mental states). In debates about self-knowledge, philosophers on all sides start from the assumption that the thinker can know the content of her beliefs (Brown 1995; McLaughlin and Tye 1998). Again, this is underpinned by the idea that inferences in thought draw on content: that self-ascription of a mental state draws on that state's content (McKinsey 1991, p. 11; Wright 2000, p. 152).[1]

This whistle-stop tour has briefly highlighted a series of different debates. The issues are different in each, but they share the underlying assumption that content is intimately involved in the way transitions between representations unfold. The point of this chapter is to show that that phenomenon comes in two importantly different varieties. This observation will not by itself resolve any of the philosophical issues canvassed above. However, we have a better chance of doing so when we have a (psychologically well-grounded) understanding of how transitions relate to content—namely that inference patterns are related to representational content in two quite different ways (in the content-specific and content-general cases).

7.3 Semantic Inference and Syntactic Inference

As we have seen, the great insight of RTM is that processing can be so-arranged that vehicle-vehicle transitions respect the contents represented. For example, when a thinker is caused to move in thought from representing p and p→q to representing q, that transition respects content in a strong way: it produces outputs that are guaranteed to be true if the inputs are true. When perceptual processing transitions from a contrast map to a representation of the location of edges, the output, although not guaranteed to be correct, is very likely to be correct, in normal environments, if the inputs are correct. So this too is a useful transition to make. Both are cases where causal transitions between vehicles unfold so as to respect content.

This is quite unlike the way things work with familiar public representations, like spoken words or written sentences. Agents react appropriately to sentences because they interpret and understand them. With mental representations there is no need for an internal understander. The meaning or content of a representation usually depends on its relational properties, characteristically on some kind of complex relation to the objects and properties it represents. Similarly, the truth or falsity of a representation usually turns on facts that are extrinsic to it. But RTM does not call for an internal homunculus that interrogates these relations. Processing proceeds in virtue of vehicle properties. If R1 being tokened causes R2

[1] Even Carruthers (2011b), who contrary to 94 per cent of philosophers (p. 17) rejects the transparency of belief self-ascription, has a view in which belief content is explanatory in the process of self-ascription (it orchestrates the process without figuring in experience).

to be tokened, that would occur even if R1 had been false, and even if the relational properties of R1 and/or R2 had been different such that their contents were different. The insight of RTM is that a physically-driven engine can, if appropriately configured, amount to a semantic engine.

RTM comes with a standard account of how thought processes are sensitive to content. Consider again the Socrates inference:

(1) All humans are mortal.
(2) Socrates is a human.
∴ (3) Socrates is mortal.

The transition from (1) and (2) to (3) occurs because the premises have a certain form (all Fs are Gs, x is F). Representations have formal properties, properties which are local properties of representational vehicles (in the representational system). These properties cause processing to unfold in certain ways. Although they are not the whole story, formal properties are involved in fixing content. They determine the way the content of the complete thought is determined by (or related to) the contents of its constituents. Manipulating representations by formal properties is a mechanism by which thought processes are sensitive to content.

This is the insight that underlies most human-created computing machines (i.e. digital computers). It has proven to be enormously powerful. Nevertheless, I want to argue that it is only half the story. The inference (1)–(3) can be performed without drawing on the meaning of ‘Socrates’, ‘human’, or ‘mortal’. For instance, if one holds that grasp or understanding is important, the inference could be performed without grasping or understanding the meaning of these concepts. It is somewhat analogous to processing a sentence without apprehending its semantics, as discussed above.

Computational approaches to language processing draw a distinction between syntactic and semantic inference (Schubert 2019). Syntactic inference starts with the grammar of a sentence and performs processes such a transforming its surface form into logical form. Semantic processing deals with the meaning of the words involved. Psycholinguistics makes a similar distinction. For example, bilingual speakers who learn their second language late treat the two languages differently. A leading model appeals to an asymmetry in the way translation between the two languages occurs. Translating from their second language to their first is accomplished merely lexically, without accessing the meaning of the words. Translating from their first language to their second is necessarily semantically mediated (Kroll and Stewart 1994; Kroll et al. 2010).

A significant problem in computational linguistics is to account for the semantic inferences people make between sentences, that is to say, the inferences that do not simply depend on syntax and logic (Pado and Dagan 2016); for example, from 'A hurricane hit Peter's town' to 'Peter's town was damaged'.[2] The standard RTM account of the role of content in thought treats all transitions between representations on the model of syntactic inference. That could have turned out to be right—it has been enormously powerful in computer science—but as we have seen, in human psychology there is also likely to be something going on that is more like semantic inference.

7.4 Content-Specific Transitions Draw on More Contents

Chapter 2 drew a graded distinction between content-specific and content-general (i.e. non-content-specific) transitions. The Socrates inference is an example of a content-general transition. The transition is faithful to content, *a fortiori* its being faithful to content depends on the content of the representations connected by the transition; however, its faithfulness to content depends only on the content of the broadly-logical concepts involved. It does not depend on the specific contents of the non-broadly-logical concepts (HUMAN, MORTAL, SOCRATES).

Which conclusion is reached of course depends on the contents entered at input. If the first premise had concerned Aristotle rather than Socrates, the conclusion would have been different (it would have concerned Aristotle). But making this inference does not demonstrate a grasp or understanding of the specific content of SOCRATES. It does not show that the thinker knows what she means in judging the conclusion (Millikan 1984, p. 322), or that she knows what she is thinking of (Millikan 2000, p. 95). It does not require her to draw on the meaning of her SOCRATES concept; nor is the content of SOCRATES needed to explain why she makes the transition. If she were making it in language, it could be a purely syntactic inference.

Contrast content-specific transitions. For example:

    (4) Fido is a dog
∴ (5) Fido barks

As argued in Chapter 3, if this transition is made without the benefit of an explicitly-represented general premise, then it is content-specific. Its faithfulness to content depends on its being about *dogs* and *barking*. By taking for granted that *dogs bark*—implicitly encoding the information that *dogs bark*—this

[2] Deep neural networks trained on huge bodies of linguistic data so as to be able to predict which words are likely to come next after any string of words or sentences given as input—large language models—now show considerable facility with semantic inference, but it is not yet clear how they achieve this feat (Bubeck et al. 2023).

transition depends on the specific contents of these concepts, and reveals a grasp of those contents.

This is an example of a direct-CS transition. We saw that there are several psychologically plausible examples (§3.4), including cases where transitions between thoughts take place in a high dimensional semantic state space. Equally important is the class of mediated-CS transitions. Chapter 5 discussed in detail how these work. One example is using sensorimotor simulation to work out whether a chair will fit in a car. Another example is relying on simulations in the hippocampal cognitive map to work out how to get from the pub to the library. These kinds of inferences (recall that 'inference' is understood broadly) rely heavily on content-specific transitions: the transition from the concept CHAIR to a sensory image; the transformation of that image under simulations of actions; and the categorisation of the resulting situation under FITS or DOES NOT FIT—all these involve content-specific transitions between representations.

Chapter 4 catalogues a large variety of informational models in which information is encoded in ways such that inferences over the model involve content-specific transitions. In all of these cases, the dispositions to transition between representations take for granted, or implicitly encode, information about specific referents. Those assumptions may be false (Machery 2017, p. 222). When they are correct, making the transition relies on a substantive fact about the referent. Making the transition demonstrates a grasp of the contents involved. It shows that the thinker knows, in some way, what it is that she is thinking of. This is a second way in which transitions between representations are sensitive to or explained by content. We can thereby, without making it mysterious, make sense of the idea that an inference—a transition between representations—draws on conceptual content.

What is this 'intimate connection' between contents and the disposition to make the transition? In many cases the thinker will have acquired the content-specific disposition because of the meaning of the concepts. Consider a child who already has the concept DOG and then learns that dogs bark. They may put this knowledge into practice so often that X IS A DOG → X BARKS becomes for them an automatic transition. The child is disposed to make that move in thought because DOG means *dog*, BARKS means *barks*, and dogs do bark. The disposition is in place and the transition is made in virtue of the content of the concepts involved. Content-specific transitions are acquired and stabilised over time in virtue of content. Exercising the disposition, then, is a way of drawing on those contents.

There is also a stronger version of this thesis according to which the relevant transitions are part of what fixes content. We can extend that idea beyond broadly-logical concepts (Rescorla 2012, 2014) and apply it to the concepts that figure in content-specific transitions. The thesis would then be that a transition like X IS A DOG → X BARKS is part of what makes it the case that DOG refers to *dogs*. That need not take us back to the old definitional theory—that the content has to meet a definition specified by the role. Not all content-specific transitions need come out as faithful. On some views, contents are assigned so as to maximise truth preservation (a principle of charity). On other views, we look to occasions when relying on the disposition led systematically to survival and reproduction, or some other process that stabilises the disposition. It is transitions made on those occasions whose faithfulness-to-content plays a role in fixing content. (Shea 2018 makes this argument in respect of subpersonal representations.)

On any of these views, content-specific transitions play a role in determining reference, and those which come out as correct, according to that content-determination, draw substantively on meaning—either because the thinker makes the transition because of a prior grip on the content, or because the disposition to make the transition is part of what gives the concept its content. Thus, a correct content-specific transition draws on the content of the concept.

A merit of the standard RTM story about the causal role of content is that it captures a role for content while holding on to the central commitment of RTM: namely that there are vehicles of content, and that transitions between representations are causal transitions between those vehicles. The standard story gives us a good account of how content is involved in transitions, but only shows us how the contents of broadly-logical concepts are involved. Content-specific transitions have a different form, but in this section we have seen that we can still vindicate a tight connection between content and transition, and do so in a way that is just as compatible with the central commitment of RTM. In the case of content-specific transitions, making the transition draws on the content of the non-broadly-logical concepts involved.

7.5 Are Content-Specific Transitions Really So Different?

The existence of inferences in thought of a kind that I would classify as content-specific is nothing new (Sellars 1953; Fodor et al. 1975). The types of representational structures, informational models, and computational processes I point to are also fairly widely accepted. What is not widely accepted is that contents are involved in representational transitions in two different ways. The difference turns on whether the information which underpins an inference—for example the fact that dogs bark—is explicitly represented or not. But is that really an important difference? And can it be that the thinker draws less on content when they represent information explicitly?

The difference is important, first, because it is important to have an empirically well-grounded account of the type of knowledge representations and cognitive-computational processes the mind actually uses. As I have argued, whether information is represented explicitly or is merely implicit in a processing disposition makes a big difference to how it figures in cognitive processing and in the mental life of the subject. The thinker can use explicit representations in a range of ways that are simply not available when information is represented only implicitly. The distinction is also empirically tractable (although difficult, as is often the case with investigating representations).

Second, recognising the class of content-specific transitions widens the way we think about the database in which information connected to a concept is stored. As well as a store of explicit memories, important information is stored in direct-CS dispositions, and also in dispositions to make transitions between conceptual and nonconceptual representations. In particular, it leaves space to encompass the rich class of what Camp calls 'characterizations': ways that we characterize the referent of a concept in sensory, motoric, evaluative, or affective terms (Camp 2015). Enlarging our theoretical perspective to include this wide range of ways in which people think about the things to which their concepts refer gives us a psychologically more plausible account. It also shows us that transitions between representations are often tied up with the content of the concepts out of which thoughts as structured (in addition to depending on their broadly-logical form).

We have seen that information that is implicit in a content-specific disposition can be made explicit and stored in memory for later use (§5.5). This brings into sharp relief the difference between the two ways that transitions can rely on content. To return to the examples in Chapter 5, I could reflect on my direct-CS dispositions and reach the conclusion that *birds have bills*; or I could rely on mediated-CS dispositions to reach the conclusion that *the chair will fit in the car*. In both cases, I now have an additional explicit representation, an extra premise that I can use in reasoning. Having formulated the conclusion about the chair explicitly, I can perform some further reasoning on it about what to do next.

My claim has a seemingly paradoxical consequence. Can it really be that when information gets made explicit the thinker operates with less understanding? Surely they now know more than they did? Granted, having made the information explicit (cf. Machery 2017, pp. 220–3), in one sense the thinker has a better understanding of the subject matter. They now know explicitly that *birds have beaks*, where as previously that was simply an implicit assumption of their dispositions to categorise things as birds. When they make an inference that has *birds have beaks* as a premise, they rely on this understanding (something that they know about birds, and about beaks). In this sense, they now know more about the subject matter.

Explicit representations can underpin further content-specific inferences, of course. But the explicit representation also makes possible broadly-logical inferences about birds, inferences that do not depend on understanding the specific content of the concept BIRD. Explicit premises give us material on which to perform that extremely general and extremely powerful form of reasoning. It is when the concept BIRD figures in inferences of this kind that the thinker is no longer drawing on its content. Making information explicit has opened up the possibility of inferences that do not draw on specific content. What has happened is that making content explicit has made reasoning less dependent on drawing on that content implicitly. The kind of grasp or understanding that exists in the content-specific dispositions involving BIRD is not being relied on when the thinker makes the information explicit and reasons with it logically. It's not that the thinker has lost this understanding—they still have the content-specific dispositions—it's just that the content-general transition does not draw on or display them. It is not actually paradoxical that explicitness should make a thinker able to make some inferences with less understanding. It is in fact a virtue, since they are not then limited by the understanding that is exhibited by their content-specific dispositions.

7.6 Conclusion

This chapter aims to bring a central but somewhat enigmatic phenomenon into slightly sharper focus. It is an assumption of some psychological theorising, and of several philosophical debates, that transitions between representations depend on content. They draw on the meaning of the concepts involved. The framework developed in the book allows us to throw some light on this phenomenon. It is more articulated than is often realised. There are in fact two rather different ways in which transitions draw on content.

This does not give us a positive theory of the causal efficacy or causal explanatoriness of content. But it does give us a more nuanced picture against which to build and assess such theories. Importantly, for present purposes, it shows how concept-driven thinking is deeply entwined with meaning. The standard RTM story about representational processing can seem anaemic. It feels as if all thinking is just logic chopping. The richness of the specific things we are thinking about seems to have been pushed off stage. The plug-and-play picture in Chapter 5 shows that much of deliberate thinking takes place in quite content-specific ways. In addition to direct-CS transitions between conceptual thoughts themselves, concept-driven simulations rely heavily on content-specific transitions: from concepts to special-purpose informational models; within those informational models; and from special-purpose informational models back into conceptual thought. This is not only a more psychologically realistic and full-blooded picture of how concept-driven thinking takes us to new conclusions. It also delivers a more satisfying account of how transitions between representations draw on meaning.

Chapter Summary

7.1 Introduction

This chapter is about the way transitions between representations draw on content, in a way that is presupposed by several different debates about the role of content in cognition. The standard account is incomplete: in the case of content-specific transitions, content is involved in a different way.

7.2 The Phenomenon: Drawing on Meaning

(p. 178)[3] This section brings into focus the claim that contents are involved in transitions between representations; it looks at some philosophical debates in which this operates as an underlying assumption.

An analogy is the contrast between processing a sentence semantically, accessing a mental representation of its meaning, or merely non-semantically, for example phonetically. We cannot make the same move at the level of thought, so the linguistic contrast presupposes that processing mental representations involves drawing on their semantic content, without having to look up their reference or access some further representation. (p. 179) Theorising about concepts also presupposes that thinking draws on the content of concept-involving thoughts. Both sides of the debate about the causal efficacy of mental content are committed to the same assumption, namely that there is some close (causal or explanatory) relation between the content of a representation and the transitions in which it figures.

Millikan's account of how a thinker knows what it is they are thinking about assumes something similar, namely that what a representation means figures in the way thinking unfolds. More recent debates about understanding presuppose that grasp of content is involved in the way thinkers make transitions between representations. Similarly, debates about self-knowledge assume that self-ascription of a mental state draws on that state's content.

(p. 180) There are different issues in play in each of these debates, but they share the underlying assumption that content is intimately involved in the way transitions between representations unfold; this chapter shows that this occurs in two different ways.

7.3 Semantic Inference and Syntactic Inference

The insight of RTM is that processing can be arranged so that vehicle-vehicle transitions respect content in the sense that the output is sure to be to true or likely to be correct if the inputs are. This is so even though processing proceeds in virtue of vehicle properties, without interpreting the representations, or checking what they refer to or whether they are true. (p. 181) The formal properties of a complex representation determine the way its content is determined by the contents of the constituents; manipulating representations in accordance with form is thus a mechanism by which transitions are sensitive to content.

[3] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

This is the insight that underlies human-created computing machines; however, it is only half the story; the Socrates inference could be made without drawing on the meaning of HUMAN, MORTAL, or SOCRATES. Both computational approaches to language, and psycholinguistics, draw a similar distinction, that between syntactic and semantic inference. The standard RTM account of the role of content in thought treats all transitions on the model of syntactic inference; it has had difficulty modelling semantic inference.

7.4 Content-Specific Transitions Draw on More Contents

(p. 182) Recalling the distinction between content-general and content-specific transitions (§3.2), the faithfulness to content of a content-general transition does not depend on the specific contents of the non-broadly-logical concepts involved. The conclusion reached does depend on all the contents represented at input, e.g. Socrates rather than Aristotle, but the specific meaning of SOCRATES does not figure in any of the ways discussed in section 7.2.

Contrast the dogs/barks content-specific transition: by implicitly encoding that *dogs bark*, this transition depends on the specific contents of the non-broadly-logical concepts involved. (p. 183) Both direct-CS and mediated-CS transitions (examples in Chapter 5) draw on the content of non-broadly-logical representations. These cases illustrate a way that transitions between representations draw upon conceptual content in a substantial sense.

There are several ways of making more precise the claim that there is an 'intimate connection' between contents and the disposition to make a transition; first off, that the disposition was acquired in virtue of the contents of the concepts involved. A second, stronger thesis is that the transitions are part of what makes it the case that the concepts have the content they do. (p. 184) On both of these views it follows that a correct content-specific transition exemplifies the thinker's understanding of, and draws on, the content of the concept. While my account here is different than the standard RTM story, it preserves the central commitment of RTM, namely that transitions take place between representational vehicles and are causally explicable in terms of vehicle properties.

7.5 Are Content-Specific Transitions Really So Different?

The existence of the types of inferences that I would classify as content-specific is nothing new, but the claim that contents are involved in transitions in two different ways is. The difference turns on whether information is represented explicitly or is merely implicit; the difference is important, first, to be psychologically realistic. (p. 185) Second, because encompassing direct-CS and mediated-CS inferential dispositions widens the class of characterizations we need to consider as encoded with a concept. The difference comes into relief when we see what it takes to make information which is implicit in content-specific transitions explicit, so that it can be used more flexibly. A seemingly paradoxical consequence is that when information gets made explicit the thinker operates with less understanding—when, in another sense, they know more than they did. However, the thinker doesn't lose the implicit understanding; what making explicit does is make available a form of inference, the broadly-logical, which is less dependent on understanding the content; it is only when operating that way that the thinker is drawing less on content.

7.6 Conclusion

(p. 186) The phenomenon of dependence on content, presupposed in several philosophical debates, is in fact achieved in two somewhat different ways. Whereas the standard RTM story seems to make representational processing rather anaemic, assimilating it all to logic chopping, the way content-specific transitions depend on content is more full-blooded, showing that meaning is deeply entwined in the way concepts operate in plug-and-play thinking.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0007

8
Metacognition

8.1 Introduction

Our topic is deliberate conscious thinking. Conceptual thoughts orchestrate executive processes and are operated on by them. Deliberation unfolds step-by-step, with the signature of type 2 cognition, while drawing liberally on automatic processes, both between conceptual representations and beyond. This kind of thinking is performed by a person, rather than happening to them. Thinking at the personal level consists in the interaction of the capacities we have discussed. In a commonsense way, I can say that it is me doing the thinking.

Part of the commonsense picture is that the person doing the thinking appreciates what is going on. To put it more psychologically, conceptual cognition is a sphere within which metacognition operates. Section 8.2 offers a very brief introduction to the psychological literature on metacognition. Sections 8.3 to 8.5 then work through three types of metacognition that are especially relevant to thinking with concepts. First, thinkers have an appreciation of the reliability of a concept—its reliability as a tool for categorisation and inference (§8.3). Second, it seems likely that thinkers are furnished with a procedural indicator of the reliability of the inferential transitions that occur in thought (§8.4). Third, it is plausible that a set of representations assembled in the cognitive playground attracts an assessment of coherence (§8.5).

This is not a complete account of metacognition in deliberate thought. Far from it. However, it does show that the account presented in previous chapters can vindicate the important idea that the goings-on of concept-driven thinking are appreciated by the person doing the thinking.

8.2 Metacognition

The book focuses on concept-involving processes at the personal level. We started with the idea that thinking with concepts is something the thinker does: it is an exercise of agency (philosophy), it engages executive functions (psychology). Dual systems theorists ascribe it to system 2. Whether or not one endorses a dual systems theory, deliberation with concepts does have many properties that are taken to be characteristic of type 2 cognitive processes. It proceeds step-by-step, effortfully, reflecting the person's goals and intentions, and is susceptible to interference by concurrent cognitive load. Along with the idea of being agentive comes the assumption that the thinker knows what is going on in their thinking. Putting it vaguely to start with: they have some appreciation of their thoughts and of their thought processes. In this section I argue that one important aspect of this appreciation is well-captured by the psychological literature on metacognition.

As a preliminary, we should distinguish metacognition from another way that a thinker can be said to appreciate their mental states and processes—by those states and processes being conscious. Conscious states are part of a thinker's mental life. Simply in virtue of being conscious, these states are in the mind, and processes involving them are processes taking place in the mind. That is not reflective or reflexive appreciation; but it is a substantial sense in which a person appreciates what they are thinking. When someone says that they are aware of ice on the driveway, for instance, that can simply mean that they have a conscious representation of ice on the driveway. Conscious states are states of awareness. Higher-order theorists of consciousness go further and argue that some second-order state is involved. Leaving this higher-order claim to one side, all theorists allow that conscious states are states of subjective awareness. In this way, an episode of conscious deliberate thinking is, by that very fact, something of which the thinker is aware. They appreciate their thoughts and thought processing in the sense that those states and processes are conscious.

However, that is not all. When it comes to deliberate thought, higher-level cognition is also involved: metacognition. Human cognition includes self-directed second-order (i.e. reflexive) representations that are formed through deliberation (i.e. are reflective). The strongest form of reflexive appreciation is reflective self-attribution of mental states and processes. A thinker can explicitly represent *I believe that p*, and thereby come to know that they do. They can also explicitly and reflectively represent thought processes, for example thinking, *I judged that he was unjust because I saw how he shared out the profits.*

In experimental psychology, metacognition is usually defined as being a matter of monitoring and control (Nelson and Narens 1990). Monitoring is keeping track of psychological states, for example by metarepresenting their contents, or keeping track of the way a psychological process operates, for example by representing the speed or fluency with which it unfolds. Control is a matter of using the result of monitoring to affect subsequent processing. In philosophy, Joëlle Proust influentially defines metacognition as 'the set of capacities through which an operating subsystem is evaluated or represented by another subsystem in a context-sensitive way' (Proust 2013, p. 4). These definitions are quite broad. Metacognition in this sense could extend to many psychological processes, including in special-purpose systems, encompassing the subpersonal. For example, I have argued that the reward prediction error signals involved in model-free reinforcement learning, which are widespread in the animal kingdom, are a form of metarepresentation (Shea 2014c). Carruthers agrees that there are forms of what he calls 'non-conceptual' or 'model-free' metacognition (Carruthers 2021; Carruthers and Williams 2022).

However, that is outside the paradigm. Research on metacognition usually has a more restricted focus. It is organised around personal-level forms of monitoring and control. 'Control' here connotes executive functions and control by the person—that is, cognitive processes which operate in the type 2 way (what Shea et al. 2014 call 'system 2 metacognition'). While special-purpose systems may include dedicated components for monitoring aspects of the circuit and affecting what happens next, it seems that we also have the capacity to monitor and control type 2 processes. We have already seen an example of the way metacognition operates in deliberation. We learn strategies for thinking and apply them deliberately: how to recall information from memory; how to take a new perspective or randomise to generate new ideas for solving a problem. While some of our metacognitive abilities are supported by mechanisms that are specific to a domain, like perceptual decision-making or memory recall, we may also have a general-purpose capacity for metacognitive monitoring and control (Mazancieux et al. 2020; Rouault, Lebreton, and Pessiglione 2023). Whether that is so remains an open empirical question.

As an aside, we can ask how there could be a general-purpose capacity for monitoring and control of deliberative processes. One plausible answer, with respect to monitoring, appeals to the general-purpose representational capacity of conceptual representation. The format of conceptual thought does not restrict the subject matter that concepts can represent. Being able to represent more-or-less anything, concepts can represent other thoughts and thought processes. So children can learn, for example, concepts of belief and desire that they use to represent other people's mental states in a reportable way (Wellman, Cross, and Watson 2001; Low and Perner 2012). These concepts seem to be acquired culturally, with variation across cultures in when and whether children come to represent others' mental states (Heyes and Frith 2014). In the populations typically studied in experimental psychology (Henrich, Heine, and Norenzayan 2010), the capacity is acquired by children between the ages of three and four years old.[1] When one has acquired concepts of mental states (e.g. belief) and processes (e.g. remembering), these concepts are available for one to apply to oneself. They can then be a basis for exercising cognitive control. Control processes are available simply because deliberate cognition is a sphere where executive processes operate, doing so in the characteristic type 2 way (with capacity limits and interference between tasks). Furthermore, since deliberate thinking takes place in the cognitive playground, control is exercised in a way that can reflect the thinker's goals and intentions. In short, there are probably good reasons why deliberate thought is subject to metacognition in the strong sense of monitoring-by-concepts and personal-level agentive control.

Returning from that aside, the discussion in this chapter does not depend on how deliberation comes to be subject to explicit monitoring and control, just that it is. Deliberate thought processes are part of the thinker's mental life in the sense of being states of conscious awareness and, furthermore, they are subject to reflexive and reflective appreciation. Explicit, conceptual self-attribution is reflective as well as reflexive. It is a form of what Joëlle Proust calls 'analytic' metacognition (Proust 2010, 2012, 2013). This contrasts with 'procedural' metacognition: signals that monitor aspects of a cognitive process, for example its fluency, and play a downstream role, but without representing it conceptually. Many procedural metacognitive signals are conscious, in the form of epistemic feelings. A familiar example is the tip-of-the-tongue phenomenon: the feeling that someone's name, or the answer to a question, is on the tip of one's tongue. People report feeling like they know the answer but can't yet bring it to mind (Schwartz 1999). This feeling is a moderately reliable guide to whether they will eventually be able to recall the information.

A central example of an epistemic feeling is confidence. The confidence attached to a representation affects how much weight is given to it when integrating information and taking decisions (Ernst and Banks 2002; Bahrami et al. 2010; Donoso, Collins, and Koechlin 2014; Lee, Shimojo, and O'Doherty 2014; Meyniel and Dehaene 2017). Conceptual representations recalled from memory—semantic memories—attract a measure of the reliability of the memory (Koriat, Lichtenstein, and Fischhoff 1980); perceptual representations and decisions come with measures of uncertainty and confidence (Fleming and Daw 2017); and conclusions reached through reasoning attract a measure of reliability in the form of a feeling of rightness (Thompson and Johnson 2014). (The usefulness of measures of confidence or reliability in weighing information may suggest that we should expect all representations in the cognitive playground to be accompanied

[1] Special-purpose systems may keep track of mental states from a much earlier age for dedicated tasks such as gaze following (Apperly and Butterfill 2009; but see Heyes 2014; Barone, Corradi, and Gomila 2019).

by such signals: Shea and Frith 2019.) Procedural signals of confidence are generated automatically (Proust 2013). Their cognitive role depends on how we have learnt to interpret them (which may also become automatic). For example, when a fact is retrieved from memory the representation is accompanied by a feeling of certainty or uncertainty (Koriat, Ma'ayan, and Nussinson 2006), and this affects whether the thinker will go on to rely on that information (Koriat and Helstrup 2007). It is likely that many of these dispositions are acquired culturally (Heyes et al. 2020).

The research that is most relevant for our purposes is the large body of work investigating metacognition of the accuracy of thoughts and thought processes. This is a rich field with lots of interesting details, but at the most general level it shows that people do keep track of the accuracy of their beliefs, in the form of judgements of truth, accuracy, or confidence in an answer they have just given (Schwarz 2015). These judgements show a moderate correlation with whether the belief really is likely to be true (metacognitive sensitivity), with some biases towards under- and over-confidence (metacognitive mis-calibration), both showing variation across different types of task. Judgements are partly based on the thinker's beliefs and other explicit information (analytic metacognition), and partly on experiential information or epistemic feelings (procedural metacognition). Fluency is a particularly potent experiential cue. Fluent processing tends to boost every kind of metacognitive assessment, whereas disfluent processing gives rise to doubts. For example, when answering a general knowledge question, disfluency causes the thinker to have less confidence in the answer they have given.

In a review paper, Schwarz (2015) lists the factors which have the largest effect on people's assessments of truth or accuracy. These all reflect the way thinking operates as a cognitive process over a collection of information in the cognitive playground. The extent to which a belief is supported by evidence is important, as is its compatibility or incompatibility with the thinker's other beliefs. Thinkers also take into account how well a belief fits into an overall narrative or other mental model. Interpersonal aspects are important, like the credibility of the source of the belief, and especially the extent to which the belief is shared by others (consensus). The overall picture in the literature is of factual representations being widely monitored for accuracy, with that happening in a variety of different ways.

Thus, metacognition in the cognitive playground, both analytic and procedural, is likely to be an important determinant of how deliberate thinking works—of how processing unfolds. Psychological research does not yet give us a complete account of all of this. Nor is there scope here comprehensively to review the rich lines of research that exist. What I will do is to discuss three kinds of metacognition that are particularly relevant: of concepts (§8.3), inference (§8.4), and coherence in the cognitive playground (§8.5). These examples will serve to make the case that metacognition is a crucial aspect of the way deliberate thinking works, thus vindicating the idea that the goings-on of concept-driven thinking are appreciated by the thinker.

8.3 Appraisal of Concepts

When theorising about the way metacognition applies to deliberate conceptual thought, an obvious first place to turn is to concepts themselves. Are concepts subject to metacognitive appraisals of any kind? A concept is a tool for thinking and so clearly could be appraised against the various uses to which it is put. Judgements are directed at forming correct representations and so can be appraised for accuracy; memory recall aimed at remembering all the items on a list can be appraised for exhaustiveness; and so on (Proust 2013). In what ways can a concept be appraised?

The use of a concept can be appraised relative to the various cognitive tasks for which it is deployed, for example categorisation, induction, and reasoning. For categorisation, we could ask how accurately the thinker is able to categorise new samples under the concept. Ruth Millikan argues that there are systems for checking whether a concept picks out the same thing when it is applied in different ways, for example through the way a thing looks and by what it sounds like (Millikan 1984, pp. 142–5). She argues that where these generate a contradiction, we become less disposed to categorise in those ways. Another way to use a concept is in induction. Relative to that use we can ask how reliably properties observed in a member of the category project to new instances, and how many properties are available for projection. We sometimes collectively assess concepts in the course of academic enquiry. For example, there is a debate in philosophy of biology about the usefulness of the concept of innateness for categorising traits and forming expectations about them.[2] We are interested in whether something similar happens individually, within the cognitive life of a thinker.

The book is focused on the way conceptual representations are involved in working out what is the case and what to do. Relative to this activity the relevant metacognitive assessments will be broadly epistemic. We can divide ways of thinking about the reliability of a concept, *qua* cognitive tool, into two broad concerns. The first concern is with the tool itself: is it a useful way of dividing up things in the world, does it support reliable inductions, does it get on to a deep and widely-applicable distinction? These can be thought of, either as meta-level questions about the concept, or as object-level questions about the category to which it refers. Is this a good category for thinking and reasoning with? The

[2] Griffiths and Linquist (2022). I share the view that it is a bad concept for these purposes, especially in the cognitive sciences, albeit with an explicable inductive underlying basis outside of humans (Shea 2012).

second concern is with the thinker's individual perhaps idiosyncratic grasp of a concept. For a given concept C, I can ask myself things like: do I have a rich collection of information stored with C, or only a thin grasp; is the information that I do represent likely to be accurate?

These suggestions are intuitively plausible but it is an empirical question whether people reliably appraise concepts in these ways. I collaborated in a series of experiments that have begun to investigate this question (Thorne et al. 2021, 2022). The large body of work on metacognition in other areas suggested that this would be a promising area of investigation. There was prior work asking people to assess how well they would be able to categorise novel items under a recently-learnt concept. This suggested that people would understand metacognitive questions specifically directed at concepts (Jacoby, Wahlheim, and Coane 2010).

We found that people do indeed reliably appraise concepts in both of the two broad ways mentioned above. As a tool, they appraise, for example, how reliable a concept is for induction: for projecting an observed property to other members of the category; and how informative it is: how much the concept tells you about what a thing is like. These appraisals are reliable in the sense that different people largely agree in their appraisals across a wide range of concepts. These assessments are also integrated with some of the central ways people use concepts. For example, as we proceed down a conceptual hierarchy (e.g. MAMMAL, DOG, SPANIEL), from superordinate through basic to subordinate level, concepts are rated as progressively more informative and as being a more secure basis for induction.

Turning to questions about how well thinkers grasp a concept, we found that people do appraise how well they understand different concepts—for example, how much they know about the category and whether most of the things they believe about it are likely to be true. Here too there is broad intersubjective agreement in how people appraise a diverse range of concepts along these dimensions. Four different ways of appraising how well understood a concept is in fact correlate, as rated by different individuals across the same group of concepts. This suggests that we have a common underlying sense of how well understood each concept is—a *sense of understanding*. We went on to discover that concepts which rate highly on sense of understanding are a preferred basis for making inductive inferences.

These results suggest that the information that a concept gives us access to includes information about its reliability for various cognitive purposes. This metacognitive information forms part of the characterizations that are connected with a concept. That is a novel point: that characterizations go beyond object-level information about the category and include meta-level information relevant to thinking with the concept. It is not yet clear whether this information takes the form of an explicit representation of reliability (analytic metacognition), or whether, alternatively or in addition, there are epistemic feelings generated by the use of a concept (procedural metacognition). In a far-from-exhaustive study we found evidence for the former and not the latter (Thorne et al. 2022). It is likely, however, that processes in which conceptual representations are involved do generate forms of procedural metacognition—processes like inference, which we turn to next (§8.4).

The studies discussed in this section just scratch the surface of a potentially rich area of investigation. Experiments on metacognition about decision confidence, reasoning, memory recall, and so on, made it highly plausible that other aspects of conceptual thought—when it also takes place deliberately, in the type 2 way—would be subject to metacognitive appraisal. These results confirm that prediction and give us a preliminary indication of some forms which that appraisal takes. They fill in another part of the picture of how the thinker appreciates what is going on when they are deliberating.

8.4 Reliability of Inference

We have seen that metacognitive evaluations attach to many of the mental states that are involved in conceptual thought: to explicit conceptual representations of facts (semantic memories), to percepts, and to concepts themselves. In this section our attention turns from states to processes. Where procedural metacognition attaches to a recalled memory, that reflects aspects of the process that produced it (e.g. fluency). An inference is a mental process. We should therefore expect that conclusions reached through inference will be subject to metacognitive monitoring. This is another way that the thinker appreciates what is going on in conceptual thought processes, as we will see in this section. Furthermore, in the case of inference, it is plausible that a metacognitive appraisal attaches, not just to the output, but to the pattern of inference itself (Shea forthcoming).

Experimental work on reasoning has found that conclusions reached in reasoning elicit a 'feeling of rightness'—an epistemic feeling which reflects the thinker's confidence in the conclusion (Thompson, Turner, and Pennycook 2011; Thompson, Evans, and Campbell 2013; Thompson and Johnson 2014). The feeling of rightness can be probed directly by asking people to report how confident they are in the conclusion of a piece of reasoning and indirectly by looking at the skin conductance response as a measure of arousal (De Neys, Moyens, and Vansteenwegen 2010; De Neys, Cromheeke, and Osman 2011).

At the level of monitoring, the feeling of rightness reflects whether the conclusion follows logically from the premises, and also whether the conclusion is plausible—whether it fits with the thinker's background beliefs. For example, given a syllogism like the following, people tend say that the conclusion follows logically from the premises (when in fact it doesn't). The fact that the conclusion does not follow logically is reflected in a reduced feeling of rightness (De Neys et al. 2011).

All birds have wings.
Crows have wings.
Therefore, crows are birds.

a. Conclusion follows logically
b. Conclusion does NOT follow logically

At the level of control, the feeling or rightness affects what the thinker does next in their reasoning. A low feeling of rightness, for example when selecting the plausible but invalid conclusion (a) above, makes the thinker more likely to pause and reflect, bringing type 2 cognition to bear on the problem (Ackerman and Thompson 2017). This effect is observed in classic 'thinking fast and slow' experiments where the intuitive, automatic answer to a problem is incorrect (De Neys 2023).[3] Thinkers who experience a low feeling of rightness are more likely to stop and check the answer.

In these studies, inferences are taking place between conceptual representations. They show another way in which the thinker appreciates what is going on in concept-involving thinking. They have varying levels of confidence in the conceptually-represented conclusions reached through inference, levels of confidence that reflect both the plausibility of the conclusion and the nature of the inference pattern that produced it.

In addition, it is likely that patterns of inference are themselves subject to metacognitive assessment. Some patterns of inference feel reliable to the thinker, others less so. These epistemic feelings change over time as a result of the thinker's experience of the results performing inferences of that form. I have made an extended argument elsewhere for the existence of these 'feelings of reliability', and for the important epistemic role they play for the thinker (Shea forthcoming). Here I will simply highlight the way feelings of reliability constitute another way that the thinker appreciates what is going on in concept-involving thinking.

As we saw in Chapter 3, there are some forms of inference that the thinker is disposed to make simply in virtue of tokening premises of the right form. Modus ponens is one example. Simply tokening thoughts of the form *if p then q*, and *p*, disposes the thinker to token the thought *q*. There is no need for the thinker to entertain a further premise to the effect that the conclusion follows from the premises. On pain of regress, there must be some inferential transitions that the thinker is disposed to make without representing anything further (Carroll 1995). These transitions are representationally basic (or 'primitively compelling'; Peacocke 1992). We saw several examples in Chapter 3: thinkers have dispositions to make a number of types of broadly-logical transition. My claim is that these

[3] E.g. 'A bat and ball cost $1.10 together. The bat costs $1 more than the ball. How much does the ball cost?' Frederick (2005). (The answer is not $0.10.)

transitions are accompanied by a feeling of reliability, or in some cases unreliability, when the disposition is activated.

These feelings change over time in a way that reflects the downstream consequences of performing the inference. Those who teach logic will have observed changes like this in the logic class. Students arrive with a disposition to react to an *if...then* premise by affirming the consequent. That disposition becomes less pronounced with experience. In the other direction, some patterns of inference that are valid but less intuitive, like modus tollens, become more fluid and feel more reliable with experience. Many simple, commonly-used patterns of reasoning with AND, OR, NOT, and IF...THEN are logically valid, so we have a wealth of experience that using them does not lead to contradictions. By contrast, if we were to start using the putative logical connective TONK (Prior 1967), whose introduction and elimination rules allow the thinker to reach any conclusion whatsoever, we would soon run up against contradictions and confusion. So the putative inference pattern for TONK would be unable to build up a feeling of reliability.

My hypothesis is that feelings of reliability attach, not just to content-general, broadly-logical transitions, but also to content-specific transitions. For example, many people are disposed to infer with the following introduction rule for the concept SET:

(1) Some things have property F.
(2) There is a set of all and only the things with property F.

Experience in philosophy, however, shows us that inferring with SET in this way leads to contradictions. We become more tentative in using this pattern of inference. It is not that the disposition to infer from (1) to (2) disappears completely. But it comes accompanied by a lower feeling of reliability, so we are more likely to stop and check before endorsing the conclusion.

Research on other types of procedural metacognition makes it plausible that patterns of inference should generate a form of procedural metacognition in their own right, for example due to their fluency. That this should show up in the mental life of a thinker in the form of an epistemic feeling is suggested by a phenomenon that has long exercised philosophers (Brewer 1995; Kornblith 2012; Boghossian 2014). The phenomenon is the contrast that we experience between trains of thought of two different kinds. Most of the time, when we reach a conclusion in inference, it seems to us that the conclusion follows from the premises. That is the default case and usually goes unremarked. In other cases we find ourselves moved to a conclusion in thought but without its seeming to follow from what we were thinking before. Boghossian gives the example of an anxious depressive character who finds himself going from the thought, *I'm having so much fun*, to the thought, *but there's so much suffering in the world*. They are disposed to make the transition. The second thought is not just some mental intrusion—it is the content of the first thought that prompts the second. But it doesn't seem to the thinker that the second thought follows from the first. A factor which is present, if unremarked-on, in most of our inferential transitions, is lacking in this case.

If a pattern of inference generates a feeling of reliability, that would explain the contrast between normal inferences and cases like the fun/suffering example. Some patterns of inference are such that, although the thinker is disposed to form the conclusion when the relevant premises are tokened, it doesn't feel reliable in the way it does in the standard case. Another example would be the way philosophers make the transition (1)–(2) involving SET only tentatively, once they have been exposed to the paradoxes that can result.

As well as generating a phenomenological difference between the cases, the feeling of reliability can play an epistemic role for the thinker. It offers them an internal signal of whether the pattern of inference is likely to be reliable. (In the cases we are interested in, an inference pattern is reliable if the conclusion is likely to be true if the premises are true.) As with epistemic feelings of other kinds, the feeling of reliability will not be perfectly correlated with whether the inference is in fact reliable. However, if these feelings are affected by experience in the way I have suggested, we can expect them to provide at least a rough guide to reliability—and a warning signal when a pattern of inference has led to trouble in the past (as it would with TONK).

Furthermore, this epistemic feeling can play an important diachronic role. We sometimes reflect on our inferential dispositions, not just in philosophy, but also in everyday life. This reflection can be individualistic but is often interpersonal. Either way, what happens when we conclude that a pattern of inference ought to be avoided? We might have noticed that affirming the consequent generates problems (unless reasoning abductively, and even then it has to be handled carefully). Or that using SET in accordance with (1)–(2) needs caution. Other content-specific cases cover social and political rather than epistemic concerns. For example, it might be pointed out to me that I use the concept WOMAN in a sexist way. Thinkers need not differentiate between the different reasons why they ought to be reluctant to use a certain inference pattern. But how does this realisation, reached by reflection, come to impact their inferential dispositions? If they have time to reflect the next time the premises are encountered, then they may well remember the problem and so remember not to draw the sexist conclusion. But often we don't have time to reflect. We might be in the middle of a conversation, say. Won't the inferential disposition still then be triggered? An episode of reflection is unlikely to be enough to retrain the thinker's automatic inferential dispositions. But it may well be enough to erode the feeling of reliability that accompanies the inference. This puts the thinker on the road to being more cautious about forming the conclusion. The feeling of reliability offers a route by which the conclusions we reach by reflecting on our reasoning can impact our inferential dispositions into the future. That is a further important epistemic role that it can play.

My suggestion is that the feeling of reliability is acting as an internal indication for the thinker of whether a pattern of inference is likely to be reliable. Although the book is not about epistemology, and there is not scope to do justice to the issues here, I will briefly gesture in the direction of a potential epistemological implication of feelings of reliability. Patterns of inference are not just externalistically assessable for their reliability. The thinker has access, in the cognitive playground, to an epistemic feeling which signals reliability versus unreliability (albeit imperfectly). Thus, there will be cases where the thinker takes the inference to be reliable, because of the feeling of reliability, and the inference pattern is in fact reliable. That offers an internalist sense in which the thinker is justified in relying on the inference pattern. Does the thinker also need to know whether or not the feeling of reliability is itself a good guide to reliability? My suggestion is that they do not. The appropriate standard for assessing the epistemic standing of the epistemic feeling is simply whether it tends to be reliable or not. Otherwise we would launch a regress of justification. This adopts a way of blocking the potential regress which is suggested by Ernest Sosa (1985, pp. 240–3). Sosa argues that, for reflective knowledge, we need the presence of a second-order element to establish justification. However, that second-order element need not itself count as reflective knowledge. It is held only to the standards of bare reliabilism. Similarly, provided the feeling of reliability is at least a rough guide to whether the pattern of inference to which it attaches is in fact reliable (as work on other epistemic feelings would suggest), it can form the basis on which a thinker is internalistically justified in relying on that pattern of inference.

This fills in the picture of the way metacognition operates on the inferences performed in deliberate thought. Experimental work shows that conclusions reached through reasoning attract a feeling of rightness which affects what happens next in thought; and philosophical reflection suggests that inference patterns themselves attract varying feelings of reliability and unreliability. This further elaborates the ways in which a thinker appreciates what is going on in concept-involving thinking, with these forms of appreciation affecting the way the thinker performs inferences.

8.5 Coherence in the Cognitive Playground

So far we have looked at how metacognition applies to component parts of deliberation: concepts, inferences, and the conclusions we draw. In this section I turn to the wider cognitive playground. In the course of the book we have seen that conceptual thought can drive the construction of a suppositional scenario in the cognitive playground: a model of an actual or postulated situation that includes both conceptual representations and representations drawn from special-purpose systems (sensory, agentive, motoric, spatial, affective, evaluative, etc.). The thinker runs inferences over this model, simulating possibilities and their consequences as a way to work out what to do (or to reach new conclusions about what is the case). These inferences can take place locally, step-by-step, from one conceptual representation to another, but they can also be non-local, depending on many different factors in parallel and/or depending on the overall structural or configural properties of the model in the playground. These inferences are loosely analogous to model-theoretic inferences in logic: representations in the model set up restrictions on the way a world can be, and inferences in the playground fill in further possibilities, given the constraints and assumptions found in special-purpose systems. This is a picture of non-local, broadly abductive inference over an entwined collection of representations in the cognitive playground.

A scenario in the cognitive playground typically contains many different representational elements, both concepts, and representations drawn from special-purpose systems. Working memory researchers distinguish between the collection of information that has been activated, forming a model of the environment, and the focus of attention, which applies successively to small parts of this model at a time (§1.2; Cowan et al. 2021, pp. 47–9; Reuter-Lorenz and Iordan 2021, p. 285). The cognitive playground corresponds to the activated model. Working memory involves the manipulation of these representations. Working memory acts through the focus of attention to add to, manipulate, and actively remove representations in the cognitive playground based on their relations to one another. Thus, the representational elements in the playground are interconnected.

How precisely to characterise this interconnectedness is a substantial topic, one on which there is no consensus. Representations in the playground can be in tension with one another—whilst it seems possible for there to be contradictory representations in the playground, the playground exerts some pressure in favour of coherence and consistency. We see that at work with binocular rivalry. Although people presented with different images in each eye transiently experience a mix of the conflicting information coming from the two eyes, conscious experience settles into a coherent representation (Haynes and Rees 2005). Fortunately, an imprecise characterisation of this phenomenon is sufficient for our purposes. The playground is a shared representational space. Relations of support and contradiction are somehow inherent in the way information is represented in the playground, so that incoherence and inconsistency are readily apparent. Contradictory representations are not impossible, but they are in tension with one another, in a way that representations tokened outside the playground in separate special-purpose systems are not. This is related to the idea of the unity of consciousness: that co-conscious representations are interconnected and together form a single mental unity or gestalt (Bayne 2010). I have been studiously avoiding putting any weight on consciousness as such in my theorising but, if construed merely functionally, the unity of consciousness is the same kind of property as that which I am claiming representations in the cognitive playground exhibit.

In this section I want to argue that metacognitive appraisals apply to the overall model in the playground: that there is monitoring and control based on relations between representations in the playground, perhaps including due to its holistic, gestalt, or configural properties. A first line of evidence is the existence of so-called content effects in reasoning—belief bias, for instance. People are more likely to draw a conclusion that coheres with their background beliefs, and are reluctant to make an inference, even if it is deductively valid, if the conclusion conflicts with their other beliefs (De Neys 2012). These coherence effects generate metacognitive appraisals. For example, Koriat's self-consistency model explains confidence in terms of whether the items of information recovered by the thinker cohere or conflict with one another (Koriat and Adiv 2015). This seems also to apply interpersonally. Sperber et al. (2010) argue that, in conversation, the listener is constantly monitoring what they hear to see if it conflicts with their own beliefs (they exercise 'epistemic vigilance'). Relatedly, responses to a question that are consensual—endorsed by most people—tend to attract higher confidence (Koriat 2012a).

Coming at the issue from philosophy, we have seen that Millikan argues that judgements made in thought are constantly being tested for consistency (Millikan 1984, pp. 142–5; 2000; 2017, p. 80). This is ultimately an empirical hypothesis, but one with some psychological support, in addition to its intuitive plausibility (Shea 2023a).[4] Millikan's motivation is the same as the one behind this chapter, namely to say how a thinker appreciates what is going on in their thinking. Millikan's specific aim is to give an account of how the thinker 'knows what they are thinking of' (Millikan 2000, pp. 95–6 and 177–92). The context is Bertrand Russell's dictum that we cannot 'make a judgement or entertain a supposition unless we know what it is we are judging or supposing about' (Russell 1912, p. 58). That looks problematic for an externalist about meaning. Millikan disagrees with Gareth Evans's account in terms of capacities to distinguish between things. Her alternative is that thinkers ensure they know what they are thinking about by testing their judgements for consistency.

Millikan's hypothesis is a more specific version of my broader claim about relations of support and inconsistency in the cognitive playground. Koriat's self-consistency model of confidence also depends on a process like this being in

[4] Millikan's more specific proposal—that registering inconsistency makes the thinker less disposed to use the concepts involved and/or to apply them in those ways—although also plausible has not, so far as I am aware, been tested empirically.

operation. Across a range of tasks, Koriat and his collaborators find that the answers which a person gives consistently are ones which attract higher confidence than those which vary across trials (Koriat and Adiv 2015). (This may seem obvious, but many commonsense claims that we make about the operation of our own minds turn out not to stand up to rigorous investigation.) In Koriat's model, the confidence people attach to retrieved memories (for example, the answer to a general knowledge question) is driven by how much coherence there is between the various items of information they consider before giving an answer (Koriat 2012b). Taking more time and thus considering more items of information tends to decrease coherence and hence reduce confidence (Koriat 2012a). These results strongly suggest that representations in the cognitive playground are monitored for coherence and consistency in a way that affects the thinker's subjective confidence. I would argue that this is evidence of a metacognitive process operating on the cognitive playground as a whole.

In short, several lines of evidence and argument suggest that there are forms of metacognition which target aspects of the overall cognitive playground. That is a further way that the thinker appreciates what is going on in concept-driven thought.

8.6 Conclusion

I started with the idea that deliberation involves executive processes, including metacognitive monitoring and control. I have argued that thinking incorporates metacognitive appraisals of the concepts we use for thinking, the conclusions we draw in thought, and the patterns of inference by which we reach them. When conceptual thought drives simulations and the construction of suppositional scenarios in the cognitive playground, that model is assessed for coherence and consistency, affecting the thinker's confidence in what they are thinking. Confidence is an epistemic signal for the thinker of whether the model they are entertaining is internally coherent and consistent with the information that is recalled from memory or sampled from special-purpose systems. That deepens the sense in which the thinker appreciates what is going on in their thinking. As well as the representations and processes being conscious, they also have an appreciation of the overall coherence of the model which they are relying on to draw new conclusions and take decisions. Part of the problematic motivating the book is that the cognitive playground integrates information drawn from a range of special-purpose systems, as well as from conceptually represented memories, and uses it to reach new conclusions, including via non-local inferences. Metacognitive processes operate on this integrated whole and affect how the thinking process unfolds. They are part of the way the thinker appreciates what is going on in their thinking.

Chapter Summary

8.1 Introduction

Concept-involving deliberation is performed by the person and is subject to various executive processes. One aspect of this is metacognitive—that the person doing the thinking appreciates what is going on; this chapter considers three forms of metacognition that are especially relevant. The examples are illustrative, not exhaustive, but they serve to vindicate the idea that the goings-on of concept-driven thinking are appreciated by the thinker.

8.2 Metacognition

(p. 192)[5] One aspect of what it is for a thinker to appreciate what is going on in their thinking is well-captured by the psychological literature on metacognition. As a preliminary, I set aside a thinner sense in which a thinker appreciates what is going on, namely that conscious thoughts are states of awareness—the thinker is aware of their contents, in the first-order sense that their contents are part of his or her conscious mental life. Higher-level representations also feature; these are both reflective, i.e. formulated through deliberation, and reflexive, i.e. second-order directed at the self.

In psychology, metacognition is usually defined as consisting of monitoring and control of cognitive processes; this is applicable broadly, including to the subpersonal. (p. 193) Paradigmatically, however, research on metacognition focuses on personal-level monitoring and executive functions ('system 2 metacognition'), which are capacities that thinkers seem to be able to apply to many different aspects of personal-level cognition. As an aside, a potential explanation of how we could have a general-purpose capacity for personal-level metacognition is the flexibility of the conceptual system, which places no strong restriction on what concepts can represent.

(p. 194) Explicit, conceptual self-attribution is both reflective and reflexive—it is a form of analytic metacognition; this contrasts with procedural metacognition (e.g. the tip-of-the-tongue feeling): signals that monitor aspects of a cognitive process without representing it conceptually. A central example of the latter is the feeling of confidence (an epistemic feeling) that accompanies, amongst other things: semantic memory recall, perceptual decisions, and conclusions reached through reasoning. (p. 195) Particularly relevant is the large body of work on judgements of truth, accuracy, or confidence, which are affected both by explicit beliefs and by

[5] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

epistemic feelings; fluency is a potent experiential cue. These assessments are affected by a collection of factors which reflect how a belief fits into the wider cognitive playground (listed). Thus, metacognition in the cognitive playground is an important determinant of how deliberate thinking unfolds; this chapter discusses three selective examples (§§8.3, 8.4. 8.5).

8.3 Appraisal of Concepts

(p. 196) An obvious first question is whether concepts themselves are appraised, and in what ways. A concept can be appraised for its reliability in categorisation or induction. For our purposes these appraisals will be broadly epistemic; they divide into concerns with the usefulness of the concept or category as a tool in cognition, and questions about the thinker's own individual understanding of the concept.

(p. 197) I collaborated in a series of experiments that have begun to investigate how people metacognise their concepts. We found that people do reliably engage in concept appraisals of both kinds, and that these appraisals are reflected in other aspects of conceptual cognition, for example in the division of concepts into hierarchical levels (superordinate, subordinate, and basic). We discovered that there is a common underlying assessment of how well understood each concept is: a sense of understanding; concepts rating high on sense of understanding are a preferred basis for making inductive inferences. This establishes that the characterizations connected to a concept go beyond object-level information and include meta-level information relevant to thinking with the concept. (p. 198) These results fill in part of the picture of how the thinker appreciates what is going on when engaged in deliberation.

8.4 Reliability of Inference

Just as epistemic feelings are generated by cognitive processes like memory recall, we should expect a form of procedural metacognition to attach to performing an inference.

Experimental work has found that conclusions reached through reasoning attract a 'feeling of rightness'. This feeling is affected both by the plausibility of the conclusion and by whether it follows logically; when a plausible but non-logical conclusion is endorsed that is reflected in an increased skin conductance response and a reduced feeling of rightness. (p. 199) This affects downstream processes, like whether the thinker is disposed to engage in more reflection about the problem. This is another way that the thinker appreciates what is going on in deliberation.

It is likely that, in addition, patterns of inference themselves are subject to metacognitive assessment, that they feel more or less reliable to the thinker. My claim is that forms of inference that the thinker is disposed to make simply in virtue of tokening premises of the right form, e.g. modus ponens, are accompanied by a feeling of reliability (or sometimes unreliability). (p. 200) The feelings are titrated by whether the inference has worked well in the past or led to contradiction and confusion—compare increasing confidence using modus tollens with the manifest unreliability of Prior's TONK. Feelings of reliability also attach to content-specific transitions, for example the low reliability philosophers come to associate with the introduction rule for the concept SET.

If the inferential process itself generates a feeling of reliability, as other work on procedural metacognition suggests, that would explain the phenomenon, which has long exercised philosophers, that the thinker is usually not simply moved from premises to conclusion, but has a sense that the conclusion follows from the premises. (p. 201) Other inferences do not feel so reliable when the conclusion is drawn. The feeling of reliability can also play an epistemic role for the thinker, signalling when a pattern of inference is likely to be unreliable. Furthermore, it plays a diachronic role: when we decide by deliberate reflection that we ought to avoid a certain pattern of inference (involving SET, say), a subsequent feeling of unreliability can affect whether we are disposed to draw the same conclusion again, in the future, when inferring automatically and unreflectively. (p. 202) Epistemically, the feeling of reliability offers the thinker an internalistically-available indication of the reliability of an inference.

In short, thinkers appreciate what is going on in drawing inferences with conceptual representations both through having a feeling of rightness in the conclusion and through a feeling of the reliability of the inference.

8.5 Coherence in the Cognitive Playground

In this section I turn to processes taking place over suppositional scenarios that depend on characteristics of the cognitive playground. (p. 203) Representations in the playground are interconnected: working memory acts through the focus of attention to add to, manipulate, and actively remove representations in the cognitive playground based on their relations to one another. How to characterise this interconnectedness is controversial, but I rely just on the idea that relations of support and inconsistency are somehow inherent in the way information is represented in the playground.

(p. 204) There is evidence that some metacognitive appraisals arise from properties of the overall model in the playground, for example the effects of consistency with background beliefs, coherence, and consensuality. Working from philosophical considerations, Ruth Millikan argues that judgements made in thought are constantly being tested for consistency. Asher Koriat's self-consistency model, and the evidence behind it, also supports the claim that representations in the cognitive playground are monitored for coherence and consistency in a way that affects the thinker's confidence.

(p. 205) In short, several lines of evidence and argument suggest that there are forms of metacognition which target aspects of the overall cognitive playground; this is a further way that the thinker appreciates what is going on in concept-driven thought.

8.6 Conclusion

This chapter fills out a way in which concept-driven thinking, as described in the book, is an executive process, carried out by the person: the thinker appreciates what is going on metacognitively and that affects how the process unfolds.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0008

9
Concluding Thoughts

9.1 Deliberating with Concepts: The Picture

The aim of the book has been to paint a richer picture of concepts and their role in our cognitive life. Psychological research has concentrated on categorisation. While we now have a good idea about the various kinds of information that are involved in applying concepts to the world, there is much less work on how people use concepts offline. Offline thinking has always been firmly in the sights of philosophy, but the paradigm here is reasoning, moving from one conceptual representation to another step-by-step. There has been much less focus on the way deliberation draws on representations and computations in special-purpose systems. The dominant tradition takes all inferences involving concepts to be of the broadly-logical kind. These are undoubtedly important. Conceptual representations are distinctive in that they are constructed using a general-purpose mode of combination and can be reasoned with in ways that are content-general—transitions that do not depend on the specific content of the concepts involved. However, by focusing on the general-purpose aspects of conceptual thought theorists have tended to overlook the way conceptual thinking draws on special-purpose resources, like sensory, motoric, evaluative, and affective representations. That is largely seen as an alternative theory of concepts, not a complement, and in any event forms a much smaller tradition (Barsalou 1999; Prinz 2002; Pulvermüller 2013). Little work has attempted to bring both together in a unified framework.

I have tried to do just that. An obstacle to providing a unified account is the diversity of representational structures, informational models, and computational processes involved. My solution is that concepts act as an interface—a link between content-neutral concatenation and reasoning, on the one hand, and content-specific computations, models, and representational structures, on the other. The interface claim is hardly earth-shattering on its own, though. The hard work is to show how these diverse resources can be integrated.

According to my plug-and-play account, a substantial number of the deliberate inferences that we make are performed within special-purpose systems. A more traditional view would have it that information from special-purpose systems is re-represented in conceptual thought. For example, a thinker could visually represent that ducks have certain properties, on that basis conceptually represent that *ducks have webbed feet* and, using that as a premise and combining it with further background information, reason their way to the conclusion that *ducks swim on water*. In my picture, by contrast, some of the inferential processes actually take place between representations in special-purpose systems. The hippocampal cognitive map runs a simulation in order to work out how to get from A to B. We run a simulation in visuo-motor systems in order to work out whether a chair will fit in the car. The intermediate representations, domain-specific information, and content-specific assumptions all remain special-purpose. They don't need to be re-represented with concepts in order to be inferred with.

Some of these inferential steps take place in the cognitive playground, others outside it. Often the computations in special-purpose systems are opaque to the thinker. (They are not at the personal level.) While looking at a photograph, I can ask myself the names of the people in the picture. The names come to me as if by magic, unencumbered by any of the rich feature processing and complex statistical inferences that have taken place. Inferences in the cognitive spatial map are probably like that. When I ask myself how to get from here to the pub, the answer just pops into my mind (turn left out of the door, etc.). Replay simulations in the hippocampus (Liu et al. 2019, 2021) do not seem to show up in the cognitive playground—just the conclusion. By contrast, the inferences I make about the chair do take place within the cognitive playground.[1] I rotate the chair in my mind's eye to see if it will fit, and the various possibilities and intermediate states figure in the playground on the way to reaching the conclusion. When I see that the chair will fit, it is only the result of these inferences that is represented conceptually (using my concepts CHAIR, FIT, and CAR), not the intermediate steps.

Suppositional scenarios in the playground are an informational model of a worldly situation or range of possibilities. Running a simulation is a way of making inferences with this model. As we have seen, these inferences can be nonlocal: taking into account many features in parallel, as with moves in a semantic space; or taking into account relational or configural aspects of the model, like the overall configuration of locations in space or the overall spatial arrangement of physical objects in a scene. At the same time, the playground does also support step-by-step reasoning. It was broadly-logical inferences that validated the very

[1] At least for most. Even in those who lack conscious imagery (Keogh and Pearson 2018), similar functional features may be present (Pounder et al. 2022).

foundation of RTM—the idea that physical transitions between representational vehicles can be so-configured as to respect their semantic contents—so it is not surprising that these have been central to the philosophical understanding of the mind. Nevertheless, it has been a major omission not to enlarge our picture of deliberation so as to encompass processing over special-purpose informational models in the cognitive playground. That is what led Fodor to a picture of the mind where he felt forced to conclude that these kinds of inferences are deeply mysterious (Fodor 2000, pp. 23–39, 99).

Abductive, content-driven inference is completely integral to my framework. Deliberate thinking includes inferences between special-purpose representations forming informational models in the cognitive playground. The transitions between informational models are content-driven in the sense that many of the computational processes are content-specific (Chapter 3), thereby drawing on meaning in a substantial way (Chapter 7). The transitions are often non-local, depending on overall features of the current model.

Deliberate thinking overall is type 2, deploying directed attention and working memory to update and manipulate the current model. The transitions that happen in an inferential step are often, however, automatic. When the visuo-motor system is given an imagined chair and a simulated action as input, it produces a representation of the physically rotated chair automatically, based on the content-specific processing dispositions that have been built into the system by experience. The disposition to move from the conceptual representation, *that is a dog*, to, *it barks*, is triggered automatically, simply in virtue of tokening the premise. The overall process of running simulations to learn about the world and plan an action is type 2 and load-dependent because it depends on directing attention and using working memory to marshal and manipulate aspects of the current model. Each step triggers a new suite of automatic processing dispositions. The type 2 process consists of a series of type 1 steps. Conceptual representations serve to structure the informational model that is constructed and to orchestrate the unfolding process.

This account shows why deliberation is model-based in all four of the senses discussed in Chapter 1 (§1.2). Most obviously, (i) it involves inferring with an informational model of the structure—often the causal and physical structure—of the environment. However, as we saw, that is barely a substantial requirement. There are many different ways of representing aspects of the structure of the world, from the very simple to the very complex, with complexities of different kinds and no simple scale (Chapter 4). However, inferences over informational models in the cognitive playground depend on working memory. They are offline, in the sense of, (ii) going beyond reacting to the current stimulus. Further, the thinker can plan into the future so as to, (iii) make choices that are stimulus-independent. The interconnectedness of the cognitive playground (§1.1, §8.5) means that this whole process can be sensitive to the agent's current goals and reflective values. Thus, (iv) the decisions the thinker takes in this way can be immediately sensitive to a change in the value of an outcome, for example, to devaluation of one of the available rewards (unlike the values encoded in special-purpose systems).

In short, deliberate thinking is a process in which conceptual representations structure and orchestrate inferences that take place over rich informational models in the cognitive playground. That dissipates the mystery of how it is that we perform abductive, content-driven inference. It also makes the account true to what deliberation is really like: it involves reasoning with concepts, but goes much wider.

9.2 Concepts Look Both Ways

Concepts are a tool by which these diverse resources are integrated in thought. A concept acts as an interface between the general-purpose and the special-purpose. It is a plug-and-play device, one that plugs into different kinds of structure at each end. Compositionality has long been held to be a key feature of conceptual thought. That is somewhat puzzling if compositionality just means the ability to combine representational constituents into a complex whole whose meaning is systematically related to the meaning of the parts. Compositionality of that kind is exhibited by many special-purpose representations (§2.2), like the hippocampal map of space, including cases where constituents are unsaturated and semantically bound, with different constituents making semantic contributions of different kinds (e.g. object vs. property).

We can now see that the compositionality of the conceptual thoughts that figure in deliberation is a richer phenomenon. It actually has three aspects. The first is the ability to hold in working memory a conceptual label—a representation that is linked to a body of stored information (both conceptually-represented facts and special-purpose characterizations). The second aspect is compositionality as standardly understood: the ability to concatenate a small number of working memory labels using a semantically-significant compositional device. The third aspect is distinctive of compositionality over conceptual labels: it is general-purpose, unlike the compositional principles of structural representations. It is language-like, obeying something like Evans's generality constraint. It involves predication, or something even more general like Merge in natural language (§2.4). Labels can be tokened in thought and combined in a way that is unconstrained by the stored information to which they are connected.

The combination brings conceptual compositionality to life. It is not just a matter of combining and recombining neutral labels. Labels held in working memory serve to retrieve information from long-term memory, representations which are fed into an informational model in the cognitive playground. When we combine labels to form a novel thought, the content we have formulated conceptually is therefore something we can think about in a rich, multi-modal way. Conceptual labels drive the construction of a suppositional scenario, retrieving special-purpose information of different kinds, as well as conceptually-represented memories, and using all that to fill in aspects of the model and draw new conclusions from it (Chapter 5). Conceptual compositionality feels special because it comes with the capacity to formulate a suppositional scenario driven by the thought.

Inferences in special-purpose informational models often take place *on*line, taking current stimuli as input and producing categorisation or behaviour at output. Content-specific dispositions in visual processing are triggered by current visual input, categorising the objects the agent encounters, which then forms the basis of how they act. A conceptual label can use the same representations offline, holding them in working memory. Deliberate thinking can thereby make use of the information encoded, implicitly and explicitly, in special-purpose systems. However, the assumptions built into a special-purpose system are also limitations, constraining which possibilities can be represented. An advantage of a conceptual label is its content neutrality. The way it represents does not constrain what it represents. Because concepts act as an interface, deliberate thinking can make use of the rich contents represented in special-purpose informational models, doing so in a way that respects the general-purpose mode of combination by which conceptual labels are concatenated (the 'dog bites man / man bites dog' phenomenon: §5.7).

The fact that concepts represent in an arbitrary code gives conceptual thinking a distinctive flexibility. Special-purpose representations can represent many different kinds of contents, but there is a link between the content represented and the way it is represented. In a structural representation, the mode of combination has specific representational significance, as we saw (§2.3). For example, firing of place cells can represent relations between locations. The vehicles are combined by the relation of co-activation and that relation stands for spatial proximity. Even in a mere organized representational system, the systematic relation between vehicles and contents is a restriction (§2.3). Computational transitions are specified at the level of a determinable, applying in a systematic way to a range of vehicles. That is both a bonus—something that is exploited computationally—and a limitation—a restriction on what each vehicle can represent. Conceptual labels do not come with those kinds of restrictions on content.

A concept thus offers the agent a way to think about a relation abstractly. Consider representing the mother-of relation, for example. A chimpanzee might represent this relation perceptually, by imagining one individual standing in some physical relation to another (Penn et al. 2008). This way of representing that X is the mother of Y depends on representing X and Y in a particular way, using perceptual representations. Thinking about *mother-of* abstractly is a matter of being able to represent the relation without having to represent anything specific about the relata between which it obtains. When a relation is represented with a concept, the concept does so abstractly, and exhibits role-filler independence. Indeed, a conceptual label can represent a relation on its own, without representing any relata between which it obtains. It is a means by which we can think about the relation as such. That in turn offers the possibility of representing higher-order relations—relations that obtain between those relations (ditto for monadic properties).

The content-neutrality of working memory labels also underpins another, closely-related phenomenon: a general-purpose capacity for analogical inference. Reasoning by analogy is a matter of drawing a parallel between a relation in one system and a relation in another. We can map the spatial ordering of points on a line to the temporal ordering of events in time. We use physical relations on the line to draw conclusions about temporal relations between events (e.g. e_1 is to the left of e_6 on the line, so it happened first). Mapping the relation in a structural representation to a concept offers a completely general-purpose way of reasoning by analogy. It is not just a matter of lining up two representational systems that have some structural correspondence. Conceptual representations can be combined in a way that is neutral as to what the relational concept picks out. So concepts allow us to analogise anything to anything else (to a first approximation).

The same contents that are represented in special-purpose systems can also be represented conceptually. Linguists teach us concepts of phonemes, but people don't need to have these concepts in order to hear different phonemes categorically. We all have concepts of many contents that we also perceive, for example properties of shapes and of spatial relations (IN FRONT, BEHIND). Is there any deep difference between the conceptual representation and the not-conceptually compositional representation of the same content? One tactic is to deny that there are such cases, restricting special-purpose systems to modality-specific properties, like visual shape and visual occlusion. That is unlikely to be decisive, since many special-purpose systems deal in properties that seem to transcend any particular perceptual modality. Two examples are representing spatial locations and recognising object categories. Furthermore, even representations that are paradigmatically driven by one particular modality are probably better thought of as supra-modal (Calzavarini 2022).

There is a clear difference, however: conceptual labels admit of general-purpose composition and can be processed in content-general ways. Many perceptual representations are realized and processed in organized families. Similar contents are represented by similar vehicles and are processed in similar ways (§2.3). Well-studied examples include colour, spatial orientation, speed of motion, and numerosity. Concepts may in practice divide up possibility space somewhat differently (e.g. colour concepts are more coarse-grained than the perceptual representation of colours). But even where the contents are identical, conceptual labels in working memory are representing in a different way than the organized representations in special-purpose systems.[2]

9.3 Tokening a Concept

What is it, then, to token a concept? In Chapter 5 I argued that a concept, within a given thinker, should be type-identified using vehicle properties (§5.7): instances of the same working memory label used within an episode of thinking and, across episodes of thinking, working memory labels that are connected to the same body of stored information. I distinguished between a concept understood as a representation a person thinks with occurrently and a concept as body of information stored together in long-term memory. I have been using 'concept' in the former sense, but the occurrent thought does include some of the stored information, representations that have been retrieved from memory on a particular occasion. Should these representations be counted as part of what tokens the concept on that occasion?

This offers us two ways to talk about tokening a concept. First, we could consider it to be just a matter of tokening the working memory label—everything else would then count as tokening connected characterizations. Alternatively, we could take the label-plus-characterizations activated on an occasion to be a tokening of the concept on that occasion. I prefer the former, since it is less confusing given the way I type-identify concepts, but there is no deep issue here. Either treatment is fine, provided we are clear about what is meant in context. The latter treatment fits better with the idea that a concept is something stored in long-term memory—a collection of interconnected characterizations. To token a concept is then to token (some of the) information stored in long-term memory. Caution is needed, though: because only a small subset of the information in memory is tokened on each occasion, we cannot type-identify token concepts in terms of the information which is activated. So if we take the latter approach and treat the label-plus-retrieved-characterizations to be what tokens a concept, then we need to insist that different sets of retrieved characterizations will still count as tokenings of the same concept. Type-identification must still be based on sameness of

[2] This offers one way to distinguish between supra-modal representations and amodal representations. (And to answer an objection: if perceptual representations are not modality-specific, why do they not count as amodal?) Supra-modal representations, while driven by more than one sensory modality, come in organized families, and the way they are organized reflects aspects of the way information is collected by sensory systems. Amodal representations either have no semantically-significant similarities (other than same-symbol/different-symbol: Shea 2023c) or, where they display organization, that organization does not reflect sensory processing. Concepts (labels in working memory) may be amodal in the latter sense, if they fall into a semantic state space, or in the former sense; amodal 'semantic hubs' in anterior temporal lobe (Lambon Ralph et al. 2017) probably fall into semantic state spaces.

the label and store of connected information. Which associated characterizations are activated will change from occasion to occasion, and during an episode of thinking.

Either way, when we think with a concept, the inferences we make will depend on the information that is recalled from memory. Different characterizations will be retrieved on different occasions. So there is a sense in which a concept is realized in a context-dependent manner (Connell and Lynott 2014; Casasanto and Lupyan 2015). The psychological consequences of tokening the same concept will vary from occasion to occasion. This is actually compatible with there being a default store of information that is always retrieved quickly and automatically, as well as further context-dependent characterizations (Machery 2015). It is obviously also compatible with there being no such default (Smith and Samuelson 1997; Malt 2010). The latter fits well with results suggesting that language comprehension involves activating a contextually communicated meaning without first going via a common literal meaning (Giora 2002). Either way, the representations which are eventually retrieved, and then guide inference, vary substantially from context to context.

There is a related debate between pluralism and hybridism about concepts. Hybridism argues that a concept is a hybrid of information stored in a variety of forms, in particular prototypes, exemplars, and mini-theories (Vicente and Martínez Manrique 2016). Pluralism argues that each of these ways of representing information about a category is a different concept (Weiskopf 2009b). A third position endorses the pluralism but goes further and argues that we should therefore eliminate the notion of *concept* from our theorising (Machery 2009, 2015). My approach sits naturally with hybridism. However, just as it is compatible with there being a core of information which is activated by default when tokening a given concept, it is also compatible with there being more than one such for a given category (pluralism). The question is an empirical one. If pluralism is right, then the different bodies of knowledge are not connected. So my vehicle-based way of individuating a concept would count them as different concept types (which in fact refer to the same referent). Where the different bodies of knowledge are connected (Malt 2010), so that they are effectively stored together in memory, there will be a single concept with a hybrid character.

Whichever way the debate between hybridism and pluralism turns out empirically, the theoretical point remains. It is important not to elide two notions of concept: as a representation in occurrent thought, and as a collection of information stored together in memory.

9.4 Doing in Thought

The book has been about the role of concepts in deliberation. The aim has been to give an account of concepts that makes sense of the way we use them in deliberate thinking. That was the motivation for a framework which unifies their role in reasoning with their role in accessing and marshalling special-purpose representations. Throughout I have been treating this as an exercise of agency. A mental action is something the thinker *does*, as a whole agent or person, like deciding or calculating, rather than something that happens to them, like falling asleep or feeling an injury. I have carefully avoided positing a homunculus—an unexplained psychological component that does the crucial thinking and deciding. But I have also avoided giving a positive account of mental agency. And I won't do so now. However, it is worth noting that my picture does contain many of the elements that will be needed to build an account of why thinking is an exercise of mental agency. That is what I briefly lay out in this section.

To avoid positing a homunculus, we can instead explain what the thinker does, as an agent, in terms of the operation of various psychological capacities. Many are the capacities that experimental psychologists study under the rubric of 'executive functions'. The strategy is to show how the phenomena that are characteristic of mental agency can emerge from the interaction of various capacities, each explicable in its own right. These are the capacities I have appealed to already in characterising conscious deliberation.

Deliberation with concepts has the properties that are taken to be characteristic of type 2 cognitive processes. Particularly pertinent is that it is effortful. Deliberation can feel like hard work. It draws on working memory and is subject to interference by concurrent cognitive load. It calls for attention to be directed so as to disengage from current stimuli, including automatic behavioural responses, so as to think things through. Deliberate thinking uses attention to enter representations into working memory, to actively erase them, and to screen out distracting information. As we have seen, variation in the capacity for attention to be directed in this way accounts for significant variation in standardised measures of fluid intelligence, which in turn predict educational outcomes and accurate performance in many different tasks (§1.2). It is at the heart of personal-level thought processes. The way attention operates depends on the thinker's current goals and values, and their occurrent beliefs. That is, it depends on what is being represented in the cognitive playground. Attention is a capacity driven by, and effective on, the contents of the playground. It is one of the capacities that operate within cognition. There is no need for a homunculus—something external determining how thoughts unfold.

Deliberate thinking marshals representations in the cognitive playground. This means that items of information are not processed separately, as in an encapsulated module, but in an integrated way. Inferential transitions can depend on overall features of the current informational model in the playground. Representations in the playground can also be integrated with the agent's occurrently-represented goals and values. So decisions taken in this way are able to reflect those values; the content-specific dispositions of special-purpose systems need not. This is why inferences drawn from informational models in the playground are also 'model-based' in the sense of being immediately sensitive to reward devaluation (i.e. 'goal-directed': Dickinson and Balleine 1994). Integration means that affective and motivational factors are part of the account of thinking, not just cognitive factors, and makes for a seamless link to the social.

Deliberate thinking is also an arena where the thinker appreciates what is going on. The representations are plausibly conscious. That is not central to my account, but I do rely on their having the functional features of being in a workspace. They are connected or unified and operate at the level of the whole person. Further, many of the transitions draw on the meaning of the representations in a substantial way (Chapter 7). And deliberate thinking is subject to metacognitive monitoring and control—not by a homunculus (the inner eye watching mental life), but in the sense that processes occurring in deliberation generate epistemic feelings which have an effect on how subsequent processing unfolds (Chapter 8). The thinker gets a signal, for instance, that a judgement they have made is likely to be inaccurate. As a result, they stop and think about it some more. They have a sense of the comparative reliability of different concepts as tools for thinking and select them in part on that basis. Inferences generate signals which give the thinker an indication of whether a particular pattern of inference is reliable. By reflecting on whether they endorse or reject an inference the thinker can affect that sense of reliability and thereby do something to change their automatic inferential dispositions into the future. They can come to align their thinking better with norms that they imbibe from their culture or choose for themselves. And the coherence of the overall situational scenario currently active in the playground is also probably something that registers with the thinker. In short, deliberate thinking is subject to monitoring and control by various psychological processes which, since they are not themselves a matter of deliberate reflection, do not launch an explanatory regress.

The strategy is to account for the agentive aspects of deliberation in terms of the interaction between all these components. The phenomena of mental agency can emerge from the operation of a suite of more basic psychological capacities interacting in the right way ('emerge' in the non-spooky sense that the complex can have capacities that are more than the sum of the capacities of the parts). Endogenous control is not a matter of an *ex machina* intervention, but consists in the operation of a complex psychological capacity of this sort. This is not the place to argue for such an account of mental agency, but it is notable that the picture I have offered here already contains many of the elements that will be needed.

9.5 The Unreasonable Power of Human Cognition

Why is the human species so special? Partly, no doubt, it's a matter of their seeming special to us—because we're human. Many of nature's earth-shattering innovations are less salient from the human perspective: the invention of the eukaryotic cell, or of photosynthesis, providing the very foundations for plant and animal life. But attempting to step away from our skewed point of view, it does still seem that humans stand out in the natural world. We have notably complex technology, life-ways, and social arrangements. We have also had a regrettably outsized impact on the environment. How has this hairless primate managed to do so much?

There might have been a magic ingredient, from which all else flows. But that now seems unlikely. Searching for the unique feature that sets us apart from all other animals has become a fool's errand. But theorising about what makes humans special is more than just a parlour game. There are probably a small number of factors that are most important. Some good candidates include our powerful vision and manual dexterity, our socially interdependent way of life, and our accumulation of skills and technology through cultural inheritance; linguistic communication too, of course—whether or not language is the basis of the general-purpose compositionality of concepts. I would argue that this small list should also include our capacity for thinking with concepts, in the way set out in the book. Deliberate concept-driven thinking is a source of the special power of human cognition.

Having representations is a clever trick in itself. Representing aspects of the world en route to producing behavioural outcomes is a way for organisms to achieve important outcomes—those that have been stabilised by evolution and/or learning—more robustly (Shea 2018). The capacity for representation is found across the animal kingdom, if not even more widely. It takes sophisticated forms in animals in a number of different clades, from mammals through birds to insects and molluscs. In primates, for example, complex representations and elaborate computations mediate between perceptual input and behaviour.

More sophisticated still is the capacity for planning: for thinking through or simulating the consequences of various possible actions before deciding what to do. Relatively sophisticated forms of planning are already found in some of our primate relatives (Passingham 2021; Tomasello 2022) and perhaps also in more distant clades (corvids, cephalopods). An animal doing planning can take offline what it can represent online, often learnt from experience, and use offline representations to anticipate the consequences of the actions they could perform. This allows 'our hypotheses to die in our stead' (Dennett 2008, p. 88; quoting Popper 1972, p. 248).

Planning is especially flexible when it is done with conceptual representations. As we have seen, concepts can be freely recombined using a general-purpose mode of combination, allowing an open-ended range of possibilities and outcomes to be represented. The capacity for general-purpose representational combination—thinking in a language of thought, in one sense of that term—is especially highly developed in humans (Dehaene et al. 2022). For the purposes of planning with concepts, it is also crucial that humans have the capacity to deploy substantial working memory, and to direct attention so as to disengage from current input and think through chains of possibilities.

There is a question about the phylogenetic and ontogenetic source of these capacities. My account can remain relatively neutral about this. Learning is doubtless involved in the development of all of these abilities, but there are also likely to be aspects that are canalized in development, on the basis of adaptive information acquired through evolution by natural selection. I have also remained neutral about whether domain-general combination of concepts—something like a language of thought—depends on natural language; or the converse; or whether they are independent capacities. In any event, these capacities are doubtless elaborated by cultural evolution so that the particular thinking and reasoning skills an individual acquires depend very much on the culture they are brought up in. For example, the specific features of deductive reasoning look to be culturally explicable (Dutilh Novaes 2020). Deductive reasoning is, however, just one form of content-general or broadly-logical inference, a category which extends more widely. This broader capacity might also be a culturally-inherited cognitive gadget (Heyes 2018), or it may instead be a more universal and canalized ability that goes along with having a type of representation—concepts—that can enter into general-purpose compositional structures. Many of the metacognitive tools that we rely on to aid our thinking (Chapter 8) could also be culturally-inherited cognitive gadgets (Heyes et al. 2020).

The ability to engage in broadly-logical reasoning over representations that display a general-purpose mode of combination is nevertheless not unique to humans. It is a capacity we share with computers—with the thinking machines we have created. They too can use symbols with a combinatorial syntax, compute with variables, and engage in long chains of reasoning. Computers can already perform a branching tree search many more steps into the future than humans can, allowing them to beat us at strategic games like chess and Go. Even so, not every task is best tackled that way. To decide using step-by-step forward planning takes time. For the types of task that an agent has to perform repeatedly, with the benefit of rich experience of what works (acquired through learning or canalized by evolution), it is more efficient to rely on if-then dispositions. These can involve complex information processing and multiple computational steps, as we have seen, but they do not require reasoning about future possibilities in a general-purpose way. The if-then solution is learning heavy but computation light at decision time. Model-based planning is much more flexible and open-ended, dealing with novel situations and allowing one-shot learning, but it is more computationally demanding at decision time.

Having access to both approaches, at least in some form, is also not unique to humans. Other species pursue the strategy of deploying each in its appropriate domain: act fast when the state is familiar or an immediate response is crucial; use an informational model to plan ahead when you can. What is special in humans is our ability to plan in a general-purpose way, using reasoning over concepts, and to deploy our special-purpose informational models flexibly in our planning, via the concepts to which they are connected. We have a domain-general way of making use of special-purpose resources in our planning, combining the results into a coherent situation model in the cognitive playground and using that to work out the consequences of our actions.

The engine of that ability is conceptual thought. Concepts interface between a domain-general capacity for combination and reasoning and the many special-purpose informational models to which they are connected. This is not just a matter of horses for courses—using the different approaches for different kinds of tasks. Concepts act as an interface that allows us to integrate the two approaches and rely on them together. Given these plug-and-play devices, with greater scope for general-purpose recombination, and the well-developed working memory capacity and executive functions to make use of them, human theoretical and practical inference can rely on both powerful content-general reasoning and the learnt experience of special-purpose informational models at the same time. It is not just the capacity for reasoning but the ability at the same time to go beyond reasoning that generates the special power of human cognition.

Special-purpose systems bring to the job of planning the information they have acquired through a wealth of experience. Using this offline gives us access to non-local processes, weighing many different considerations at once, or making inferences by moving through a high dimensional semantic space. The general-purpose recombinability of concepts allows us to formulate entirely new possibilities. Content-general, broadly-logical reasoning allows us to perform useful inferences on novel representations, even when the possibility represented falls far outside the range of experience for which we have acquired content-specific dispositions. That is valuable when we encounter a novel situation. We can represent it, remember it, and plan with it. It is also helpful when planning in a familiar situation, because it allows us to formulate new plans of action that proceed through novel world states. The connection of concepts to special-purpose informational models allows us to attempt to make sense of a novel conceptual representation (e.g. pursuing a light beam at the speed of light) by constructing a suppositional scenario in the cognitive playground. Humans' hybrid system gives us a practical way of taking good decisions. We can use a variety of perspectives to retrieve from memory considerations that may be important, and that are contextually relevant in different ways. We can use reasoning to work out the consequences of these factors: the likely outcomes of various actions we could take, how other people would react, and how the outcomes strike us affectively. Finally, we can rely on systems that perform multiple-constraint satisfaction to weigh these factors in parallel to produce an overall assessment or feeling about what to do, in a way that step-by-step reasoning cannot accomplish on its own.

The capacity for content-general reasoning from stored memories brings in train the frame problem—the problem of retrieving a tractable set of relevant representations to reason with. Because concepts interface with special-purpose systems, conceptual thought can recycle the context-specific if-then dispositions of special-purpose informational models as a method to perform relevance-based retrieval. Thinking with concepts in this hybrid way thus throws up the frame problem but also contains a partial solution. It is this combination of features that makes human cognition so powerful. Or so I suggest.

So what exactly is the special ingredient? There isn't one—it's the way they're combined. The picture is slightly complex, but let me oversimplify briefly for emphasis. What sets humans apart is not performing complex computations with organized representations and structured representations. Many other animals have special-purpose systems which do that. And in recent years we have seen that deep neural networks—given large computational resources and enormous amounts of training on huge databases of information—can use this computational principle to solve what were always thought to be really difficult tasks. Trained if-then dispositions are in one sense a simple solution, but they can produce behavioural performance that exceeds the skill of humans in many domains. Most deep neural networks do not yet engage in planning, but that doesn't set humans apart either. As we have seen, several non-human animals can do prospection or model-based forward planning before deciding what to do. The capacity for general-purpose, broadly-logical reasoning may set us apart from other animals. Non-human animals either lack it entirely or have it in a less sophisticated form. However, this is not unique to us either. Computers have it in spades—it is the whole basis of classical computation. But they haven't yet managed to match human flexible planning and fluid intelligence in a generally applicable way.

What sets us apart from all of these, both other animals, and computing machines (at least for now), is the ability to link up highly-flexible general-purpose recombination and content-general reasoning with special-purpose informational models. Engaging in thinking in this hybrid way is plausibly a human speciality. Concepts are at the heart of it. Conceptual thinking is the engine of distinctively human cognition. Concepts are keys that unlock the mind's resources.

Chapter Summary

9.1 Deliberating with Concepts: The Picture

Research on concepts has focused on categorisation; work on inference has focused on reasoning, taking place step-by-step between conceptual representations; the role of special-purpose resources has been to support an alternative theory of concepts, which has not become central. The book aims to unify these, with concepts at the centre, linking content-neutral concatenation and reasoning with content-specific computations, models, and representational structures; the task is to show how these diverse resources can be integrated.

(p. 212)[3] In my picture, information represented in special-purpose systems does not need to be re-represented with concepts in order to be inferred with in deliberation. Inferential steps take place outside (hippocampal spatial map) or inside (mental rotation) the cognitive playground, but with only the conclusion of the inference, not the intermediate steps, represented using concepts. The cognitive playground supports both step-by-step reasoning, and non-local inferences; the latter may take into account relational or configural aspects of the model, or may process many features in parallel; it is a major omission not to include the latter in our picture of deliberation. (p. 213) Abductive, content-driven inference is integral to my picture.

Deliberation overall is type 2, deploying directed attention and relying on working memory, and is made up of a series of automatic, type 1 steps. This account shows why deliberation is model-based (§1.2): (i) inferring with an informational model of the world, (ii) calculating over those representations in working memory, (iii) making choices that are stimulus-independent, (iv) exhibiting immediate sensitivity to a change in the value of outcomes. (p. 214) In short, deliberation involves reasoning with concepts but goes much wider, into special-purpose systems and rich informational models in the cognitive playground; that dissipates the mystery of how we perform abductive, content-driven inferences.

9.2 Concepts Look Both Ways

Concepts integrate these diverse resources, interfacing between the special-purpose and the general-purpose. Concept compositionality involves: an ability to hold in working memory labels that are connected to bodies of stored information, to combine them compositionally, and to do so using general-purpose compositional principles, unconstrained by the stored information to which each is connected. This allows for the formulation of novel conceptual representations; what brings them to life is the ability of working memory labels to drive the construction of suppositional scenarios and inferences over special-purpose informational models. (p. 215) This allows deliberate thinking to rely on information encoded explicitly or implicitly in special-purpose informational models, while also transcending their assumptions and limitations.

[3] Each sentence of the summary corresponds to one paragraph. Page numbers indicate where the paragraphs begin.

In special-purpose informational models using structural representations or organized representations, the structures that make them computationally useful also impose limitations on what they can represent; concepts are not so-restricted. Concepts allow us to think about a relation abstractly, without representing anything about the relata between which it obtains. (p. 216) This allows us to formulate analogies between any two relations.

The same contents that are represented in special-purpose systems can also be represented by concepts. Concepts are, however, content neutral vehicles, thus representing in a different way from special-purpose representations that form organized families.

9.3 Tokening a Concept

(p. 217) Taking a concept to be an occurrent representation, rather than a stored body of information, what it is to token a concept is to token a working memory label, and thereby to retrieve and activate a small subset of the information to which the label is connected in long-term memory. We can think of the token concept as being just the working memory label, or the label plus retrieved information; if the latter, type-identification proceeds at the level of the label, not the information retrieved. (p. 218) There is thus a sense in which a concept is realized in a context-dependent manner—the representations that are retrieved, and then guide inference, vary substantially from occasion to occasion. My approach sits naturally with concept hybridism; it is compatible with pluralism, if it turns out empirically that there are different bodies of stored information about a category that are not connected in memory. Either way, it is important not to elide the two notions of concept: as a representation in occurrent thought, and as a collection of information stored together in memory.

9.4 Doing in Thought

Using concepts in deliberation is something the thinker does; I have not offered a positive account of mental agency, but my picture does contain many of the elements that will be needed for a non-homuncular account. (p. 219) The strategy is to show how the interaction of various psychological capacities, each explicable in its own right, gives rise to the phenomena that are characteristic of mental agency.

Deliberate thinking uses directed attention, shaped by the thinker's goals and values, to disengage from current stimuli, and to maintain and manipulate representations in working memory. Inferential processes in the cognitive playground can reflect the thinker's current goals and values. (p. 220) Deliberate thinking takes place in an arena where the thinker appreciates what is going on, not just in the sense that it operates in part in consciousness (i.e. in an interconnected global workspace), but also in the sense that it is subject to metacognitive monitoring and control, allowing the agent better to align their thinking with norms, individual or social.

Thus, my picture already contains many of the elements that will be needed to formulate an account of endogenous control as consisting in the operation of a complex psychological capacity.

9.5 The Unreasonable Power of Human Cognition

Why is the human species so special? (p. 221) There is no magic ingredient, but there are probably only a small number of factors that are most important; this list should include our capacity for thinking with concepts. Having representations is a clever trick in itself, widespread in the animal kingdom. More sophisticated still is the capacity for planning, which allows 'our hypotheses to die in our stead'. Planning is especially flexible when it is done with conceptual representations, with their power of general-purpose recombination, when used with directed attention and substantial working memory capacity. (p. 222) I remain neutral on the phylogenetic and ontogenetic sources of these capacities; there is doubtless some mix of developmentally canalized outcomes based on gene-based evolution, culturally-inherited adaptations, and individual learning.

Computers can also engage in logical reasoning and step-by-step forward planning, and they can laboriously learn if-then solutions to repeatedly-presented problems. Some other animals have access to both approaches, deploying each in its appropriate domain: act fast when the state is familiar or an immediate response is crucial, use an informational model to plan ahead when you can; human cognition is not limited to switching between the two approaches, but can employ both together in a hybrid, since we have a domain-general way of making use of special-purpose resources in our planning. (p. 223) The engine of that ability is conceptual thought, which gives us both the capacity for reasoning and at the same time the ability to go beyond reasoning. We can represent an entirely novel possibility conceptually and then attempt to make sense of it by constructing a suppositional scenario from special-purpose informational models; the hybrid architecture gives us a practical way to take good decisions. (p. 224) Content-general reasoning faces the frame problem; thinking with concepts in this hybrid way allows us to recycle the assumptions and constraints inherent in special-purpose informational models as a partial solution.

What's special to humans is not any individual ingredient—each is displayed by other animals and/or computing machines—but the way they're combined. What sets us apart is the ability of concepts both to enable general-purpose recombinability and content-general reasoning, and to link up with special-purpose informational models: concepts are the keys that unlock the mind's resources and the engine of distinctively human cognition.

Concepts at the Interface. Nicholas Shea, Oxford University Press. © Nicholas Shea 2024.
DOI: 10.1093/9780191997167.003.0009

Acknowledgements

I have been thinking about concepts and meaning ever since I first started reading philosophy as a student. I may not have been able to reach any firm conclusions about how concepts get their content, but I have learnt a lot about concepts, and about the mind, on the way. An early inspiration was an MA seminar taught by Sarah Patterson and Barry Smith at Birkbeck. I also learnt an awful lot from members of the London Concepts Group—I am very grateful to psychologists James Hampton and Nick Braisby for welcoming this philosopher to their meetings. Many of my early ideas were tested and shaped by David Papineau and Gabriel Segal during my PhD at King's College London. Most are still to see the light of day, but some of the ideas in this book are distant descendants of those discussions. David Papineau has been a regular sounding board and generous colleague ever since. I remain immensely grateful for his support and inspiration.

The Institute of Philosophy has provided a fertile environment for working on these issues. I am grateful to Barry Smith and Ophelia Deroy for facilitating the collaboration with Oxford University and City University for my European Research Council project, and to Barry for his continuing support. The IP's weekly 'lab' meeting has been a constant source of feedback, philosophical and scientific, and of good advice. I have tried out almost every part of the book in some form there. I would also like to thank my London colleagues Adrian Alsmith, Jonathan Birch, Ali Boyle, Alex Grzankowski, and James Stazicker, from whom I have learnt a lot (and to apologise that our discussions have so often seemed to circle back to issues that were preoccupying me for the book).

I am very grateful for support for this project from the European Research Council (under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 681422 MetCogCon). James Hampton at City University was involved in the project from the outset. My thanks are due to him, and to Sapphira Thorne who carried out experiments for the project as a postdoctoral researcher at City. Joulia Smortchkova and Jake Quilty-Dunn worked on the project as postdoctoral researchers at Oxford. I am very grateful to them for the many constructive discussions we had, both collectively and individually; and to the Faculty of Philosophy in Oxford for hosting them and me during the project. Chris Frith, Celia Heyes, and Matthew Rushworth were each generous with their time in collaborating on various aspects of the project. It was a privilege to be able to draw on the expertise of such respected scientists in their respective fields. I also greatly enjoyed working with Kentaro Miyamoto in the later stages of the project.

I am very grateful to the following colleagues for discussion, references, and comments on written drafts: Marc Artiga, Fabrizio Calzavarini, Rosa Cao, Peter Carruthers, Robyn Carston, Dimitri Coelho Mollo, Peter Godfrey-Smith, Steven Frankland, Chris Frith, Colin Klein, Eric Mandelbaum, Manolo Martinez, Raphaël Millière, Ruth Millikan, Ross Pain, Seongmin Park, Christopher Peacocke, Kim Plunkett, Brant Pridmore, Joëlle Proust, Daniel Rothschild, Josh Shepherd, Daniel Stoljar, Michael Strevens, and Agustin Vicente; and to audiences at: the Philosophy Departmental seminar and the *Neural Representation and Neural Computation* workshop at the ANU; the reading group organised by Ross Pain at ANU on the first five chapters of the manuscript; the St Andrews *Arché Mind & Language* seminar; the Arché Open Online Conference on *Applied Conceptual Engineering*; the Association for the Scientific Study of Consciousness annual conferences in Amsterdam and Krakow; the UT Austin *MindWork* seminar; the Barcelona *Information and Representation* workshop; the *Cognition Academy* of the Max Planck School of Cognition, Berlin; the Philosophy Departmental Seminar at the University of Birmingham; the *Concepts: Philosophical Perspectives* workshop at Ruhr-Universität Bochum; the Cambridge *Kinds of Intelligence* workshop: *Cognitive Science, Beyond the Human*; the Columbia online Deep Learning Discussion Group; a graduate class and the Cognitive Science Seminar at CUNY Graduate Center; the Cognition group at DeepMind in London; the European Society for Philosophy and Psychology annual conferences in Prague and Milan; the Philosophy Department Senior Seminar and the Philosophy, Psychology and Neuroscience seminar at the University of Glasgow; the Hamburg workshop *Thought in Humans and Machines: Multidisciplinary Perspectives*; the Philosophy Colloquium at the University of Maryland; the *Mind & Language* workshop on *Metacognition* in London; the *Concepts, Pointers and Occasion-Specific Thoughts* conference at the Institute of Philosophy, University of London; the 'MAD' online discussion group and an in-person seminar at Monash University; the CVBE and Crowd Cognition Lab meeting at LMU Munich (twice); the Mind discussion group and an invited seminar at NYU; the *Philosophy of Deep Learning* conference at NYU; the *Colloquium in Honour of Ruth Millikan* at Oberlin; the Oxford Mind work-in-progress group (twice); the Philosophy of Biology at Dolphin Beach workshop; the Philosophy of Science Association biennial conference in Pittsburgh; the Philosophy Departmental Seminar at the University of Sydney; the Traincrease Rome Conference on *New/Fresh Perspectives on Abstract Concepts*; the *Varieties of Aboutness* workshop in Turin; the conference on *The Representational Penumbra* in Valencia; and the workshop on *Concepts, Meaning and Polysemy* at UPV/EHU Vitoria-Gastiez. I am especially grateful to Liz Camp for being so generous with her time and so thoughtful with her feedback. Her advice about the project was particularly formative.

Some of the ideas in the book received an earlier treatment in 'Concepts as Plug & Play Devices', *Philosophical Transactions of the Royal Society B*, 378:

20210353; 'Moving Beyond Content-Specific Computation in Artificial Neural Networks', *Mind & Language*, 38: 156–77; 'Millikan's Consistency Testers and the Cultural Evolution of Concepts', *Evolutionary Linguistic Theory*, 5: 81–103; 'Organized Representations Forming a Computationally Useful Processing Structure', *Synthese*, 202: 1–20; and 'Metacognition of Inferential Transitions', forthcoming in the *Journal of Philosophy*.

My thanks are due to the three readers for Oxford University Press who offered valuable feedback on the book proposal. One in particular gave thoughtful advice which shaped how I came to conceive of the whole project. I'm also immensely grateful to two readers for the Press who offered detailed comments on the full manuscript. Peter Momtchiloff has been an outstanding editor at OUP, gently and patiently supporting the entire project. I was pleased to be able to send the final version of the book into production in Peter's last week at the Press. He will be much missed. I am also grateful to Jamie Mortimer for shepherding the book through its final stages, Jo North for copyediting, Vasuki Ravichandran for overseeing production, and Iwan Williams for his sterling work on the index.

Philosophy is an inspiring day job, but a quiet one. Almost the whole time I have been thinking about concepts and meaning I have also been tied up with the busy whirl of parenthood. My children Adam and Clara are at the heart of what has made my wider life meaningful—as well as fun. This book is dedicated, with love, to them.

Figure Credits

The following are reproduced with thanks under the terms of the Creative Commons Attribution Licence https://creativecommons.org/licenses/by/4.0/

Figure 2.2 from Khajeh-Alijani et al. (2015).

Figure 3.1 from Güçlü and van Gerven (2015).

Figure 4.1 from Whittington et al. (2020).

Figure 4.2 from Park et al. (2021) is reproduced with permission of Springer Nature. This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

Figure 4.3 from Lovett & Forbus (2017) is reproduced with permission of the American Psychological Association. This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

Figure 4.4 from Krawczyk (2012) is reproduced with permission of Elsevier. This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

Figure 6.1 from Grand et al. (2022) is reproduced with permission of Springer Nature. This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

References

Ackerman, Rakefat, and Valerie A. Thompson. 2017. 'Meta-reasoning: Monitoring and control of thinking and reasoning', *Trends in Cognitive Sciences*, 21: 607–17.

Aguilera, Mariela. 2021. 'Heterogeneous inferences with maps', *Synthese*, 199: 3805–24.

Aho, Kaarina, Brett D. Roads, and Bradley C. Love. 2023. 'Signatures of cross-modal alignment in children's early concepts', *Proceedings of the National Academy of Sciences of the United States of America*, 120: e2309688120.

Akam, Thomas, Ines Rodrigues-Vaz, Ivo Marcelo, Xiangyu Zhang, Michael Pereira, Rodrigo Freire Oliveira, Peter Dayan, and Rui M. Cost. 2021. 'The anterior cingulate cortex predicts future states to mediate model-based action selection', *Neuron*, 109: 149–63, e7.

Allott, Nicholas, and Mark Textor. 2012. 'Lexical pragmatic adjustment and the nature of ad hoc concepts', *International Review of Pragmatics*, 4: 185–208.

Amalric, Marie, Liping Wang, Pierre Pica, Santiago Figueira, Mariano Sigman, and Stanislas Dehaene. 2017. 'The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers', *PLOS Computational Biology*, 13: e1005273.

Andrews, Glenda. 2010. 'Belief-based and analytic processing in transitive inference depends on premise integration difficulty', *Memory & Cognition*, 38: 928–40.

Antony, Louise, and Georges Rey. 2016. 'Philosophy and psychology', in Herman Cappelen, Tamar Szabó Gendler, and John Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology* (New York: Oxford University Press).

Apperly, Ian A., and Stephen A. Butterfill. 2009. 'Do humans have two systems to track beliefs and belief-like states', *Psychological Review*, 116: 953–70.

Aronowitz, Sara, and Tania Lombrozo. 2020. 'Learning through simulation', *Philosophers' Imprint*, 20: 1–18.

Ashby, F. Gregory, and W. Todd Maddox. 2011. 'Human category learning 2.0', *Annals of the New York Academy of Sciences*, 1224: 147–61.

Ashby, F. Gregory, and Vivian V. Valentin. 2017. 'Multiple systems of perceptual category learning: Theory and cognitive tests', in Henri Cohen and Claire Lefebvre (eds.), *Handbook of Categorization in Cognitive Science* (2nd edition) (Amsterdam: Elsevier).

Azzopardi, Leif. 2021. 'Cognitive biases in search: A review and reflection of cognitive biases in information retrieval', *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (New York: ACM), 27–37.

Baddeley, Alan. 2012. 'Working memory: Theories, models, and controversies', *Annual Review of Psychology*, 63: 1–29.

Baddeley, Alan, Graham Hitch, and Richard Allen. 2021. 'A multicomponent model of working memory', in Robert H. Logie, Valerie Camos, and Nelson Cowan (eds.), *Working Memory: The State of the Science* (New York: Oxford University Press).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. 'Neural machine translation by jointly learning to align and translate', *arXiv:1409.0473*.

Bahrami, Bahador, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith. 2010. 'Optimally interacting minds', *Science*, 329: 1081–5.

Ball, Linden J., Valerie A. Thompson, and Edward J. N. Stupple. 2018. 'Conflict and dual process theory: The case of belief bias', in W. De Neys (ed.), *Dual Process Theory 2.0* (Abingdon: Routledge).

Balleine, Bernard W., and Anthony Dickinson. 1998. 'Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates', *Neuropharmacology*, 37: 407–19.

Banks, Martin S., David M. Hoffman, Joohwan Kim, and Gordon Wetzstein. 2016. '3D displays', *Annual Review of Vision Science*, 2: 397–435.

Baram, Alon Boaz, Timothy Howard Muller, Hamed Nili, Mona Maria Garvert, and Timothy Edward John Behrens. 2021. 'Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems', *Neuron*, 109: 713–23, e7.

Barone, Pamela, Guido Corradi, and Antoni Gomila. 2019. 'Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis', *Infant Behavior and Development*, 57: 101350.

Barron, Helen C., Raymond J. Dolan, and Timothy E. J. Behrens. 2013. 'Online evaluation of novel choices by simultaneous representation of multiple memories', *Nature Neuroscience*, 16: 1492–8.

Barron, Helen C., Hayley M. Reeve, Renée S. Koolschijn, Pavel V. Perestenko, Anna Shpektor, Hamed Nili, Roman Rothaermel, Natalia Campo-Urriza, Jill X. O'Reilly, and David M. Bannerman. 2020. 'Neuronal computation underlying inferential reasoning in humans and mice', *Cell*, 183: 228–43, e21.

Barsalou, Lawrence W. 1983. 'Ad hoc categories', *Memory & Cognition*, 11: 211–27.

Barsalou, Lawrence W. 1999. 'Perceptual symbol systems', *Behavioral and Brain Sciences*, 22: 577–660.

Barsalou, Lawrence W. 2003. 'Situated simulation in the human conceptual system', *Language and Cognitive Processes*, 18, Special Issue on Semantic and Conceptual Representation: 513–62.

Barsalou, Lawrence W. 2008. 'Grounded cognition', *Annual Review of Psychology*, 59: 617–45.

Barsalou, Lawrence W. 2009. 'Simulation, situated conceptualization, and prediction', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364: 1281–9.

Barsalou, Lawrence W. 2016. 'Situated conceptualization: Theory and applications', in Yann Coello and Martin Fischer (eds.), *Perceptual and Emotional Embodiment* (Abingdon: Routledge).

Barth, Hilary, Kristen La Mont, Jennifer Lipton, Stanislas Dehaene, Nancy Kanwisher, and Elizabeth Spelke. 2006. 'Non-symbolic arithmetic in adults and young children', *Cognition*, 98: 199–222.

Barwise, Jon. 1986. 'Information and circumstance', *Notre Dame Journal of Formal Logic*, 27: 324–38.

Barwise, Jon, and John Etchemendy. 1996. 'Heterogeneous logic', in Gerard Allwein and Jon Barwise (eds.), *Logical Reasoning with Diagrams* (New York: Oxford University Press).

Bascandziev, Igor, and Susan Carey. 2022. 'Young children learn equally from real and thought experiments', *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44: 142–8.

Bayne, Tim. 2010. *The Unity of Consciousness* (New York: Oxford University Press).

Bayne, Tim, and Michelle Montague. 2011. 'Cognitive phenomenology: An introduction', in Tim Bayne and Michelle Montague (eds.), *Cognitive Phenomenology* (New York: Oxford University Press).

Beck, Jacob. 2018. 'Analog mental representation', *Wiley Interdisciplinary Reviews: Cognitive Science*, 10: e1479.

Beck, Jeffrey M., Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K. Churchland, Jamie Roitman, Michael N. Shadlen, Peter E. Latham, and Alexandre Pouget. 2008. 'Probabilistic population codes for Bayesian decision making', *Neuron*, 60: 1142–52.

Bendaña, Joseph, and Eric Mandelbaum. 2021. 'The fragmentation of belief', in Cristina Borgoni, Dirk Kindermann, and Andrea Onofri (eds.), *The Fragmented Mind* (New York: Oxford University Press).

Bergen, Benjamin K. 2012. *Louder Than Words* (New York: Basic Books).

Block, Ned. 1990. 'Can the mind change the world?', in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam* (Cambridge: Cambridge University Press).

Block, Ned. 1993. 'Holism, hyper-analyticity and hyper-compositionality', *Philosophical Issues*, 3: 37–72.

Block, Ned. 2011. 'Perceptual consciousness overflows cognitive access', *Trends in Cognitive Sciences*, 15: 567–75.

Blouw, Peter, Eugene Solodkin, Paul Thagard, and Chris Eliasmith. 2016. 'Concepts as semantic pointers: A framework and computational model', *Cognitive Science*, 40: 1128–62.

Boghossian, Paul. 2014. 'What is inference?', *Philosophical Studies*, 169: 1–18.

Bonnay, Denis. 2008. 'Logicality and invariance', *Bulletin of Symbolic Logic*, 14: 29–68.

Bottou, Léon. 2014. 'From machine learning to machine reasoning', *Machine learning*, 94: 133–49.

Botvinick, Matthew, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. 2019. 'Reinforcement learning, fast and slow', *Trends in Cognitive Sciences*, 23: 408–22.

Bouchacourt, Flora, and Timothy J. Buschman. 2019. 'A flexible model of working memory', *Neuron*, 103: 147–60, e8.

Brewer, Bill. 1995. 'Mental causation: Compulsion by reason', *Proceedings of the Aristotelian Society*, *Supplementary Volume*, 69: 237–53.

Brody, Gabor, and Roman Feiman. 2023. 'Polysemy does not exist, at least not in the relevant sense', *Mind & Language*. https://doi.org/10.1111/mila.12474.

Brown, Jessica. 1995. 'The incompatibility of anti-individualism and privileged access', *Analysis*, 55: 149–56.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. 'Sparks of artificial general intelligence: Early experiments with GPT-4', *arXiv:2303.12712v5*.

Burge, Tyler. 2010. *Origins of Objectivity* (Oxford: Oxford University Press).

Burge, Tyler. 2018. 'Iconic representation: Maps, pictures, and perception', in S. Wuppuluri and F. Doria (eds.), *Map and Territory: Exploring the Foundations of Science, Thought and Reality* (Heidelberg: Springer).

Butlin, Patrick. 2021. 'Cognitive models are distinguished by content, not format', *Philosophy of Science*, 88: 83–102.

Butlin, Patrick. 2022. 'Machine learning, functions and goals', *Croatian Journal of Philosophy*, 22: 351–70.

Calzavarini, Fabrizio. 2022. 'The conceptual format debate and the challenge from (global) supramodality', *British Journal for the Philosophy of Science*. Forthcoming.

Camp, Elisabeth. 2006. 'Metaphor in the mind: The cognition of metaphor', *Philosophy Compass*, 1/2: 154–70.

Camp, Elisabeth. 2007. 'Thinking with maps', *Philosophical Perspectives*, 21: 145–82.

Camp, Elisabeth. 2015. 'Logical concepts and associative characterizations', in Eric Margolis and Stephen Laurence (eds.), *Conceptual Mind: New Directions in the Study of Concepts* (Cambridge, MA: MIT Press).

Camp, Elisabeth. 2018. 'Why maps are not propositional', in Alex Grzankowski and Michelle Montague (eds.), *Non-Propositional Intentionality* (New York: Oxford University Press).

Camp, Elisabeth. 2019. 'Perspectives and frames in pursuit of ultimate understanding', in Stephen R. Grimm (ed.), *Varieties of Understanding* (New York: Oxford University Press).

Camp, Elisabeth. 2022. '*Representation in Cognitive Science* by Nicholas Shea: Organization and structure in the service of systematicity and productivity', *Studies in History and Philosophy of Science*, 92: 264–66.

Cao, Rosa, and Daniel Yamins. 2021. 'Explanatory models in neuroscience: Part 1 – Taking mechanistic abstraction seriously', *arXiv2104.01490v2*.

Carandini, Matteo, and David J. Heeger. 2012. 'Normalization as a canonical neural computation', *Nature Reviews Neuroscience*, 13: 51–62.

Carey, Susan. 2009. *The Origin of Concepts* (New York: Oxford University Press).

Carey, Susan, and Elizabeth Spelke. 1996. 'Science and core knowledge', *Philosophy of Science*, 63: 515–33.

Carroll, Lewis. 1995. 'What the tortoise said to Achilles', *Mind*, 104: 691–3.

Carruthers, Peter. 2003. 'On Fodor's problem', *Mind & Language*, 18: 502–23.

Carruthers, Peter. 2011a. 'Creative action in mind', *Philosophical Psychology*, 24: 437–61.

Carruthers, Peter. 2011b. *The Opacity of Mind: An Integrative Theory of Self-Knowledge* (Oxford: Oxford University Press).

Carruthers, Peter. 2015. *The Centered Mind: What the Science of Working Memory Shows Us about the Nature of Human Thought* (New York: Oxford University Press).

Carruthers, Peter. 2021. 'Explicit nonconceptual metacognition', *Philosophical Studies*, 178: 2337–56.

Carruthers, Peter, and D. M. Williams. 2022. 'Model-free metacognition', *Cognition*, 225: 105117.

Carston, Robyn. 2010. 'Explicit communication and "free" pragmatic enrichment', in Belén Soria and Esther Romero (eds.), *Explicit Communication* (Basingstoke: Palgrave Macmillan).

Casasanto, Daniel, and Gary Lupyan. 2015. 'All concepts are ad hoc concepts', in Eric Margolis and Stephen Laurence (eds.), *The Conceptual Mind* (New York: Oxford University Press).

Castro-Rodrigues, Pedro, Thomas Akam, Ivar Snorasson, Marta Camacho, Vitor Paixão, Ana Maia, J. Bernardo Barahona-Corrêa, Peter Dayan, H. Blair Simpson, and Rui M. Costa. 2022. 'Explicit knowledge of task structure is a primary determinant of human model-based action', *Nature Human Behaviour*, 6: 1126–41.

Charest, Ian, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. 2014. 'Unique semantic space in the brain of each beholder predicts perceived similarity', *Proceedings of the National Academy of Sciences of the United States of America*, 111: 14565–70.

Chomsky, Noam. 2017. 'Language architecture and its import for evolution', *Neuroscience & Biobehavioral Reviews*, 81: 295–300.

Chow, Sheldon J. 2013. 'What's the problem with the frame problem?', *Review of Philosophy and Psychology*, 4: 309–31.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. 2022. 'PaLM: Scaling language modeling with pathways', *arXiv:2204.02311v2.*

Churchland, Paul M. 1998. 'Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered', *The Journal of Philosophy*, 95: 5–32.

Churchland, Paul M. 2012. *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals* (Cambridge, MA: MIT Press).

Cichy, Radoslaw Martin, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. 2016. 'Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence', *Scientific Reports*, 6: 27755.

Clarke, Sam. 2021. 'Mapping the visual icon', *Philosophical Quarterly*, 72: 552–77.

Clayton, Nicola S., and Anthony Dickinson. 1998. 'Episodic-like memory during cache recovery by scrub jays', *Nature*, 395: 272–4.

Coelho Mollo, Dimitri, and Alfredo Vernazzani. 2023. 'The Formats of Cognitive Representation: A Computational Account', *Philosophy of Science*. https://doi.org/ 10.1017/psa.2023.123

Cohen, Michael A., Cole Dembski, Kevin Ortego, Clay Steinhilber, and Michael Pitts. 2023. 'A novel signature of visual awareness independent of post-perceptual processing', *bioRxiv*: 2023.06.09.543951.

Collins, John. 2011. *The Unity of Linguistic Meaning* (New York: Oxford University Press).

Connell, Louise, and Dermot Lynott. 2014. 'Principles of representation: Why you can't represent the same concept twice', *Topics in Cognitive Science*, 6: 390–406.

Conwell, Colin, and Tomer Ullman. 2022. 'Testing relational understanding in text-guided image generation', *arXiv:2208.00005.*

Corbett, Faye, Elizabeth Jefferies, Sheeba Ehsan, and Matthew A. Lambon Ralph. 2009. 'Different impairments of semantic cognition in semantic dementia and semantic aphasia: Evidence from the non-verbal domain', *Brain*, 132: 2593–608.

Corneil, Dane S., and Wulfram Gerstner. 2015. 'Attractor network dynamics enable preplay and rapid path planning in maze-like environments', *Advances in Neural Information Processing Systems*, Conference proceedings: 1684–92.

Courtin, J., Y. Bitterman, S. Muller, J. Hinz, K. M. Hagihara, C. Muller, and A. Luthi. 2022. 'A neuronal mechanism for motivational control of behavior', *Science*, 375: eabg7277.

Cowan, Nelson. 2008. 'What are the differences between long-term, short-term, and working memory?' in Wayne S. Sossin, Jean-Claude Lacaille, Vincent F. Castellucci, and Sylvie Belleville (eds.), *Progress in Brain Research* (Amsterdam: Elsevier).

Cowan, Nelson, Candice C. Morey, and Moshe Naveh-Benjamin. 2021. 'An embedded-processes approach to working memory', in Robert H. Logie, Valerie Camos, and Nelson Cowan (eds.), *Working Memory: State of the Science* (New York: Oxford University Press).

Cox, James R., and Richard A. Griggs. 1982. 'The effects of experience on performance in Wason's selection task', *Memory & Cognition*, 10: 496–502.

Creswell, Antonia, Murray Shanahan, and Irina Higgins. 2023. 'Selection-inference: Exploiting large language models for interpretable logical reasoning', *ICLR*, *arXiv:2205.09712*.

Daw, Nathaniel D., Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. 2011. 'Model-based influences on humans' choices and striatal prediction errors', *Neuron*, 69: 1204–15.

De Neys, Wim. 2012. 'Bias and conflict: A case for logical intuitions', *Perspectives on Psychological Science*, 7: 28–38.

De Neys, Wim. 2023. 'Advancing theorizing about fast-and-slow thinking', *Behavioral and Brain Sciences*, 46: 1–68.

De Neys, Wim, Sofie Cromheeke, and Magda Osman. 2011. 'Biased but in doubt: Conflict and decision confidence', *PLOS One*, 6: e15954.

De Neys, Wim, Elke Moyens, and Debora Vansteenwegen. 2010. 'Feeling we're biased: Autonomic arousal and reasoning conflict', *Cognitive, Affective, & Behavioral Neuroscience*, 10: 208–16.

Dehaene, Stanislas, Fosca Al Roumi, Yair Lakretz, Samuel Planton, and Mathias Sablé-Meyer. 2022. 'Symbols and mental programs: A hypothesis about human singularity', *Trends in Cognitive Sciences*, 26: 751–66.

Dennett, Daniel C. 2008. *Kinds of Minds: Toward an Understanding of Consciousness* (New York: Basic Books).

Dickinson, Anthony, and Bernard Balleine. 1994. 'Motivational control of goal-directed action', *Animal Learning & Behavior*, 22: 1–18.

Donoso, Mael, Anne G. E. Collins, and Etienne Koechlin. 2014. 'Foundations of human reasoning in the prefrontal cortex', *Science*, 344: 1481–6.

Dragoi, George, and Susumu Tonegawa. 2011. 'Preplay of future place cell sequences by hippocampal cellular assemblies', *Nature*, 469: 397–401.

Drożdżowicz, Anna. 2019. 'Do we hear meanings? Between perception and cognition', *Inquiry*, 66: 196–228.

Duncan, John. 2001. 'An adaptive coding model of neural function in prefrontal cortex', *Nature Reviews Neuroscience*, 2: 820–9.

Dutilh Novaes, Catarina. 2020. *The Dialogical Roots of Deduction: Historical, Cognitive, and Philosophical Perspectives on Reasoning* (Cambridge: Cambridge University Press).

Egan, Frances. 1992. 'Individualism, computation, and perceptual content', *Mind*, 101: 443–59.

Einstein, Albert. 1970. 'Autobiographical notes', in P. A. Schilpp (ed.), *Albert Einstein—Philosopher Scientist* (New York: MJF Books).

Eliasmith, Chris. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition* (Oxford: Oxford University Press).

Engle, Randall W. 2010. 'Role of working-memory capacity in cognitive control', *Current Anthropology*, 51: S17–S26.

Ernst, Marc O., and Martin S. Banks. 2002. 'Humans integrate visual and haptic information in a statistically optimal fashion', *Nature*, 415: 429–33.

Evans, Gareth. 1982. *The Varieties of Reference* (Oxford: Oxford University Press).

Evans, Jonathan St B. T., and Keith E. Stanovich. 2013. 'Dual-process theories of higher cognition: Advancing the debate', *Perspectives on Psychological Science*, 8: 223–41.

Evans, Vyvyan. 2015. 'What's in a concept? Analog versus parametric concepts in LCCM theory', in Eric Margolis and Stephen Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (Cambridge, MA: MIT Press).

Fedorenko, Evelina, and Rosemary Varley. 2016. 'Language and thought are not the same thing: Evidence from neuroimaging and neurological patients', *Annals of the New York Academy of Sciences*, 1369: 132–53.

Figdor, Carrie. 2009. 'Semantic externalism and the mechanics of thought', *Minds and Machines*, 19: 1–24.

Fleming, Stephen M., and Nathaniel D. Daw. 2017. 'Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation', *Psychological Review*, 124: 91–114.

Fodor, J. D., J. A Fodor, and M. F. Garrett. 1975. 'The psychological unreality of semantic representations', *Linguistic Inquiry*, 6: 515–31.

Fodor, Jerry A. 1975. *The Language of Thought* (Cambridge, MA: Harvard University Press).

Fodor, Jerry A. 1985. 'Précis of *The Modularity of Mind*', *Behavioral and Brain Sciences*, 8: 1–42.

Fodor, Jerry A. 1987. 'Modules, frames, fridgeons, sleeping dogs, and the music of the spheres', in J. Garfield (ed.), *Modularity in Knowledge Representation and Natural-Language Understanding* (Cambridge, MA: MIT Press).

Fodor, Jerry A. 1991. 'A modal argument for narrow content', *The Journal of Philosophy*, 88: 5–26.

Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong* (New York: Oxford University Press).

Fodor, Jerry A. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology* (Cambridge, MA: MIT Press).

Fodor, Jerry A. 2003. *Hume Variations* (New York: Oxford University Press).

Fodor, Jerry A. 2008. *LOT 2: The Language of Thought Revisited* (Oxford: Oxford University Press).

Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. 'Connectionism and cognitive architecture: A critical analysis', *Cognition*, 28: 3–71.

Frankland, Steven M., and Joshua D. Greene. 2015. 'An architecture for encoding sentence meaning in left mid-superior temporal cortex', *Proceedings of the National Academy of Sciences of the United States of America*, 112: 11732–7.

Frankland, Steven M., and Joshua D. Greene. 2020. 'Concepts and compositionality: In search of the brain's language of thought', *Annual Review of Psychology*, 71: 273–303.

Frankland, Steven M., Taylor W. Webb, and Jonathan D. Cohen. Preprint. 'No coincidence, George: Capacity limits are the curse of compositionality'. https://doi.org/10.31234/osf.io/cjuxb.

Frederick, Shane. 2005. 'Cognitive reflection and decision making', *Journal of Economic Perspectives*, 19: 25–42.

Fricker, Elizabeth. 2003. 'Understanding and knowledge of what is said', in A. Barber (ed.), *Epistemology of Language* (Oxford: Oxford University Press).

Frith, Christopher D., and Uta Frith. 1978. 'Feature selection and classification: A developmental study', *Journal of Experimental Child Psychology*, 25: 413–28.

Frith, Uta, and Maggie Snowling. 1983. 'Reading for meaning and reading for sound in autistic and dyslexic children', *British Journal of Developmental Psychology*, 1: 329–42.

Gabay, Yafit, Casey L. Roark, and Lori L. Holt. 2023. 'Impaired and spared auditory category learning in developmental dyslexia', *Psychological Science*, 34: 468–80.

Gallistel, Charles R. 2008. 'Learning and representation', in R. Menzel and J. Byrne (eds.), *Learning and Memory: A Comprehensive Reference* (Amsterdam: Elsevier).

Gallistel, Charles R., and Adam Philip King. 2009. *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience* (Malden, MA: Wiley-Blackwell).

Gazzaniga, Michael S., Richard B. Ivry, and George R. Mangun. 2019. *Cognitive Neuroscience: The Biology of the Mind* (New York: W. W. Norton).

Gendler, Tamar Szabó. 2000. 'The puzzle of imaginative resistance', *The Journal of Philosophy*, 97: 55–81.

Gendler, Tamar Szabó. 2004. 'Thought experiments rethought—and reperceived', *Philosophy of Science*, 71: 1152–63.

Gentner, Dedre, and Michael Jeziorski. 1993. 'The shift from metaphor to analogy in Western science', in A. Ortony (ed.), *Metaphor and Thought* (Cambridge: Cambridge University Press).

Gentner, Dedre, and Francisco Maravilla. 2018. 'Analogical reasoning', in Linden J. Ball and Valerie A. Thompson (eds.), *Routledge International Handbook of Thinking and Reasoning* (New York: Routledge).

Gershman, Samuel J., Arthur B. Markman, and A. Ross Otto. 2014. 'Retrospective revaluation in sequential decision making: A tale of two systems', *Journal of Experimental Psychology: General*, 143: 182.

Gigerenzer, Gerd, and Klaus Hug. 1992. 'Domain-specific reasoning: Social contracts, cheating, and perspective change', *Cognition*, 43: 127–71.

Giora, Rachel. 2002. *On Our Mind: Salience, Context, and Figurative Language* (New York: Oxford University Press).

Glanzberg, Michael. 2008. 'Metaphor and lexical semantics', *The Baltic International Yearbook of Cognition, Logic and Communication*, 3: 1–47.

Glass, Arnold L., and Keith J. Holyoak. 1974. 'Alternative conceptions of semantic theory', *Cognition*, 3: 313–39.

Glenberg, Arthur M., and Michael P. Kaschak. 2002. 'Grounding language in action', *Psychonomic Bulletin & Review*, 9: 558–65.

Glosser, Guila, Rhonda B. Friedman, Patrick K. Grugan, Jefferson H. Lee, and Murray Grossman. 1998. 'Lexical semantic and associative priming in Alzheimer's disease', *Neuropsychology*, 12: 218–24.

Godfrey-Smith, Peter. 2017. 'Senders, receivers, and symbolic artifacts', *Biological Theory*, 12: 275–86.

Gold, Joshua I., and Michael N. Shadlen. 2001. 'Neural computations that underlie decisions about sensory stimuli', *Trends in Cognitive Sciences*, 5: 10–16.

Gold, Joshua I., and Michael N. Shadlen. 2007. 'The neural basis of decision making', *Annual Review of Neuroscience*, 30: 535–74.

Goodman, Nelson. 1955. *Fact, Fiction, & Forecast* (Cambridge, MA: Harvard University Press).

Goodman, Nelson. 1968. *Languages of Art* (New York: Bobbs-Merrill).

Goodman, Noah D., Joshua B. Tenenbaum, and Tobias Gerstenberg. 2015. 'Concepts in a probabilistic language of thought', in Eric Margolis and Stephen Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (Cambridge, MA: MIT Press).

Gopnik, Alison, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. 2004. 'A theory of causal learning in children: Causal maps and Bayes nets', *Psychological Review*, 111: 3–32.

Goswami, Usha. 2001. 'Analogical reasoning in children', in Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov (eds.), *The Analogical Mind: Perspectives from Cognitive Science* (Cambridge, MA: MIT Press).

Grand, Gabriel, Idan A. Blank, Francisco Pereira, and Evelina Fedorenko. 2022. 'Semantic projection recovers rich human knowledge of multiple object features from word embeddings', *Nature Human Behaviour*, 6: 975–87.

Graves, Alex, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, and John Agapiou. 2016. 'Hybrid computing using a neural network with dynamic external memory', *Nature*, 538: 471–6.

Grieves, Roddy M., and Kate J. Jeffery. 2017. 'The representation of space in the brain', *Behavioural Processes*, 135: 113–31.

Griffiths, Paul, and Stefan Linquist. 2022. 'The distinction between innate and acquired characteristics', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition). https://plato.stanford.edu/archives/spr2024/entries/innate-acquired/.

Grimm, Stephen. 2012. 'The value of understanding', *Philosophy Compass*, 7: 103–17.

Grush, Rick. 2004. 'The emulation theory of representation: Motor control, imagery, and perception', *Behavioral and Brain Sciences*, 27: 377–96.

Güçlü, Umut, and Marcel A. J. van Gerven. 2015. 'Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream', *Journal of Neuroscience*, 35: 10005–14.

Hadjiosif, Alkis M., John W. Krakauer, and Adrian M. Haith. 2021. 'Did we get sensorimotor adaptation wrong? Implicit adaptation as direct policy updating rather than forward-model-based learning', *Journal of Neuroscience*, 41: 2747–61.

Halford, Graeme S., William H. Wilson, and Steven Phillips. 2010. 'Relational knowledge: The foundation of higher cognition', *Trends in Cognitive Sciences*, 14: 497–505.

Hampton, James A. 2015. 'Concepts in the semantic triangle', in Eric Margolis and Stephen Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (Cambridge, MA: MIT Press).

Hampton, James A., and Alessia Passanisi. 2016. 'When intensions do not map onto extensions: Individual differences in conceptualization', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42: 505–23.

Harmon-Jones, Eddie, and Judson Mills. 2019. 'An introduction to cognitive dissonance theory and an overview of current perspectives on the theory', in Eddie Harmon-Jones (ed.), *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology* (2nd edition) (Washington, DC: American Psychological Association).

Harris, Daniel W. 2020. 'Semantics without semantic content', *Mind & Language*, 37: 304–28.

Haslanger, Sally. 2000. 'Gender and race: (What) are they? (What) do we want them to be?', *Noûs*, 34: 31–55.

Hasson, Uri, Janice Chen, and Christopher J. Honey. 2015. 'Hierarchical process memory: Memory as an integral component of information processing', *Trends in Cognitive Sciences*, 19: 304–13.

Haynes, John-Dylan, and Geraint Rees. 2005. 'Predicting the stream of consciousness from activity in human visual cortex', *Current Biology*, 15: 1301–7.

Heit, Evan, and Caren M. Rotello. 2010. 'Relations between inductive reasoning and deductive reasoning', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36: 805–12.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. 'The weirdest people in the world?', *Behavioral and Brain Sciences*, 33: 61–83.

Heyes, Cecilia M. 2014. 'False belief in infancy: A fresh look', *Developmental Science*, 17: 647–59.

Heyes, Cecilia M. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking* (Cambridge, MA: Harvard University Press).

Heyes, Cecilia M., D. Bang, Nicholas Shea, Chris D. Frith, and S. M. Fleming. 2020. 'Knowing ourselves together: The cultural origins of metacognition', *Trends in Cognitive Sciences*, 24: 349–62.

Heyes, Cecilia M., and Chris D. Frith. 2014. 'The cultural evolution of mind reading', *Science*, 344: 1243091.

Hummel, John E., and Keith J. Holyoak. 2003. 'A symbolic-connectionist theory of relational inference and generalization', *Psychological Review*, 110: 220–64.

Hummel, John E., Keith J. Holyoak, Collin B. Green, Leonidas A. A. Doumas, Derek Devnich, Aniket Kittur, and Donald J. Kalar. 2004. 'A solution to the binding problem for compositional connectionism', *AAAI Technical Report*, 3: 31–4.

Huth, Alexander G., Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. 'Natural speech reveals the semantic maps that tile human cerebral cortex', *Nature*, 532: 453–8.

Huth, Alexander G., Shinji Nishimoto, An T. Vu, and Jack L. Gallant. 2012. 'A continuous semantic space describes the representation of thousands of object and action categories across the human brain', *Neuron*, 76: 1210–24.

Icard, Thomas. 2021. 'Why be random?', *Mind*, 130: 111–39.

Jackson, Rebecca L., Timothy T. Rogers, and Matthew A. Lambon Ralph. 2021. 'Reverse-engineering the cortical architecture for controlled semantic cognition', *Nature Human Behaviour*, 5: 774–86.

Jacoby, Larry L., Christopher N. Wahlheim, and Jennifer H. Coane. 2010. 'Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36: 1441–51.

Johnson-Laird, Philip N., and Ruth M. J. Byrne. 2002. ‘Conditionals: A theory of meaning, pragmatics, and inference’, *Psychological Review*, 109: 646–78.

Jorba, Marta, and Agustin Vicente. 2014. ‘Cognitive phenomenology, access to contents, and inner speech’, *Journal of Consciousness Studies*, 21: 74–99.

Kan, Irene P., Lawrence W. Barsalou, Karen Olseth Solomon, Jeris K. Minor, and Sharon L. Thompson-Schill. 2003. ‘Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge’, *Cognitive Neuropsychology*, 20: 525–40.

Keil, Frank C. 1992. *Concepts, Kinds, and Cognitive Development* (Cambridge, MA: MIT Press).

Keogh, Rebecca, and Joel Pearson. 2018. ‘The blind mind: No sensory visual imagery in aphantasia’, *Cortex*, 105: 53–60.

Khajeh-Alijani, Azadeh, Robert Urbanczik, and Walter Senn. 2015. ‘Scale-free navigational planning by neuronal traveling waves’, *PLOS One*, 10: e0127269.

Khaligh-Razavi, Seyed-Mahdi, and Nikolaus Kriegeskorte. 2014. ‘Deep supervised, but not unsupervised, models may explain IT cortical representation’, *PLOS Computational Biology*, 10: e1003915.

King, Jeffrey C. 2009. ‘Questions of unity’, *Proceedings of the Aristotelian Society*, 109: 257–77.

Knowlton, Barbara J., Robert G. Morrison, John E. Hummel, and Keith J. Holyoak. 2012. ‘A neurocomputational system for relational reasoning’, *Trends in Cognitive Sciences*, 16: 373–81.

Kolling, Nils, Jacqueline Scholl, Adam Chekroud, Hailey A. Trier, and Matthew F. S. Rushworth. 2018. ‘Prospection, perseverance, and insight in sequential behavior’, *Neuron*, 99: 1069–82, e7.

Koriat, Asher. 2012a. ‘The self-consistency model of subjective confidence’, *Psychological Review*, 119: 80–113.

Koriat, Asher. 2012b. ‘The subjective confidence in one’s knowledge and judgments: Some metatheoretical considerations’, in Michael J. Beran, Johannes L. Brandl, Josef Perner, and Joëlle Proust (eds.), *The Foundations of Metacognition* (New York: Oxford University Press).

Koriat, Asher. 2016. ‘Metacognition: Decision making processes in self-monitoring and self-regulation’, in Gideon Keren and George Wu (eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (Malden, MA: Wiley-Blackwell).

Koriat, Asher, and Shiri Adiv. 2015. ‘The self-consistency theory of subjective confidence’, in John Dunlosky and Sarah K. Tauber (eds.), *The Oxford Handbook of Metamemory* (New York: Oxford University Press).

Koriat, Asher, and Tore Helstrup. 2007. ‘Metacognitive aspects of memory’, in Svein Magnussen and Tore Helstrup (eds.), *Everyday Memory* (Hove: Psychology Press).

Koriat, Asher, Sarah Lichtenstein, and Baruch Fischhoff. 1980. ‘Reasons for confidence’, *Journal of Experimental Psychology: Human Learning and Memory*, 6: 107–18.

Koriat, Asher, Hilit Ma’ayan, and Ravit Nussinson. 2006. ‘The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior’, *Journal of Experimental Psychology: General*, 135: 36–69.

Kornblith, Hilary. 2012. *On Reflection* (Oxford: Oxford University Press).

Krausz, Timothy A., Alison E. Comrie, Loren M. Frank, Nathaniel D. Daw, and Joshua D. Berke. 2023. ‘Dual credit assignment processes underlie dopamine signals in a complex spatial environment’, *bioRxiv*: 2023.02.15.528738.

Krawczyk, Daniel C. 2012. ‘The cognition and neuroscience of relational reasoning’, *Brain Research*, 1428: 13–23.

Kriegeskorte, Nikolaus, and Rogier A. Kievit. 2013. ‘Representational geometry: Integrating cognition, computation, and the brain’, *Trends in Cognitive Sciences*, 17: 401–12.

Kriete, Trenton, David C. Noelle, Jonathan D. Cohen, and Randall C. O’Reilly. 2013. ‘Indirection and symbol-like processing in the prefrontal cortex and basal ganglia’, *Proceedings of the National Academy of Sciences of the United States of America*, 110: 16390–5.

Kripke, Saul A. 1982. *Wittgenstein on Rules and Private Language: An Elementary Exposition* (Cambridge, MA: Harvard University Press).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ‘Imagenet classification with deep convolutional neural networks’, in Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25* (New York: Curran Associates).

Kroll, Judith F., and Erika Stewart. 1994. 'Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations', *Journal of Memory and Language*, 33: 149–74.

Kroll, Judith F., Janet G. Van Hell, Natasha Tokowicz, and David W. Green. 2010. 'The revised hierarchical model: A critical review and assessment', *Bilingualism: Language and Cognition*, 13: 373–81.

Kurth-Nelson, Zeb, Timothy Behrens, Greg Wayne, Kevin Miller, Lennart Luettgau, Ray Dolan, Yunzhe Liu, and Philipp Schwartenbeck. 2023. 'Replay and compositional computation', *Neuron*, 111: 454–69.

Laakso, Aarre, and Garrison Cottrell. 2000. 'Content and cluster analysis: Assessing representational similarity in neural systems', *Philosophical Psychology*, 13: 47–76.

Lackner, James R., and Merrill F. Garrett. 1972. 'Resolving ambiguity: Effects of biasing context in the unattended ear', *Cognition*, 1: 359–72.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. 'Building machines that learn and think like people', *Behavioral and Brain Sciences*, 40: e253.

Lambon Ralph, Matthew A., Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. 2017. 'The neural and computational bases of semantic cognition', *Nature Reviews Neuroscience*, 18: 42–55.

Lande, Kevin J. 2021. 'Mental structures', *Noûs*, 55: 649–77.

Langdon, Angela, Matthew Botvinick, Hiroyuki Nakahara, Keiji Tanaka, Masayuki Matsumoto, and Ryota Kanai. 2022. 'Meta-learning, social cognition and consciousness in brains and machines', *Neural Networks*, 145: 80–9.

Langdon, Christopher, Mikhail Genkin, and Tatiana A. Engel. 2023. 'A unifying perspective on neural manifolds and circuits for cognition', *Nature Reviews Neuroscience*, 24: 363–77.

Laurence, Stephen, and Eric Margolis. 1999. 'Concepts and cognitive science', in Eric Margolis and Stephen Laurence (eds.), *Concepts: Core Readings* (Cambridge, MA: MIT Press).

Lea, R. Brooke, Elizabeth J. Mulligan, and Jennifer Lee Walton. 2005. 'Accessing distant premise information: How memory feeds reasoning', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31: 387–95.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep learning', *Nature*, 521: 436–44.

Lee, Andrew Y., Joshua Myers, and Gabriel Oak Rabin. 2023. 'The structure of analog representation', *Noûs*, 57: 209–37.

Lee, Hongmi, and Brice A. Kuhl. 2016. 'Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex', *Journal of Neuroscience*, 36: 6069–82.

Lee, Sang Wan, Shinsuke Shimojo, and John P. O'Doherty. 2014. 'Neural computations underlying arbitration between model-based and model-free learning', *Neuron*, 81: 687–99.

Lewandowsky, Stephan, and Kim Kirsner. 2000. 'Knowledge partitioning: Context-dependent use of expertise', *Memory & Cognition*, 28: 295–305.

Liu, Michelle. 2024. 'Mental simulation and language comprehension: The case of copredication', *Mind & Language*, 39: 2–21.

Liu, Yunzhe, Raymond J. Dolan, Zeb Kurth-Nelson, and Timothy E. J. Behrens. 2019. 'Human replay spontaneously reorganizes experience', *Cell*, 178: 640–52, e14.

Liu, Yunzhe, Marcelo G. Mattar, Timothy E. J. Behrens, Nathaniel D. Daw, and Raymond J. Dolan. 2021. 'Experience replay is associated with efficient nonlocal learning', *Science*, 372: eabf1357.

Longworth, Guy. 2016. 'Understanding what was said', *Synthese*, 195: 815–34.

Lovett, Andrew, and Kenneth Forbus. 2017. 'Modeling visual problem solving as analogical reasoning', *Psychological Review*, 124: 60–90.

Low, Jason, and Josef Perner. 2012. 'Implicit and explicit theory of mind: State of the art', *British Journal of Developmental Psychology*, 30: 1–13.

McCoy, R. Thomas, Tal Linzen, Ewan Dunbar, and Paul Smolensky. 2019. 'RNNs implicitly implement tensor product representations'. International Conference on Learning Representations, *arXiv:1812.08718*.

McCrink, Koleen, and Elizabeth S. Spelke. 2010. 'Core multiplication in childhood', *Cognition*, 116: 204–16.

Machery, Edouard. 2009. *Doing without Concepts* (New York: Oxford University Press).

Machery, Edouard. 2015. 'By default: Concepts are accessed in a context-independent manner', in Eric Margolis and Stephen Laurence (eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (Cambridge, MA: MIT Press).

Machery, Edouard. 2017. *Philosophy within Its Proper Bounds* (New York: Oxford University Press).

McKinsey, Michael. 1991. 'Anti-individualism and privileged access', *Analysis*, 51: 9–16.

McLaughlin, Brian P., and Michael Tye. 1998. 'Is content-externalism compatible with privileged access?', *The Philosophical Review*, 107: 349–80.

Magidor, Ofra. 2009. 'II—The last dogma of type confusions', *Proceedings of the Aristotelian Society*, 109: 1–29.

Mahowald, Kyle, Evgeniia Diachek, Edward Gibson, Evelina Fedorenko, and Richard Futrell. 2022. 'Grammatical cues are largely, but not completely, redundant with word meanings in natural language', *arXiv:2201.12911*.

Maley, Corey J. 2011. 'Analog and digital, continuous and discrete', *Philosophical Studies*, 155: 117–31.

Maley, Corey J. 2023. 'Analogue computation and representation', *British Journal for the Philosophy of Science*, 74: 739–69.

Malt, Barbara C. 2010. 'Why we should do without concepts', *Mind & Language*, 25: 622–33.

Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. 'Context-dependent computation by recurrent dynamics in prefrontal cortex', *Nature*, 503: 78–84.

Marcel, Anthony J. 1980. 'Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context', in R. S. Nickerson (ed.), *Attention and Performance VIII* (Hillsdale, NJ: Erlbaum).

Margolis, Eric, and Stephen Laurence. 2019. 'Concepts', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition). https://plato.stanford.edu/archives/fall2023/entries/concepts/.

Margolis, Eric, and Stephen Laurence. 2023. 'Making sense of domain specificity', *Cognition*, 240: 105583.

Markman, Arthur B., and Dedre Gentner. 1993. 'Structural alignment during similarity comparisons', *Cognitive Psychology*, 25: 431–67.

Masse, Nicolas Y., Guangyu R. Yang, H. Francis Song, Xiao-Jing Wang, and David J. Freedman. 2019. 'Circuit mechanisms for the maintenance and manipulation of information in working memory', *Nature Neuroscience*, 22: 1159–67.

Mattar, Marcelo G., and Nathaniel D. Daw. 2018. 'Prioritized memory access explains planning and hippocampal replay', *Nature Neuroscience*, 21: 1609–17.

Mazancieux, Audrey, Stephen M. Fleming, Céline Souchay, and Chris J. A. Moulin. 2020. 'Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks', *Journal of Experimental Psychology: General*, 149: 1788–99.

Medin, Douglas L., and Edward J. Shoben. 1988. 'Context and structure in conceptual combination', *Cognitive Psychology*, 20: 158–90.

Meyniel, Florent, and Stanislas Dehaene. 2017. 'Brain networks for confidence weighting and hierarchical inference during probabilistic learning', *Proceedings of the National Academy of Sciences of the United States of America*, 114: E3859–E3868.

Miall, R. Christopher, and Daniel M. Wolpert. 1996. 'Forward models for physiological motor control', *Neural Networks*, 9: 1265–79.

Miller, Earl K., and Jonathan D. Cohen. 2001. 'An integrative theory of prefrontal cortex function', *Annual Review of Neuroscience*, 24: 167–202.

Millikan, Ruth Garrett. 1984. *Language, Thought and Other Biological Categories* (Cambridge, MA: MIT Press).

Millikan, Ruth Garrett. 2000. *On Clear and Confused Ideas* (Cambridge: Cambridge University Press).

Millikan, Ruth Garrett. 2017. *Beyond Concepts: Unicepts, Language, and Natural Information* (New York: Oxford University Press).

Miyamoto, Kentaro, Matthew F. S. Rushworth, and Nicholas Shea. 2023. 'Imagining the future self through thought experiments', *Trends in Cognitive Sciences*, 27: 446–55.

Mok, Robert M., and Bradley C. Love. 2019. 'A non-spatial account of place and grid cells based on clustering models of concept learning', *Nature Communications*, 10: 5685.

Montague, Richard. 1974. 'English as a formal language', in Richmond H. Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague* (New Haven: Yale University Press).

Moore, George Edward. 1953. 'Propositions', in *Some Main Problems of Philosophy* (New York: Macmillan).

Moortgat, Michael. 2010. 'Typelogical grammar', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition). https://plato.stanford.edu/archives/spr2014/entries/typelogical-grammar/.

Mueller, Martin, and Radiger Wehner. 1988. 'Path integration in desert ants, Cataglyphis fortis', *Proceedings of the National Academy of Sciences of the United States of America*, 85: 5287–90.

Mulcahy, Nicholas J., and Josep Call. 2006. 'How great apes perform on a modified trap-tube task', *Animal Cognition*, 9: 193–9.

Murphy, Gregory L. 2002. *The Big Book of Concepts* (Cambridge, MA: MIT Press).

Musslick, Sebastian, and Jonathan D. Cohen. 2021. 'Rationalizing constraints on the capacity for cognitive control', *Trends in Cognitive Sciences*, 25: 757–75.

Na, Soojung, Dongil Chung, Andreas Hula, Ofer Perl, Jennifer Jung, Matthew Heflin, Sylvia Blackmore, Vincenzo G. Fiore, Peter Dayan, and Xiaosi Gu. 2021. 'Humans use forward thinking to exploit social controllability', *Elife*, 10: e64983.

Nanda, Neel, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. 2023. 'Progress measures for grokking via mechanistic interpretability', *arXiv:2301.05217*.

Nelli, Stephanie, Lukas Braun, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. 2023. 'Neural knowledge assembly in humans and neural networks', *Neuron*, 111: 1504–16, e9.

Nelson, Thomas O., and Louis Narens. 1990. 'Metamemory: A theoretical framework and new findings', in Gordon H. Bower (ed.), *The Psychology of Learning and Motivation*, *Vol. 26* (San Francisco: Academic Press).

Nersessian, Nancy J. 2018. 'Cognitive science, mental modeling, and thought experiments', in Michael T. Stuart, Yiftach Fehige, and James Robert Brown (eds.), *The Routledge Companion to Thought Experiments* (Abingdon: Routledge).

Newell, Ben R., and David R. Shanks. 2014. 'Unconscious influences on decision making: A critical review', *Behavioral and Brain Sciences*, 37: 1–19.

Nieder, Andreas. 2016. 'The neuronal code for number', *Nature Reviews Neuroscience*, 17: 366–82.

Nieder, Andreas, and Stanislas Dehaene. 2009. 'Representation of number in the brain', *Annual Review of Neuroscience*, 32: 185–208.

Norman, Donald A., and Tim Shallice. 1986. 'Attention to action: Willed and automatic control of behavior', in Richard J. Davidson, Gary E. Schwartz, and David Shapiro (eds.), *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4* (New York: Springer).

Norman, Yitzhak, Erin M. Yeagle, Simon Khuvis, Michal Harel, Ashesh D. Mehta, and Rafael Malach. 2019. 'Hippocampal sharp-wave ripples linked to visual episodic recollection in humans', *Science*: eaax1030.

O'Doherty, John P., Jeffrey Cockburn, and Wolfgang M. Pauli. 2017. 'Learning, reward, and decision making', *Annual Review of Psychology*, 68: 73–100.

O'Doherty, John P., Peter Dayan, Karl Friston, Hugo Critchley, and Raymond J. Dolan. 2003. 'Temporal difference models and reward-related learning in the human brain', *Neuron*, 38: 329–37.

O'Keefe, John, and Neil Burgess. 1996. 'Geometric determinants of the place fields of hippocampal neurons', *Nature*, 381: 425–8.

O'Keefe, John, and Lynn Nadel. 1978. *The Hippocampus as a Cognitive Map* (Oxford: Clarendon Press).

Oaksford, Mike, and Nick Chater. 1994. 'A rational analysis of the selection task as optimal data selection', *Psychological Review*, 101: 608–31.

Oberauer, Klaus. 2009. 'Design for a working memory', *Psychology of Learning and Motivation*, 51: 45–100.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. 'Zoom in: An introduction to circuits', *Distill*, 5: e00024, 001.

Oosterhof, Nikolaas N., and Alexander Todorov. 2008. 'The functional basis of face evaluation', *Proceedings of the National Academy of Sciences of the United States of America*, 105: 11087–92.

Orilia, Francesco, and Michele Paolini Paoletti. 2020. 'Properties', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition). https://plato.stanford.edu/archives/spr2022/entries/properties/.

Pado, Sebastian, and Ido Dagan. 2016. 'Textual entailment', in Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics* (Oxford: Oxford University Press).

Paivio, Allan. 1986. *Mental Representations: A Dual Coding Approach* (New York: Oxford University Press).

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. 'Generative agents: Interactive simulacra of human behavior', *arXiv:2304.03442*.

Park, Seongmin A., Douglas S. Miller, and Erie D. Boorman. 2021. 'Inferences on a multidimensional social hierarchy use a grid-like code', *Nature Neuroscience*, 24: 1292–301.

Passingham, Richard. 2021. *Understanding the Prefrontal Cortex: Selective Advantage, Connectivity, and Neural Operations* (Oxford: Oxford University Press).

Peacocke, Christopher. 1992. *A Study of Concepts* (Cambridge, MA: MIT Press).

Peacocke, Christopher. 1993. 'Externalist explanation', *Proceedings of the Aristotelian Society*, 93: 203–30.

Peacocke, Christopher. 2019. 'Spatial perception, magnitudes, and analogue representation', in Tony Cheng, Ophelia Deroy, and Charles Spence (eds.), *Spatial Senses: Philosophy of Perception in an Age of Science* (New York: Routledge).

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference* (Cambridge: Cambridge University Press).

Penn, Derek C., Keith J. Holyoak, and Daniel J. Povinelli. 2008. 'Darwin's mistake: Explaining the discontinuity between human and nonhuman minds', *Behavioral and Brain Sciences*, 31: 109–30.

Piantadosi, Steven T., and Robert A. Jacobs. 2016. 'Four problems solved by the probabilistic language of thought', *Current Directions in Psychological Science*, 25: 54–9.

Pietroski, Paul M. 2016. 'Logical form', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). https://plato.stanford.edu/archives/fall2021/entries/logical-form/.

Pietroski, Paul M. 2018. *Conjoining Meanings: Semantics without Truth Values* (Oxford: Oxford University Press).

Pinker, Steven. 1991. 'Rules of language', *Science*, 253: 530–5.

Pinker, Steven. 2007. *The Stuff of Thought: Language as a Window into Human Nature* (London: Viking Penguin).

Planer, Ronald J., and Peter Godfrey-Smith. 2021. 'Communication and representation understood as sender–receiver coordination', *Mind & Language*, 36: 750–70.

Polich, John. 2007. 'Updating P.300: An integrative theory of P.3a and P3b', *Clinical Neurophysiology*, 118: 2128–48.

Popper, Karl R. 1972. 'Of clouds and clocks', in *Objective Knowledge* (Oxford: Clarendon Press).

Pounder, Zoë, Jane Jacob, Samuel Evans, Catherine Loveday, Alison F. Eardley, and Juha Silvanto. 2022. 'Only minimal differences between individuals with congenital aphantasia and those with typical imagery on neuropsychological tasks that involve imagery', *Cortex*, 148: 180–92.

Prinz, Jesse. 2002. *Furnishing the Mind* (Cambridge, MA: MIT Press).

Prior, Arthur. 1967. ‘The runabout inference ticket’, in P Strawson (ed.), *Philosophical Logic* (Oxford: Oxford University Press).

Pritzel, Alexander, Benigno Uria, Sriram Srinivasan, Adria Puigdomenech, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. 2017. ‘Neural episodic control’, *arXiv:1703.01988.*

Proust, Joëlle. 2010. ‘Metacognition’, *Philosophy Compass*, 5: 989–98.

Proust, Joëlle. 2012. ‘Metacognition and mindreading: One or two functions?’, in Michael J. Beran, Johannes L. Brandl, Josef Perner, and Joëlle Proust (eds.), *Foundations of Metacognition* (Oxford: Oxford University Press).

Proust, Joëlle. 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness* (Oxford: Oxford University Press).

Pulvermüller, Friedemann. 2013. ‘How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics’, *Trends in Cognitive Sciences*, 17: 458–70.

Pylyshyn, Zenon. 1989. ‘The role of location indexes in spatial perception: A sketch of the FINST spatial-index model’, *Cognition*, 32: 65–97.

Quilty-Dunn, Jake. 2016. ‘Iconicity and the format of perception’, *Journal of Consciousness Studies*, 23: 255–63.

Quilty-Dunn, Jake. 2020. ‘Perceptual pluralism’, *Noûs*, 54: 807–38.

Quilty-Dunn, Jake. 2021. ‘Polysemy and thought: Toward a generative theory of concepts’, *Mind & Language*, 36: 158–85.

Quilty-Dunn, Jake, and Eric Mandelbaum. 2019. ‘Non-inferential transitions’, in Timothy Chan and Anders Nes (eds.), *Inference and Consciousness* (New York: Routledge).

Quilty-Dunn, Jake, Nicolas Porot, and Eric Mandelbaum. 2023. ‘The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences’, *Behavioral and Brain Sciences*, 46: e261.

Quine, Willard Van Orman. 1968. ‘Ontological relativity’, *Journal of Philosophy*, 65: 185–212.

Quine, Willard Van Orman. 1969. ‘Natural kinds’, in Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel* (Dordrecht: D. Reidel).

Ramsey, William. 2007. *Representation Reconsidered* (Cambridge: Cambridge University Press).

Recanati, François. 2012. *Mental Files* (Oxford: Oxford University Press).

Redding, Gordon M., and Benjamin Wallace. 1997. ‘Prism adaptation during target pointing from visible and nonvisible starting locations’, *Journal of Motor Behavior*, 29: 119–30.

Rescorla, Michael. 2009. ‘Predication and cartographic representation’, *Synthese*, 169: 175–200.

Rescorla, Michael. 2012. ‘Are computational transitions sensitive to semantics?’, *Australasian Journal of Philosophy*, 90: 703–21.

Rescorla, Michael. 2014. ‘The causal relevance of content to computation’, *Philosophy and Phenomenological Research*, 88: 173–208.

Rescorla, Michael. 2024. ‘Neural implementation of (approximate) Bayesian inference’, in Tony Cheng, Ryoji Sato, and Jakob Hohwy (eds.), *Expected Experiences: The Predicative Mind in an Uncertain World* (New York: Routledge).

Reuter-Lorenz, Patricia A., and Alexandru D. Iordan. 2021. ‘Remembering over the short and long term: Empirical continuities and theoretical implications’, in Nelson Cowan, Candice C. Morey, and Moshe Naveh-Benjamin (eds.), *Working Memory: State of the Science* (New York: Oxford University Press).

Reverberi, Carlo, Doris Pischedda, Michele Burigo, and Paolo Cherubini. 2012. ‘Deduction without awareness’, *Acta Psychologica*, 139: 244–53.

Rouault, Marion, Maël Lebreton, and Mathias Pessiglione. 2023. ‘A shared brain system forming confidence judgment across cognitive domains’, *Cerebral Cortex*, 33: 1426–39.

Rushworth, Matthew F. S., Rogier B. Mars, and Christopher Summerfield. 2009. ‘General mechanisms for making decisions?’, *Current Opinion in Neurobiology*, 19: 75–83.

Russell, Bertrand. 1912. *The Problems of Philosophy* (New York: Henry Holt).

Ryoo, Michael S., Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. 2022. ‘Token Turing machines’, *arXiv:2211.09119v1.*

Sablé-Meyer, Mathias, Kevin Ellis, Josh Tenenbaum, and Stanislas Dehaene. 2022. 'A language of thought for the mental representation of geometric shapes', *Cognitive Psychology*, 139: 101527.

Sainsbury, Mark, and Michael Tye. 2007. *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts* (Oxford: Oxford University Press).

Samsonovich, Alexei V., and Giorgio A. Ascoli. 2005. 'A simple neural network model of the hippocampus suggesting its pathfinding role in episodic memory retrieval', *Learning & Memory*, 12: 193–208.

Samuels, Richard. 2010. 'Classical computationalism and the many problems of cognitive relevance', *Studies in History and Philosophy of Science Part A*, 41: 280–93.

Santoro, Adam, David Raposo, David G. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. 'A simple neural network module for relational reasoning.' In *31st conference on neural information processing systems (NIPS 2017)*, 4967–76. Long Beach, CA, USA.

Schneider, Wolfgang, and Elisabeth Löffler. 2016. 'The development of metacognitive knowledge in children and adolescents', in John Dunlosky and Sarah K. Tauber (eds.), *The Oxford Handbook of Metamemory* (New York: Oxford University Press).

Schubert, Lenhart. 2019. 'Computational Linguistics', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/computational-linguistics/.

Schuck, Nicolas W., and Yael Niv. 2019. 'Sequential replay of nonspatial task states in the human hippocampus', *Science*, 364: eaaw5181.

Schultz, Wolfram, Peter Dayan, and P. Read Montague. 1997. 'A neural substrate of prediction and reward', *Science*, 275: 1593–9.

Schwartenbeck, Philipp, Alon Baram, Yunzhe Liu, Shirley Mark, Timothy Muller, Raymond Dolan, Matthew Botvinick, Zeb Kurth-Nelson, and Timothy Behrens. 2023. 'Generative replay underlies compositional inference in the hippocampal-prefrontal circuit', *Cell*, 186: 4885–97, e14.

Schwartz, Bennett L. 1999. 'Sparkling at the end of the tongue: The etiology of tip-of-the-tongue phenomenology', *Psychonomic Bulletin & Review*, 6: 379–93.

Schwartz, Daniel L., and Tamara Black. 1999. 'Inferences through imagined actions: Knowing by simulated doing', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25: 116–36.

Schwarz, Norbert. 2015. 'Metacognition', in Eugene Borgida and John A. Bargh (eds.), *APA Handbook of Personality and Social Psychology, Vol. 1: Attitudes and Social Cognition* (Washington, DC: American Psychological Association).

Sellars, Wilfrid. 1953. 'Inference and meaning', *Mind*, 62: 313–38.

Sergent, Claire, Martina Corazzol, Ghislaine Labouret, François Stockart, Mark Wexler, Jean-Rémi King, Florent Meyniel, and Daniel Pressnitzer. 2021. 'Bifurcation in brain dynamics reveals a signature of conscious processing independent of report', *Nature Communications*, 12: 1149.

Seyfarth, Robert M., Dorothy L. Cheney, and Peter Marler. 1980. 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Animal Behaviour*, 28: 1070–94.

Shagrir, Oron. 2001. 'Content, computation and externalism', *Mind*, 110: 369–400.

Shagrir, Oron. 2012. 'Structural representations and the brain', *British Journal for the Philosophy of Science*, 63: 519–45.

Shanahan, Murray. 2016. 'The frame problem', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition). https://plato.stanford.edu/archives/spr2016/entries/frame-problem/.

Shanahan, Murray, and Bernard Baars. 2005. 'Applying global workspace theory to the frame problem', *Cognition*, 98: 157–76.

Shea, Nicholas. 2007. 'Content and its vehicles in connectionist systems', *Mind & Language*, 22: 246–69.

Shea, Nicholas. 2012. 'Genetic representation explains the cluster of innateness-related properties', *Mind & Language*, 27: 466–93.

Shea, Nicholas. 2013. 'Millikan's isomorphism requirement', in Dan Ryder, Justine Kingsbury, and Kenneth Williford (eds.), *Millikan and Her Critics* (Malden, MA: Wiley-Blackwell).

Shea, Nicholas. 2014a. 'Exploited isomorphism and structural representation', *Proceedings of the Aristotelian Society*, 64: 123–44.

Shea, Nicholas. 2014b. 'Neural signaling of probabilistic vectors', *Philosophy of Science*, 81: 902–13.

Shea, Nicholas. 2014c. 'Reward prediction error signals are meta-representational', *Noûs*, 48: 314–41.

Shea, Nicholas. 2015. 'Distinguishing top-down from bottom-up effects', in Dustin Stokes, Mohan Matthen, and Stephen Biggs (eds.), *Perception and Its Modalities* (New York: Oxford University Press).

Shea, Nicholas. 2016. 'Representational development need not be explicable-by-content', in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Cham: Springer).

Shea, Nicholas. 2018. *Representation in Cognitive Science* (Oxford: Oxford University Press).

Shea, Nicholas. 2022a. '*Representation in Cognitive Science* by Nicholas Shea: Reply by the author', *Studies in History and Philosophy of Science*, 92: 270–73.

Shea, Nicholas. 2022b. 'Concepts as plug & play devices', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378: 20210353.

Shea, Nicholas. 2023a. 'Millikan's consistency testers and the cultural evolution of concepts', *Evolutionary Linguistic Theory*, 5: 79–101.

Shea, Nicholas. 2023b. 'Moving beyond content-specific computation in artificial neural networks', *Mind & Language*, 38: 156–77.

Shea, Nicholas. 2023c. 'Organized representations forming a computationally useful processing structure', *Synthese*, 202: 1–20.

Shea, Nicholas. Forthcoming. 'Metacognition of inferential transitions', *Journal of Philosophy*.

Shea, Nicholas, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D. Frith. 2014. 'Supra-personal cognitive control and metacognition', *Trends in Cognitive Sciences*, 18: 186–93.

Shea, Nicholas, and Chris D Frith. 2019. 'The global workspace needs metacognition', *Trends in Cognitive Sciences*, 23: 560–71.

Shepard, Roger N., and Jacqueline Metzler. 1971. 'Mental rotation of three-dimensional objects', *Science*, 171: 701–3.

Shepherd, Joshua. 2023. 'Disappearing agents, mental action, rational glue', in Michael Brent and Lisa Miracci Titus (eds.), *Mental Action and the Conscious Mind* (Abingdon: Routledge).

Silliman, Daniel C., and Kenneth J. Kurtz. 2019. 'Evidence of analogical re-representation from a change detection task', *Cognition*, 190: 128–36.

Sloutsky, Vladimir M. 2010. 'From perceptual categories to concepts: What develops?', *Cognitive Science*, 34: 1244–86.

Smith, Edward E., and Murray Grossman. 2008. 'Multiple systems of category learning', *Neuroscience & Biobehavioral Reviews*, 32: 249–64.

Smith, Linda B., and Larissa K. Samuelson. 1997. 'Perceiving and remembering: Category stability, variability and development', in Koen Lamberts and David Shanks (eds.), *Knowledge, Concepts, and Categories* (Hove: Psychology Press).

Smolensky, Paul. 1988. 'On the proper treatment of connectionism', *Behavioral and Brain Sciences*, 11: 1–74.

Smolensky, Paul. 1995. 'Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture', in Cynthia Macdonald and Graham Macdonald (eds.), *Connectionism: Debates on Psychological Explanation, Volume 2* (Oxford: Blackwell).

Smolensky, Paul, and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar, Volume 1: Cognitive Architecture* (Cambridge, MA: MIT Press).

Søgaard, Anders. 2023. 'Grounding the vector space of an octopus: Word meaning from raw text', *Minds and Machines*, 33: 33–54.

Solomon, Karen O., and Lawrence W. Barsalou. 2004. 'Perceptual simulation in property verification', *Memory & Cognition*, 32: 244–59.

Sosa, Ernest. 1985. ‘Knowledge and intellectual virtue’, *The Monist*, 68: 226–45.

Sperber, Dan, Fabrice Clement, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. ‘Epistemic vigilance’, *Mind & Language*, 25: 359–93.

Spiro, Rand J., Walter P. Vispoel, John G. Schmitz, Ala Samarapungavan, and A. E. Boerger. 1987. ‘Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains’, in Bruce K. Britton and Shawn M. Glynn (eds.), *Executive Control Processes in Reading* (Hillsdale, NJ: Erlbaum).

Sprevak, Mark. 2005. ‘The frame problem and the treatment of prediction’, in Lorenzo Magnani and Riccardo Dossena (eds.), *Computing, Philosophy and Cognition* (London: King’s College Publications).

Stahlberg, Felix. 2020. ‘Neural machine translation: A review’, *Journal of Artificial Intelligence Research*, 69: 343–418.

Stanovich, Keith E. 2009. ‘Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?’, in Jonathan St. B. T. Evans and Keith Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (New York: Oxford University Press).

Stich, Stephen P. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief* (Cambridge, MA: MIT Press).

Strawson, Peter. 1962. ‘Freedom and resentment’, in Gary Watson (ed.), *Proceedings of the British Academy, Volume 48* (Oxford: Oxford University Press).

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation* (Cambridge, MA: Harvard University Press).

Strevens, Michael. 2010. ‘Varieties of understanding’, paper presented at the Pacific Division Meeting of the American Philosophical Association.

Strevens, Michael. 2019. *Thinking Off Your Feet: How Empirical Psychology Vindicates Armchair Philosophy* (Cambridge, MA: Harvard University Press).

Sun, Lin, and Sanjay G. Manohar. 2023. ‘Syntax through rapid synaptic changes’, *bioRxiv*: 2023.12.21.572018.

Sutton, Richard S., and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press).

Swoyer, Chris. 1991. ‘Structural representation and surrogative reasoning’, *Synthese*, 87: 449–508.

Tanaka, Kazumasa Z., and Thomas J. McHugh. 2018. ‘The hippocampal engram as a memory index’, *Journal of Experimental Neuroscience*, 12: 1179069518815942.

Tang, Jerry, Amanda LeBel, Shailee Jain, and Alexander G. Hu. 2023. ‘Semantic reconstruction of continuous language from non-invasive brain recordings’, *Nature Neuroscience*, 26: 858–66.

Teyler, Timothy J., and Jerry W. Rudy. 2007. ‘The hippocampal indexing theory and episodic memory: Updating the index’, *Hippocampus*, 17: 1158–69.

Thompson, Valerie A., Jonathan St. B. T. Evans, and Jamie I. D. Campbell. 2013. ‘Matching bias on the selection task: It’s fast and feels good’, *Thinking & Reasoning*, 19: 431–52.

Thompson, Valerie A., and Stephen C. Johnson. 2014. ‘Conflict, metacognition, and analytic thinking’, *Thinking & Reasoning*, 20: 215–44.

Thompson, Valerie A., Jamie A. Prowse Turner, and Gordon Pennycook. 2011. ‘Intuition, reason, and metacognition’, *Cognitive Psychology*, 63: 107–40.

Thorne, Sapphira R., Jake Quilty-Dunn, Joulia Smortchkova, Nicholas Shea, and James A. Hampton. 2021. ‘Concept appraisal’, *Cognitive Science*, 45: e12978.

Thorne, Sapphira R., Joulia Smortchkova, Jake Quilty-Dunn, Nicholas Shea, and James A. Hampton. 2022. ‘Is concept appraisal modulated by procedural or declarative manipulations?’, *Frontiers in Psychology*, 13: 774629.

Thura, David, Jean-François Cabana, Albert Feghaly, and Paul Cisek. 2022. ‘Integrated neural dynamics of sensorimotor decisions and actions’, *PLOS Biology*, 20: e3001861.

Todorov, Alexander, Christopher Y. Olivola, Ron Dotsch, and Peter Mende-Siedlecki. 2015. ‘Social attributions from faces: Determinants, consequences, accuracy, and functional significance’, *Annual Review of Psychology*, 66: 519–45.

Tomasello, Michael. 2022. *The Evolution of Agency* (Cambridge, MA: MIT Press).

Travis, Charles. 1997. ‘Pragmatics’, in Bob Hale and Crispin Wright (eds.), *Companion to the Philosophy of Language* (Oxford: Blackwell).

Traylor, Aaron, Roman Feiman, and Ellie Pavlick. 2021. 'AND does not mean OR: Using formal languages to study language models' representations', *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*: 158–67.

Treisman, Anne. 1996. 'The binding problem', *Current Opinion in Neurobiology*, 6: 171–8.

Treisman, Anne, and Janet Souther. 1985. 'Search asymmetry: A diagnostic for preattentive processing of separable features', *Journal of Experimental Psychology: General*, 114: 285–310.

Tversky, Barbara, and Elizabeth J. Marsh. 2000. 'Biased retellings of events yield biased memories', *Cognitive Psychology*, 40: 1–38.

Usher, Marius, Zohar Russo, Mark Weyers, Ran Brauner, and Dan Zakay. 2011. 'The impact of the mode of thought in complex decisions: Intuitive decisions are better', *Frontiers in Psychology*, 2: 37.

Vicente, Agustín, and Fernando Martínez Manrique. 2016. 'The big concepts paper: A defence of hybridism', *The British Journal for the Philosophy of Science*, 67: 59–88.

Wang, Mengni, David J. Foster, and Brad E. Pfeiffer. 2020. 'Alternating sequences of future and past behavior encoded within hippocampal theta oscillations', *Science*, 370: 247–50.

Weiskopf, Daniel A. 2009a. 'Atomism, pluralism, and conceptual content', *Philosophy and Phenomenological Research*, 74: 131–63.

Weiskopf, Daniel A. 2009b. 'The plurality of concepts', *Synthese*, 169: 145–73.

Wellman, Henry M., David Cross, and Julanne Watson. 2001. 'Meta-analysis of theory-of-mind development: The truth about false belief', *Child Development*, 72: 655–84.

Whitney, Paul, Timothy McKay, George Kellas, and William A. Emerson. 1985. 'Semantic activation of noun concepts in context', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11: 126–35.

Whittington, James C. R., Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. 2020. 'The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation', *Cell*, 183: 1249–63, e23.

Wolfe, Jeremy M., Keith R. Kluender, Dennis M. Levi, Linda M. Bartoshuk, Rachel S. Herz, Roberta L. Klatzky, and D. M. Merfeld. 2018. *Sensation & Perception* (Sunderland, MA: Sinauer Associates).

Wolpert, Daniel M., Jörn Diedrichsen, and J. Randall Flanagan. 2011. 'Principles of sensorimotor learning', *Nature Reviews Neuroscience*, 12: 739–51.

Wright, Crispin. 2000. 'Cogency and question-begging: Some reflections on McKinsey's paradox and Putnam's proof', *Philosophical Issues*, 10: 140–63.

Xu, Yingjin, and Pei Wang. 2012. 'The frame problem, the relevance problem, and a package solution to both', *Synthese*, 187: 43–72.

Yablo, Stephen. 2002. 'Coulda, woulda, shoulda', in Tamar S. Gendler and John Hawthorne (eds.), *Conceivability and Possibility* (Oxford: Oxford University Press).

Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. 'Performance-optimized hierarchical models predict neural responses in higher visual cortex', *Proceedings of the National Academy of Sciences of the United States of America*, 111: 8619–24.

Yang, Dujia, Xiaowei Qin, Xiaodong Xu, Chensheng Li, and Guo Wei. 2020. 'Sample efficient reinforcement learning method via high efficient episodic memory', *IEEE Access*, 8: 129274–84.

Yousefi-Azar, Mahmood, and Len Hamey. 2017. 'Text summarization using unsupervised deep learning', *Expert Systems with Applications*, 68: 93–105.

Index

Since the index has been created to work across multiple formats, indexed terms for which a page range is given (e.g., 52–53, 66–70, etc.) may occasionally appear only on some, but not all of the pages within the range.