MICHAEL SCHEFCZYK &
CHRISTOPH SCHMIDT-PETRI
Editors

Utility, Progress, and Technology

Proceedings of the 15th Conference of the International Society for Utilitarian Studies

KIT Scientific Publishing

Michael Schefczyk & Christoph Schmidt-Petri (eds.)

Utility, Progress, and Technology

Proceedings of the 15th Conference of the
International Society for Utilitarian Studies

Utility, Progress, and Technology

Proceedings of the 15th Conference of the
International Society for Utilitarian Studies

edited by
Michael Schefczyk & Christoph Schmidt-Petri

KIT Scientific Publishing

Cover design by Nico Brähler

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

www.ksp.kit.edu

Print on Demand 2021 – Gedruckt auf FSC-zertifiziertem Papier

ISBN 978-3-7315-1108-3
DOI 10.5445/KSP/1000134479

Preface

On behalf of the International Society for Utilitarian Studies, I would like to express our deep gratitude to Michael Schefczyk and Christoph Schmidt-Petri and their team at Karlsruhe for organizing such a wonderful conference, where a degree of humour was never very far from the surface. Michael and I talked about the organization of the conference while attending Malik Bozzo-Rey's ISUS conference at Lille, and while we did not get quite so far as implementing our idea of accommodating participants in a panopticon-style tent village, we did get to enjoy a panopticon dinner, where unseen ears listened to our conversations and noted bad behaviour (as one might expect in a company well-populated by moral philosophers, there was no more than a modicum of that)—though the food was probably somewhat more appetising than that which Bentham would have served to his inmates. The panopticon dinner took place at the famous Karlsruhe ZKM/Centre for Art and Media, where we were able to enjoy a tour of the fascinating 'Open Codes: Living in Digital Worlds' exhibition—Bentham would have loved all that data.

Karlsruhe was the fifteenth conference in the series since they began with a very modest affair, attended, as I recall, by about 30 person, at University College London in 1987. Subsequent conferences (with regnal numbers) have been as follows:

ISUS II (1989):	King's College, Cambridge
ISUS III (1992):	University of Western Ontario
ISUS IV (1994):	Chuo University, Tokyo
ISUS V (1997):	Tulane University, New Orleans
ISUS VI (2000):	Wake Forest University, North Carolina
ISUS VII (2003):	Lisbon, Portugal
ISUS VIII (2005):	Dartmouth College, New Hampshire
ISUS IX (2006):	University College London: The John Stuart Mill Bicentenary
ISUS X (2008):	University of California at Berkeley

ISUS XI (2011):	Lucca, Italy
ISUS XII (2012):	New York University
ISUS XIII (2014):	Yokohama National University
ISUS XIV (2016):	Catholic University of Lille

All have been memorable in various ways (am I the only person to have attended all of them?), but Karlsruhe was the first that has included a ride on a narrow-gauge railway (future conference organizers please note!). The train took our group around the grounds of Karlsruhe castle, which themselves are arranged on a semi-panoptic plan.

I should add that experiencing the hospitality at the Karlsruhe Institute of Technology made me a little more aware of why the good folk of Germany are renowned for their organizational efficiency.

You might be forgiven for thinking that the conference was merely an excuse to have a good time and to meet up with old and the opportunity to make new friends, and you would not be wrong! However, I include in the idea of having a good time, the pleasure of hearing a selection of excellent lectures and papers, discussing ideas with people who have something to say, and experiencing, if not contributing to, the republic of scholarship. The conference was characterized by the usual mixture of philosophical and historical, theoretical and applied, perspectives that makes ISUS so special. The current collection of essays reflects that mixture and is a worthy tribute to the vibrancy and vitality of utilitarian studies.

The conference scheduled for 2020 at the University of Illinois at Chicago had to be cancelled, due to circumstances beyond our control, but I fervently hope that we will meet again for ISUS XVI at Rome in 2022.

Philip Schofield

Honorary Secretary and Treasurer
International Society for Utilitarian Studies
London, 21 January 2021

Acknowledgements

Michael Schefczyk and Christoph Schmidt-Petri

Organising the 15th conference of the International Society for Utilitarian Studies has been a great pleasure for us. We are honoured the ISUS committee trusted us with this project and sincerely thank all participants for joining us in Karlsruhe from all over the world.

Legend has it that the founder of Karlsruhe, Margrave of Baden-Durlach Karl Wilhelm, fell asleep in the forest and dreamt of a palace. He had the palace of Karlsruhe built in 1715, in that very forest, as the epicentre of a new city. The absolutist idea reflected in the fan-shaped map of Karlsruhe, with the ruler overlooking its streets from the tower's castle, of course blended in well with the visionary plan of a panopticon-style tent camp (as mentioned in Philip Schofield's preface), which subsequently gave way to more viable aspects of the conference taking equally peculiar shapes.

Our conference theme, ‚Utility, Progress, and Technology' was intended to emphasise that any reflection on technology necessarily requires normative dimensions that even the best scientific education or training cannot provide. Does some innovation, useful as it may seem, actually constitute progress? Does it increase human happiness? We hope ISUS 2018 has helped improving thinking on these (and other) philosophical problems.

For financial support we thank the German Research Foundation (Deutsche Forschungsgemeinschaft), the Karlsruhe Institute of Technology, its Institute for Technology Futures, the KIT Fördergesellschaft, the Kursbuch Kulturstiftung, as well as the International Society for Utilitarian Studies and the German Society for Utilitarian Studies.

The success of the conference required the commitment of many student assistants and helping hands. In particular, we thank Michael W. Schmidt, who coordinated the core organising group consisting of Marie-Claire Haag (née Baur), Dorothee Bleisch, Nico Brähler, Max Hagelstein, and Sina Schmitt – thank you all! During the event, we could not have done without the many helping hands: thanks to Alessa Auerswald, Sarah Schwarz (née Boussard), Jonas Bühler, Johanna Gramacho Narloch, Jean Gras, Simeon Imhoff, Mona Meixner, Lilly Osburg, Andie Rothenhäusler, Nora Steinhäuser, Constantin Weeber, Kai Wieland, and Elizaveta Zaidman.

Contents

Utilitarianism and Responsibility for the Future[1]

Dieter Birnbacher, University of Düsseldorf, Germany

Abstract

Differently from scientific contexts utilitarianism continues to be a stumbling-block in many public debates, partly because of misunderstandings, partly because of conflicts with widespread moral convictions. These concern both its axiology and its theory of normativity. On the other hand, there are several context of ethical and public discussion in which characteristic elements of utilitarianism and widely shared normative position come remarkably close, such as the growing recognition of the moral status of nonhuman animals and the recognition of the responsibility for a sustainable use of natural resources. Historically, representatives of utilitarianism had an important share in driving this development. Furthermore, there is a remarkable affinity between utilitarianism and the „principle of responsibility" highlighted, among others, by Hans Jonas. First, there is an affinity between the concept of a prospective responsibility and the utilitarian conception of responsibility as directed at future events and states rather than at future actions and omissions. Another affinity is the utilitarian principle of extending responsibility to all foreseeable consequences instead of, as the theory of double effect has it, restricting responsibility to intended consequences. Finally, utilitarianism is more than its rivals able to satisfy the demands of universalizability implied by the moral nature of prospective responsibility by making the value of subjective well-being its one and only intrinsic value. There does not seem to be any other value on which the same degree of *consensus gentium* can be expected.

I Utilitarianism – Between Academia and the Public

Every practice-oriented ethicist knows the gap that from time to time requires an intellectual balancing act between the culture of discussion in the academic world and that of the public sphere: on the one hand a disciplinary expert, on the other a moralist. Many ethical theories discussed objectively and dispassionately in philosophical or economic seminars are met by the public with rejection or outrage, for example, when they conflict with common sense notions of everyday morality or with fundamental political norms.

[1] Translated from German by Paul Lauer.

Utilitarianism is the ethical doctrine that this characterization best fits. In numerous academic debates, and in particular in those about the big problems concerning our future, utilitarian premises are the more or less unquestioned starting point of the discussion. This is, for example, the case when in academic discussions about climate ethics the question is asked which strategies seem most promising to limit climate change and its consequences for all of those affected directly and indirectly. Utilitarianism is already, as it were, part of the question. The debate is less about whether utilitarianism or another ethical theory should be the basis of a search for an answer to the problem, but more about which *version* of utilitarianism is best able to do justice to the problem.

Indeed some of the best known models for a successful climate politics are based on different versions of a utilitarian ethics. A positive version is found in the so-called *DICE Model*, a dynamic integrated model of climate and the economy, proposed by the American economist William Nordhaus (2013). The goal of this model is to determine – by means of simulating the consequences of different climate policy strategies – how it is possible to maximize the total utility of everyone in the world affected by climate change. This model makes use of a not entirely unproblematic methodological simplification, namely, the *monetarization* of all crucial dimensions of utility. This involves the assessment of all of the different positive and negative impacts of climate change and efforts to combat it – including climate-induced migration and the erosion of social and political institutions – by means of comparable monetary units. A comparable concept, which uses a 'utility currency' instead of monetary units to evaluate the commensurability of many dimensions of positive and negative utilities, has been developed in the philosophy of Christoph Lumer (2002). By contrast, the so-called Stern Review – the report by a commission evaluating the economic consequences of climate change led by the former chief economist of the World Bank – makes use of a *negative* version of utilitarianism (Stern, 2006). Only the negative consequences of, on the one hand, a laisser-faire climate policy and, on the other, an ambitious future-oriented climate policy are compared with each other. Only those climate strategies that minimize long-term costs are worthy of consideration. Based on detailed calculations, the Stern Review calls for a stronger and more rigorous climate policy to reduce emissions and provide more comprehensive support for adapting to climate change in affected regions. Yet another version of utilitarianism is the basis of Bernward Gesang's concept of climate ethics. Gesang (2011, 43) extends the classic utilitarian principles of long-term and universal maximization of utility by means of a catastrophe avoidance principle, which prohibits initiating processes with potentially catastrophic consequences, regardless of whether such an avoidance strategy is economically viable over the long run.

All of these models have considerable differences in their conclusions – which are however only partially due to the fact that they are based on different versions of a utilitarian ethics.

It is also due to their providing different answers to the open questions of this ethics, for example, the extent to which a utility calculation 'discounts' the value of benefits and harms lying in the distant future in comparison to their current value. That the *Stern Report* model arrives at completely different results than the Nordhaus model is largely due to the fact that it 'discounts' the future benefits at a very much higher rate than Nordhaus does. Unlike Nordhaus, the authors of the Stern Report do not consider it justified to value the benefits and harms of future generations at a considerably lower rate than the benefits and harms of currently living generations of human beings.

In contrast to academic contexts, utilitarianism is in many public debates still a stumbling block – partially due to misunderstandings such as utilitarianism assessing consequences on their objective instead of their expected and so more or less probable consequences, but also due to a firm rejection of some of the characteristic norms of utilitarianism when they are strongly and undeniably opposed to popular convictions. Two examples are the value theory of utilitarianism, which has only a single value, that of a positive state of consciousness – however this might be understood – and the normative theory that prioritizes those actions that under given conditions will in all probability maximize the gains as well as minimize the losses in subjective well-being for all those affected by the action.

Utilitarianism, as a value theory, is often objected to because of the perception that it does not give any *intrinsic* but only a *derived* value to life (the life of human beings but also that of higher forms of animal life). Not life, but the quality of life has an intrinsic value according to the hedonistic value teachings of classic utilitarianism. Solely as a condition of the quality of life is life worth being preserved, protected, initiated, and made possible. The consequence of this is that utilitarians take up liberal positions in many controversial issues in bioethics – such as abortion, reproductive medicine or euthanasia – which encounter resistance in Germany, especially by the Catholic Church but also in political parties influenced by Christian morality.

Utilitarianism, as a normative theory, is objected to for two of its characteristic tendencies. First is the tendency not to recognize any absolute moral boundaries to what is permitted (such as, in law the concept of *human dignity*) and instead to consider all value dimensions relevant for a particular decision as subject to calculation; second is the tendency to abolish the distinction in the moral evaluation of acting and not acting, of active doing and the passive allowing of an action. This distinction is firmly established in everyday morality as well as in the law, with the consequence that many people strongly reject *active* euthanasia just as they approve *passive* euthanasia.

This image of a division between academia and the public in its perception of utilitarianism is however one-sided. It neglects the fact there are undoubtedly contexts in academia and in the public where there are remarkable convergences between characteristically utilitarian positions and those normative positions currently shared at least among the *intelligentsia*.

One of which is the increasing recognition of the *moral status of higher forms of animal life*, another is the *increasing recognition of a responsibility for the preservation of the basis for human life in future generations*. Both are facets of what Wilhelm Kamlah called the "unbordering of responsibility" (1973, 105), which has been happening in our culture since the Enlightenment and has been most fully developed in the classics of utilitarianism.

An 'unbordering of responsibility' beyond the human species was first broached in the preface to Rousseau's (1973, 72-73) essay on inequality in 1755, in which he speaks of animals also partaking in natural law. In 1785, for the first time, the recognition of animal rights was expressly called for in Wilhelm Dietler's treatise with the title *Justice Towards Animals*. It did not go so far as to prohibit the killing of animals for food or safety. But it should only be allowed for animals to be killed in 'the fastest, most gentle and least painful way'. Nor were people allowed to hunt animals solely in the pursuit of pleasure or to abuse their pets (Dietler 1997, 26). In the revolutionary year of 1789, Bentham's (1948, 311) major work was published, *The Principles of Morals and Legislation*, containing the argument that morality was not about whether a creature can reason or talk but solely about whether they can *suffer*.

This inclusion of creatures capable of suffering, beyond the borders of species, can be found in modern laws governing animal protection, which without exception identify such animals as deserving of protection and aim to avoid, prevent and alleviate animal suffering. The reason that in Germany cephalopods such as octopuses are now protected is largely due to the supposition that their sensory capabilities justify their being included, alongside vertebrate animals, among non-human creatures capable of suffering. It is entirely another matter whether this generous standard defining which animals are capable of suffering will be applied in practice. That there is a 'lack of enforcement' is obvious.

II The Sentimental Future of Utilitarianism

The second area in which everyday moral thinking and utilitarian principles are in contact is in the recognition of our *responsibility for the long-term preservation of the biological and*

civilizational basis of human life and for the successive humanization of living conditions on a global scale. Growing prosperity increasing social security has lead, in wealthy countries, to a willingness to recognize and take on 'responsibility at a distance', and in three different ways: beyond the borders of species to a responsibility for the well-being of sentient animals and the preservation of the *natural world*; beyond the borders of one's own group to an extension of solidarity with a *global community* and so overcoming the evolutionary relict of tribalism and the limitation of solidarity to members of one's own family, kinship, clan, or tribe; and finally beyond the borders of the present to an expansion of responsibility in the direction of an endless future.

Advocates of a utilitarian ethics were crucially involved in the *expanding circle* of human responsibility. From Bentham to John Stuart Mill and Henry Sidgwick to Peter Singer, there is an unbroken line of thinkers who, going beyond the human welfare, took up the cause of the welfare of animals and in the case of John Stuart Mill the preservation of natural diversity. The young Peter Singer spoke so powerfully about the problem of world hunger in 1972 in an article considered sensational at the time, "Famine, Affluence, and Morality", that a number of other utilitarian authors hurriedly followed in its wake. And the perspective of *'in the long run'*, a perspective that when taking decisions or making strategies primarily concerned with the present we should also, to the extent possible, consider their consequences in the future is a hallmark of the writings of John Stuart Mill on economics, politics and social policy.

In Mill we find – in almost all of his writings in which he took on the role of a public moralist – him expressing a sentimental belief in a better future. Mill's style of writing about people and about what is good and right for them is not only confined to the people of his time but to humanity as a process, or as he would put it, *man as a progressive being* (CW XVIII, 224). Without doubt Mill understood utilitarianism from a long-term perspective and against a background of an ideal of civilization as increasingly intergenerational and global. Humanity is for Mill – I would like to use a category of Ernst Bloch here – a latency, a *dynamis* and not an *energeia*. It is a continuing, if also very gradual and always beset by reversals, transition to a successively higher level of perfection, a continuous emancipation from the limitations of an earlier state. It was not the social utility of the here and now that was decisive for him but the importance of human action for the future of humanity. The dominant medium of this progress for Mill – although in England he was a contemporary witness of the first industrial revolution – was not technology and economic productivity but education.

Given this viewpoint it is hardly surprising that Mill was an avowed opponent of any anthropology that sought to determine human nature or the essence of being human. For Mill there is neither a nature nor a God that determines the future of humankind. Humans are

instead literally an 'animal that is not determined', one that is free to set their own tasks for themselves.

The moralist Mill goes so far as to not only encourage his fellow human beings to transcend the borders drawn for them by anthropologists but obligated them to do so. In his essay Nature he wrote:

> [T]he duty of man is the same in respect to his own nature as in respect to the nature of all other things, namely not to follow but to amend it. (CW X, 397)

The most notorious example of a false determination of this type was for Mill the Aristotelian characterisation of women as an 'inherently' inferior being to men. Mill was convinced that women were, not only in his time but in the past as well, prevented from achieving their potential by mistaken norms and conventions. The seeming inferiority of women was, for Mill, nothing other than a projection of social coercion in an image of nature that legitimized this relationship: the seeming inferiority of women as a product of the *subjection of women*.

It was however characteristic for Mill that the inequality of the sexes, which is still prevalent around the world, was not only an example of a false philosophy but also a practical political challenge. Mill had the good fortune that when he was a member of parliament the electoral franchise was debated. This gave him the opportunity to hold a number of speeches on women's right to vote. Mill believed that – as he wrote in his autobiography – his advocacy of female suffrage was "by far the most important, perhaps the only really important public service I performed in the capacity of a Member of Parliament" (CW I, 285) . He had a symbolic victory. His amendment to enfranchise women – it was to replace the word 'man' with the word 'person' in the Reform Bill – received an impressive 73 votes, a third of the members of parliament present. However, women did not receive the right to vote in the United Kingdom until 1928.

Mill's speech to his fellow parliamentarians still resonates today. To demonstrate the backwardness of the occupational regulations of the time, he gave the example of a young woman who wanted to become a physician and – thanks only to a legal loophole – was able to become a pharmacist. As soon as the Society of Apothecaries noticed that a woman had slipped into their ranks, they quickly passed a statute prohibiting women from becoming members. Mill's sarcastic comment was,

> No sooner do women show themselves capable of competing with men in any career, than that career, if it be lucrative or honourable, is closed to them. (CW XXVIII, 160)

Mill's advocacy for the emancipation of women has more to do with his sentimental view of the future than it may appear at first sight. Mill's motive was not that English women were longing for emancipation. He knew and admitted that there was minimal interest in political equality among the great majority of English women and that gaining the right to vote was by no means a priority for them. Mill's motive seems to have been another: not to forgo the promise that women's participation in leading positions in politics, education, business and science had for the future of society. Self-realization, as we would say today, was for Mill, as a student of Humboldt, not only an individual but also – and especially – a collective goal.

Many seeming inconsistencies and tensions in Mill's writings are resolved if we place them next to the sentimentality of his view of the future and make this a key to interpreting him. His glorification of individualism and his encouragement of stubbornness – as expressed in *On Liberty* – are often seen as contradictions to his faith in utilitarianism. Indeed, Mill does seem to argue in long passages that the development of the productive potential of the individual is an end in itself and not solely a means to increasing social welfare. Even admitting that Mill's own biography may have been a cause of his showing a degree of sympathy with social outsiders – he lived for many years with a married woman, a relationship condemned in the Victorian era, and was areligious as few of his contemporaries were – there seems to be more to his argumentation, namely that it is outsiders who contribute through their intellectual, social, scientific and economic innovations to improving the lives of people over the long run. Liberty is an essential condition for the unlocking of the creative potential that will advance humanity. Not harmony, adaptation, or complacent satisfaction but instead restlessness, criticism and dissatisfaction are the yeast of progress. He was certain, as he once wrote, that

> [N]othing is more certain than that improvement in human affairs is wholly the work of the uncontented characters. (CW XIX, 407)

Liberty is not, in this utilitarian perspective, a luxury. On the one hand, with increasing affluence and education the need for individual liberty also grows. On the other, only independent thought is able to bring forth future-oriented ideas and work that lead to social progress. One of Mill's fundamental convictions is that the active character type has advanced humankind to a much greater extent than the passive – even if he is not always popular and is sometimes seen as rival and a threat. This makes competition vital to progress, all the more so when it is a competition of plural opinions, perspectives and ways of living.

Mill's ever-present references to the future may also help explain his seemingly arbitrary and eclectic use of the Platonic theory of the three parts of the soul to explain the sources of pleasure in Chapter 2 of *Utilitarianism*: different types of pleasure have different effects on an individual's personality and motivation. Someone who enjoys science or technology – or music, art or theatre – generally also gives pleasure to others or makes their lives easier – in contrast to someone who only finds enjoyment in more ephemeral sensual pleasures.

Also in his political economy, Mill's arguments that certain institutional arrangements hinder 'social justice' are never solely about the present but invariably also about the future. They are about the functions an institution has for securing universal social welfare over the long term or, given social reform, could secure. What is important is that the reform of economic institutions is, as we would now say, 'sustainable' for those affected, for the 'stakeholders'.

Mill's arguments for a radical taxation of inheritances are a masterpiece of sophistry. In Mill's day inheritance was a major condition for the accumulation of wealth and political influence in the hands of an unproductive landed gentry. In our time inheritance is a major factor in the maintenance and intensification of economic inequality. Large inheritances, like today, were bequeathed to individuals belonging to the same class as the person giving the inheritance. These people were rarely reliant on the inheritance for their prosperity. Mill made use of a daring conceptual construction to argue for the legal restriction of the inheritance of property or even its entire prohibition. Although he had to allow that the concept of property encompassed the right of the property owner to pass on property as he thought best, this did not mean that society had to recognize a corresponding right on the part of the heirs to receive the inheritance, or to receive it in full. While the liberty to *bequaethe an inheritance* is irrevocably tied to the institution of property, the freedom to *receive an inheritance* is subject to moral restrictions:

> The guarantee to them of the fruits of the labour and abstinence of others, transmitted to them without any merit or exertion of their own, is not of the essence of the institution, but a mere incidental consequence, which, when it reaches a certain height, does not promote, but conflicts with, the ends which render private property legitimate. (CW II, 208)

This opens the way for the radical (in Mill's day) proposal to increase the taxation of inheritances of property to a level that would bring about fundamental changes in its distribution.

It is clear that Mill's sentimentality about the future and his appeals to the responsibility to promote – and not to impede or hinder – progress for future generations is not specific to utilitarianism. It is a commonplace in Enlightenment philosophy. In the 18th and 19th centuries this meant the rapid accumulation of theoretical and practical knowledge in the wake

of the liberation of thought from the dogmatic bonds of religion and despotism. What is important is that as the presence of the future has a growing importance in philosophical thought an 'ethics at a distance' takes shape, one that encourage the current generation to reflect on the long-term consequences of their actions and inactions and to exercise as much wisdom in providing for future generations as an individual does for his own personal future.

A temporal 'ethics at a distance' was already proposed, in a certain sense, by Kant. In his philosophy of history he advanced the proposition that the so-called 'pure' time preference – that is the preference for a near rather than a distant future regardless of all other factors – is not only not an essential human characteristic (as Spinoza claimed) it is foreign to his nature:

> Moreover, human nature is so constituted that we cannot be indifferent to the most remote epoch our race may come to, if only we may expect it with certainty. (Kant 1902, 27)

This statement is however at most acceptable as an ethical ideal. Empirical anthropology has shown that, without exception, the value of future events are 'discounted' whatever the certainty of their occurrence may be. The discount rate is not a linear one; it is a hyperbolic function. Whether a future good or harm occurs in one hundred or one thousand years does not have a large effect on how they are valued. There is however a great difference in whether it occurs in ten or one hundred years.

III Three Affinities between Prospective Responsibility and a Utilitarian Ethics

Utilitarianism is a paradigmatic example of a consequentialist ethics that measures the moral correctness or wrongness of something – actions and strategies as well as action guidelines, moral attitudes and moral emotions – solely according to the value of the expected consequences. This makes the justification of moral evaluations principally oriented toward the future and not the present or the past. This meant that the only justification for state punishment acceptable to utilitarians such as Bentham, Beccaria and Mill were oriented towards the future, that is the expected consequences of this practice for society as a whole. State punishment at the time was, for a number of different reasons, almost completely unable to satisfy this condition.

This orientation of justifying and criticizing moral institutions towards their consequences opens it to an empirical grounding. This aspect of consequential ethics is, however, both a blessing and a curse. Moral judgements, on the one hand, are unable to rely on intuitions, religious convictions and moral coercion. On the other, they are – to the extent that estimates or evaluations of consequences are uncertain – also uncertain.

In this context I would like to draw attention to another characteristic of utilitarian, and with it all other consequentialist ethics: the fact that the future consequences decisive for moral evaluation are expected positive or negative future events or states. The value of an action is not to be found in future actions but in evaluations of future *events* and *states*.

This means that a consequentialist ethics touches on central elements of the concept of *prospective* responsibility – in contrast to *retrospective* responsibility for actions lying in the past. The *first affinity* can be expressed as follows: The primary meaning of assuming or accepting responsibility directed towards the future is (though not exclusively) ensuring that certain positive or negative *events* will or will not occur and that certain *states* will or will not come into existence. The meaning and purpose of attributing or assuming responsibility is primarily the generation of certain goods and the avoidance of certain harms, not the execution or non-execution of certain actions. We speak of 'responsibility' not when we expect certain actions from someone bearing responsibility but when we oblige him to bring about certain events or states, without however specifying which actions will bring them about. In contrast to specific norms of behaviour a person who bears a responsibility is not obliged to act in a specific way but more generally to reach a specific goal with purposeful actions. To say that someone is responsible for one's children, a device, world peace or the reduction of greenhouse gases does not oblige that person to specific actions directly implied in the responsibility itself but to actions (including non-actions) that bring about or contribute to the bringing about of specified or implied objectives – whether the well-being of a child, the functionality or safety of the device, the non-occurrence of armed conflict or a reduction in emissions.

This is the justification of the concept of an 'ethics of responsibility' as opposed to what Max Weber called an 'ethics of conviction'. While a deontological ethics typically proscribes or prescribes certain actions regardless of the purposes or intentions for which they were carried out and a virtue ethics the formation and exercise of certain moral behavioural dispositions, an 'ethics of responsibility' largely leaves open how the goals a person is responsible for achieving should be reached and which behavioural dispositions and attitudes are necessary to that end. At the same time the person bearing responsibility is confronted with the difficult task of determining whether the means are morally justified given the moral

importance of the goals and placing the value and nonvalue of the means in an appropriate relation to the value and nonvalue of the ends.

A *second affinity* between utilitarianism and the concept of responsibility lies in the equivalence of intentional and unintentional but expected consequences. This equivalence is characteristic for a consequentialist ethics, in particular for utilitarianism. It is one of the central characteristics of a utilitarian ethics that *intentional* and *indirect consequences* are weighted equally and that foreseeable consequences that are accepted are not 'discounted' in relation to those that are intended. This leads to a conflict, on the one hand, with the so-called 'double effect principle' in the tradition of Catholic moral theology, according to which intended consequences count significantly more than unintended ones, but also with elements of everyday morality and criminal law. On the other hand, there is a correspondence between this characteristic and the prevailing meaning of the term responsibility. Whoever is responsible for achieving certain goals is not only responsible for not using means that are so negative that they outweigh the positive value of the goal but also for not accepting any harms that might be expected as a consequence of using this means and that are so negative that they outweigh the positive value of the goal to be achieved.

Whoever is responsible for, say, keeping the peace is not only also responsible for not using means that would cause more harm than the harm being avoided but also for the *indirect* consequences of the harm caused not being worse than the harms being avoided. The bearer of responsibility cannot excuse himself that he accepted the indirect consequences of the means 'for the sake of a good cause' – at least not in a utilitarian perspective. The best intentions do not change anything on the moral wrongness of an action. Actions, action strategies or rules for which it is foreseeable that their bad indirect consequences – including the bad indirect consequences of the means employed to reach a goal – outweigh their well-intentioned consequences are no less morally wrong than actions that are intended, from the start, to inflict harm.

A *third affinity* between a utilitarian ethics and the concept of (prospective) responsibility may have far greater consequences for the practice of the attribution, assumption and acceptance of responsibility. This affinity is founded in a structural characteristic of responsibility related to the specific moral character of this responsibility.

It is an essential characteristic for every form of *moral* responsibility that it is not only attributable to oneself but to others. To the extent that it is morally necessary to bring about or prevent certain future events or states, or advance their occurrence or non-occurrence, this responsibility is not only to be borne by each person equally, but those who respond to this moral imperative may and must attribute this responsibility to each other reciprocally.

Attributing responsibility to others must satisfy stricter conditions than to oneself. While for self-attributions attributing and assuming responsibility are more or less the same thing, when attributing responsibility to others they are not. It is not necessary to provide any special justification when placing oneself under a particular responsibility, but we can only expect another person to assume a responsibility we are placing on him if we provide reasons that allow him to understand and, without direct or indirect coercion, accept them as reasonable. Whoever proposes to another person that he assume a responsibility will have to provide reasons why it should be accepted.

This condition limits the possible content of reciprocally attributed moral responsibility in two ways. First, if A attributes a corresponding responsibility to B, A will not be allowed to appeal to values that he can only justify by appealing to *authority*. A cannot expect that B accepts the authority he is appealing to, regardless of whether it is a law, a cultural tradition or a religious authority. Moral responsibility can only have its source in an authority that is acceptable to everyone whatever their specific traditions or loyalties. Second, he will also not be able to appeal to values that are only understandable or acceptable given certain metaphysical presuppositions. A cannot assume that B will share his specific metaphysical convictions.

Of course this last condition implies restrictions on the value theory foundations of attributing responsibility if values can only be justified by recourse to authorities or metaphysical assumptions. It is likely that most values justified by *de facto* appeals to authority or metaphysical assumptions can also be justified without recourse of this kind. In most cases the appeal to authority or metaphysical assumptions serves only to strengthen their rhetorical and persuasive effect. For example, it is undoubtedly possible to imagine other justifications for the values postulated by Hans Jonas in his avowedly metaphysical theory of a 'responsibility to the future' – such as the preservation of a higher human civilization.

Alongside this critical argument there is a more positive and substantial plausible argument to be gained from examining the conditions of the reciprocal attribution of moral responsibility. A can only expect B to accept the responsibility A is asking of him if its assumption and acceptance – however it might be requested – promotes the realization of a value that A can assume B would accept. Is there such a value? There is much to suggest that there is only one value, namely the value of subjective well-being, the experience of states of consciousness subjectively assessed as positive. It is only this elementary value that can claim to be accepted by any individual B. That it is fundamentally better that someone feels better than worse – in his own estimation – is such an elementary value assumption that it can be attributed to all axiological systems both past and present and regardless of their other differences. It seems to be the only value assumption that can be agreed upon by otherwise

so heterogeneous value theories in subjective and objective, ascetic and hedonistic, minimalistic and maximalistic ethical traditions. While there is much wide-ranging and problematic dissension about the intrinsic value of virtue, dignity, justice, harmony and beauty, the assumption that what a subject feels for himself and regardless of the consequences as a positive state of consciousness – and so is also objectively something positive – can be considered a good candidate as a something held in common by every axiology ever proposed.

The consequence is that it is easier to justify attributing responsibility to another person if the responsibility can be related to the intrinsic value of subjective well-being. The core of moral responsibility would be, in this regard, responsibility for the subjective well-being of conscious beings, including the creation and maintenance of its conditions and the prevention or amelioration of impediments and threats to its continuation. All intrinsic responsibility is at its core responsibility for the subjective well-being and happiness of conscious beings, all extrinsic responsibility is responsibility for its direct and indirect preconditions. Between utilitarianism and responsibility there is thus an extremely close relationship. The concept of responsibility seems to lend itself to a utilitarian interpretation.

IV Responsibility in the Long Run: The Problem of Motivation

'Motivation problem' is not a commonly used term in ethics. Nevertheless, it can serve as a convenient label for an inquiry into the conditions that have to be fulfilled in order to make a norm, prescription, recommendation or any other action-guiding statement effective in the sense of making the addressee behave in conformity with it.

This question arises because ought statements, like requests to assume and accept responsibility, are in themselves unable to compel compliant behaviour but instead are dependent on a corresponding willingness on the part of the addressee. Even ought statements in the form of a categorical proposition – those including a 'must' – give the addressee the freedom to say no. Nevertheless, it seems obvious that more should be said about the motivation to follow moral demands, even when they are meant merely as guides to action and not as requirements. Whoever accepts a moral demand has a reason to orient his behaviour towards it and so is at least partially motivated to follow it.

However, the recognition of an obligation is in general not a *sufficient condition* for the actual exercise of a responsibility. Even if in agreement with the more plausible and in moral

psychology more widely accepted concept of *internalism* we assume that recognition in itself contains an element of moral motivation, we cannot assume that this is sufficient to bring about a given behaviour. In order to make moral principles or resolutions effective there must as a rule be additional motivation.

For many forms of attribution of 'responsibility at a distance', just as for the responsibility for avoiding future harmful consequences of climate change, the problem of motivation is especially acute because there seems to be a wider gap between the willingness to recognize this obligation and the willingness to act according to this obligation than there is in other areas of morality. Even in Germany – a country pledged to renewable energy and often seen by other nations as a pioneer of forward-looking climate protection – climate policy is more about words than deeds.

Psychologically, the discrepancy between words and deeds in climate policy is easily explained. In the climate problem a number of factors come together that are known to have an inhibiting effect on the motivation to behave in accordance with one's own moral norms: the strong *future dimension* of responsibility for the climate, the *social distance* to those who are affected, and the menace of the changes needed to one's habitual *lifestyle*.

The first factor – the relation to a future that we will not see ourselves, one that is abstract and difficult to imagine – is something the responsibility for the climate has in common with the responsibility for the long-term preservation of the basis of human life and biodiversity. In both cases a responsibility over the long run has been acknowledged as urgent. That there is still a 'motivation gap' between acceptance and acting in accordance with this duty over the long term can be explained by the special features of these duties.

The first special feature is that obligations related to a distant future are necessarily *non-reciprocal*. From future generations we can expect nothing in return for present sacrifices but we also need not fear sanctions. Neither can they do something for present generations, nor can they be compensated by the present generation for irreversible harms they will suffer. They are unilateral beneficiaries but also unilateral victims. On the positive side, they have the present to thank for an enormous growth of knowledge and technology, which they can at most symbolically thank the present for; on the negative side, they will suffer an enormous loss of exhaustible resources and biodiversity, which they can at most symbolically deplore. They are unable to be compensated for the harms we are inflicting upon them. They are unable to even demand from us reparation. When climate change has its most deleterious consequences those who bear responsibility will long since be among the deceased. While children and grandchildren today can claim their due and protest future burdens, great-grandchildren do not have a voice. If they have a voice then at most that

their interests and rights are anticipated and their claims put forward by those in the present advocating on their behalf.

A second aspect is that future developments are less certain than those of the present and that the causality of actions taken today on future life conditions are more difficult to estimate than the causality of past or present actions on spatially far off regions in the world. Even if some estimates are more certain than others – and projections of demographic trends until mid-century show much less variation than estimates of the destruction of biodiversity resulting from climate change – uncertainty is a more than negligible variable.

Uncertainty affects a number of dimensions. *First*, there is a residual uncertainty regarding the reliability of the scientific scenarios risk forecasts are based on. Even if there is little room for doubt about the physics of climate change, there is a much greater scope concerning the question of how increasing temperatures will affect the economy, the living conditions and – the most relevant dimension ethically – the quality of life. Since the motivation to reduce greenhouse gas emissions is dependent on the consequences for oneself and one's immediate descendants, uncertainties about their local and regional effects are of great importance, especially with regard to the exodus of climate refugees that can be expected to besiege the wealthy fortress Europe, straining its assimilation capacities.

The *second* dimension of uncertainty is the unpredictability of technical progress. It cannot be ruled out that less risky technological solutions for neutralizing carbon dioxide will be found than those being currently discussed under the heading geo-engineering.

Third, it is uncertain whether today's efforts to reduce emissions will have any appreciable effect on ethically relevant objectives. We are confronted with a systematic lack of feedback on the success and failure of long-term provisions for the future. 'Control beliefs' are missing, and these are crucial for the willingness of an individual to adapt his behaviour to his own principles. Without suitable convictions about our ability to control our environment our behavioural motivation is necessarily unstable.

A further aspect is the uncertainty about the extent to which successive generations will continue strategies initiated today. We cannot be certain that our descendants will share our values and norms and continue a transition process we have begun. Much depends on how well the current generation succeeds in demonstrating to future generations that they are able to forego fossil fuels without having to accept losses in prosperity or disappoint expectations about future growth.

The second factor, the *social distance* to those most affected, has similarities with the futurity factor yet goes further. As a large part of those most affected live in the future they

are necessarily anonymous. They appear as 'statistical' instead of 'identified' victims. Motivating feelings leading to solidarity action is however more likely to be triggered by people 'in front of our very own eyes' suffering great harm or on the point of a catastrophe (in a mining accident, an earthquake or an epidemic), even in cases in which an unemotional utilitarian calculation would tell us to use the resources not for saving others but for prevention efforts.

Another specific aspect of responsibility for the climate is that those most affected are likely to belong to other cultures, and so they are outside the empathy horizon of the main actors.

The third difficulty, the necessity to adapt our habits and lifestyle, may have an even greater effect on the possibilities of a rigorous adaptation strategy. Changing habitual lifestyles is a difficult undertaking, and in democracies politicians have understandable reservations about distancing themselves too far from the average voter. The surprisingly fast implementation of the social preference against smoking is not a suitable counter-example. Carbon dioxide emissions do not endanger an individual's own health or the immediate environment. The risks remain abstract and so appeal to 'cold reason' rather than to the heart. A visceral response that could motivate appropriate behaviour is missing. And so in today's industrialised nations the ambition to protect the climate exists side-by-side with the desire to preserve our habits and lifestyles, like fossil fuel mobility.

All three factors – the *future reference* of climate responsibility, the social distance to those most affected and the *conservatism of lifestyle* – contribute to our more easily repressing awareness of even unambiguously recognised dangers than of dangers that are more immediately threatening, that affect us personally or those close to us, or that can be handled without extensive changes to our behaviour. That warnings about potential catastrophes in the future are less likely to trigger solidarity than catastrophes occurring in the present can be better understood by examining three separate aspects: non-reciprocity, uncertainty and the anonymity of future generations. These are all reasons for doubting that motivations arising from the recognition of responsibility for the climate will be intensive and reliable enough – given the bombardment of competing moral and non-moral values and goals – to have an effective impact on behaviour.

This does not mean however that the prognosis for a motivation to care for the climate is completely bleak. A more favourable outlook than *direct* motivation is to consider *indirect* motivation to assume responsibility for the climate. The object of indirect motivation is not caring for future generations themselves but achieving other more immediate goals that we assume will contribute to caring for future generations. The decisive advantage of indirect motivation is its more stable emotional foundation. Indirect motivation can, generally

speaking, rely on a number of emotional factors that are unavailable to direct motivation in the same way; it can make use of 'quasi-moral motives' – motives such as love, pity, caring and solidarity – which are to a large extent functionally equivalent to moral motives but are more closely bound to affectively coloured relationships and needs.

The best-known model of indirect motivation for responsibility in the long run is the *chain of love*, the intergenerational linking of caring and precaution by each generation for the following generation. In this model each generation merely cares for the generation of their children. The 'linking' of generations, each caring for the next, has the same effect as a first generation assuming a hypothetical responsibility for all succeeding generations. Assuming the behaviour of the parents has a role model function and their children will care for their own children in the same way that their parents cared for them, then the same care will be given to the great-grandchildren as if each generation had oriented itself toward an abstract moral principle of caring for all future generations. The point of this model is that if the current generation does not care for the great-grandchildren's generation and only for their children's generation then the great-grandchildren may very well do better than if all had followed a more ambitious intergenerational moral principle.

A second form of indirect responsibility in the long run – and one without moral motives in a strict sense – is the preservation and care of intrinsic cultural values. The appreciation of cultural values – such as, certain forms of art, music, literature, philosophy and science as well as social virtues and political institutions – is anthropologically closely linked to the motive to preserve these values and know that they will be preserved over the long term. Whoever loves Bach's music also has, as a rule, an interest in preserving this music and ensuring that future generations, even if they have little appreciation for it, will also preserve and hand down this music to successive generations. It is hard to imagine that someone can seriously prize values such as scientific truth, artistic perfection or the principles of democracy and not at the same time at least hope that they – in analogy to Nietzsche's 'all desire wants eternity' – never pass away.

The most important project of this kind for climate ethics is a permanent respect for human rights. Human rights do not have a timeline. As fragile an achievement as they may be and the result of an arduous historical process of humanisation, which is by no means at its end, they have a timeless validity. Whoever values them now will always value them.

A further indirect motive that can serve as a stable basis, both today and tomorrow, for the assumption of responsibility for the climate is the human need for overarching goals that go beyond one's self, one's own community and one's own lifetime. In a secularised and globally networked world this need can best be fulfilled with universal future-oriented

goals. Ernest Partridge has called such goals motives of self-transcendence. They could also be called motives of *sense-making*. Caring for a universal future is an especially appropriate example of these motives as it is through his engagement for the future that an individual confirms his own value and feels secure in being part of a greater overarching context providing meaning. He is a link in a generational chain that is held together by an inter-generational sense of community, of which gratitude looking backwards is just as much a part as is recognition of obligations looking forward. This motive can be especially strong when it is supported by membership in a like-minded community. A not unimportant factor is also that moral engagement for a future that is beyond our experience – in this respect similar to a transcendental god – is *undisappointable*.

Also *self-binding commitment* through future-proof institutions represents a way of transferring the moral burden of caring for the future to indirect motivation. Self-commitment can be understood as the long-term replacement of direct by indirect motivation, which is always advisable when direct motivation is not reliable enough for individuals to promptly assume particular responsibilities. Whoever enters a long-term contract – whether for life assurance or a regular charitable donation – finds it easier to remain true to an obligation and makes it more difficult to give in to the temptation to relinquish a commitment once made. He limits the scope of his future decisions and actions by committing himself to general guidelines and replaces direct motivation for long-term provisions or charitable actions by indirect motivation to prevent the undesirable short-term consequences of a breach of contract or the cancellation of an agreement. Whoever knows that he that his motivation flags and he tends to give into impulsive moments but also knows that this will endanger his long-term goals will in general be better off by structuring his options so that impulsive motives are directed over the long term in the 'right' way and serve to promote rather than hinder his overarching goals. In the sense of a utilitarian 'ethics at a distance', he would do well, in other words, to live *virtuously*.

This brings us at the end – by roundabout ways – to the utilitarian John Stuart Mill and his praise of virtue as an end in itself:

> It maintains not only that virtue is to be desired, but that it is to be desired disinterestedly, for itself. Whatever may be the opinion of utilitarian moralists as to the original conditions by which virtue is made virtue … they not only place virtue at the very head of the things which are good as means to the ultimate end, but they also recognise as a psychological fact the possibility of its being, to the individual, a good in itself, without looking to any end beyond it; and hold, that the mind is not in a right state, not in a state conformable to Utility, not in the state most conducive to the general happiness, unless it does love virtue in this manner – as a thing desirable in itself, even although, in the individual instance, it should not produce those

other desirable consequences which it tends to produce, and on account of which it is held to be virtue. (CW X, 235)

References

[1] Bentham, Jeremy. 1948. *An Introduction to the Principles of Morals and Legislation*. New York: Hafner Pub. Co.

[2] Dietler, Wilhelm. 1997. *Gerechtigkeit gegen Thiere*. Bad Nauheim: Asku Presse.

[3] Gesang, Bernward. 2011. *Klimaethik*. Berlin: Suhrkamp.

[4] Kamlah, Wilhelm. 1973. *Philosophische Anthropologie. Sprachliche Grundlegung und Ethik*. Mannheim: Bibliographisches Institut.

[5] Kant, Immanuel. 1902. *Werke, Akademie-Ausgabe Bd. 8.* Berlin.

[6] Lumer, Christoph. 2002. *The Greenhouse. A welfare assessment and some morals*. Lanham, MD: University Press of America.

[7] Mill, John Stuart. 1963-1991. *The Collected Works of John Stuart Mill*, 33 vols. Toronto: University of Toronto Press / London: Routledge and Kegan Paul.

[8] Nordhaus, William D. 2013. *The Climate Casino: Risk, Uncertainty, and Economics for a Warming World*. New Haven, CT: Yale University Press.

[9] Prichard, Harold A. 1950. *Knowledge and Perception*, Oxford: Clarendon Press.

[10] Rousseau, Jean-Jacques. 1978. “Über den Ursprung der Ungleichheit unter den Menschen.” In: *J. J. Rousseau: Schriften zur Kulturkritik*, edited by Kurt Weigand, 61-269. Hamburg: Meiner.

[11] Singer, Peter. 1972. “Famine, Affluence, and Morality.” *Philosophy & Public Affairs* 1 (3): 229-243.

[12] Stern, Nicholas (2006): *The Economics of Climate Change: The Stern Review*. Cambridge, UK: Cambridge University Press.

Was Abraham Lincoln a Utilitarian?[1]

Thomas L. Carson, Loyola University Chicago, USA

Abstract

There is considerable *prima facie* evidence that Lincoln was a utilitarian. He said that we should judge actions by their "fruits" (consequences). He also said:

> I hold that while a man exists, it is his duty to improve not only his own condition, but to assist in ameliorating mankind; and, therefore ... I am for those means which will give the greatest good to the greatest number.
> The true rule, in determining to embrace, or reject any thing, is not whether it have any evil in it; but whether it have more of evil than of good. There are few things wholly evil, or wholly good. Almost everything, especially governmental policy, is an inescapable compound of the two; so that our best judgment of the preponderance between them is continually demanded.
> I would consent to any GREAT evil, to avoid a GREATER one.

However, Lincoln endorsed other moral principles that can sometimes conflict with utilitarianism. He said that we should obey the law and follow God's will. He also thought that he was morally obligated to abide by his oath of office to execute the law faithfully and defend the US Constitution. Lincoln didn't have a fully consistent moral philosophy. But, while he was President of the United States, Lincoln was a utilitarian, *in practice*. In all of his important decisions and policies regarding slavery and the American Civil War, he tried to do what would have the best consequences. In these cases, he saw no conflict between utilitarianism and the other moral principles that he endorsed. Lincoln thought that it is very seldom possible to discern God's will. He also believed that, in order for his policies to succeed, he needed to act in accordance with the law and his oath of office.

Introduction

Was Abraham Lincoln a utilitarian? The evidence is mixed, but, while he was President of the United States, Lincoln was an act-utilitarian, in practice.

[1] Tim Mulgan has asked me whether Lincoln might have been some kind of rule-utilitarian, or rule-consequentialist. I hope to address this question in a longer version of this paper.

I Some Evidence for Thinking That Lincoln Was an Act-Utilitarian

Lincoln endorsed explicitly act-utilitarian moral principles at different times in his life. In a speech in February 1861, shortly before he became president, he said:

> I hold that while a man exists, it is his duty to improve not only his own condition, but to assist in ameliorating mankind; and, therefore ... I am for those means which will give the greatest good to the greatest number. (Lincoln 1989, II, 203)

In a speech in the US House of Representatives in 1848, he said:

> The true rule, in determining to embrace, or reject any thing, is not whether it have any evil in it; but whether it have more of evil than of good. There are few things wholly evil, or wholly good. Almost everything, especially governmental policy, is an inescapable compound of the two; so that our best judgment of the preponderance between them is continually demanded. (Lincoln 1989, I, 192)

In October 1854, he said:

> I would consent to any GREAT evil, to avoid a GREATER one. (Lincoln 1989, I, 333)

In an early speech from 1842, he talked about the great good that the temperance movement had done without causing much harm and said:

> If the relative grandeur of revolutions shall be estimated by the great amount of human misery they alleviate, and the small amount of harm they inflict, then, indeed, will this be the grandest the world shall ever have seen. (Lincoln 1989, I, 89)

The context of this passage makes it clear that he *is* saying that we should estimate the goodness of social movements by how much human misery they cause and relieve. In 1852, Lincoln praised Henry Clay for not wanting to *immediately* eradicate slavery because it would produce "a greater evil, even to the cause of human liberty itself." (Lincoln 1989, I, 269)

II Some Apparent Evidence That Lincoln Was Not an Act-Utilitarian - That, on Examination, Is Actually Strong Evidence for Thinking That He Was One

Sometimes following act-utilitarianism requires that one break promises. As a young man, Lincoln had a very strong commitment to keeping his promises come what may. In an early letter, concerning his promise to wed Mary Owens, a promise that he had come to regret, Lincoln wrote, "I made a point of honor and conscience in all things to stick to my word, especially if others had been induced to act on it ..."[2] Later, he felt honor bound to keep the promise he had made to wed Mary Todd, even though he had grave doubts about doing so (Carson 2015, 318-26). In connection with this promise, he quoted and endorsed his father's saying "If you make a bad bargain, *hug* it the tighter." (Lincoln 1989, I, 91)

But his views about the moral obligation to keep promises changed considerably by the end of his life. In his last public speech on April 11, 1865, just three days before his assassination, he discussed a plan he had proposed to Congress earlier. He said that this proposal had made a promise to states that had seceded from the Union and added:

> But as bad promises are better broken than kept, I shall treat this as a bad promise, and break it, whenever I shall be convinced that keeping it is adverse to the public interest. (Lincoln, 1989, II, 698)

This is a sharp departure from his earlier view that bad bargains should be kept and held all the tighter. His criteria for bad promises are explicitly utilitarian. He says that his earlier promise should be regarded as a bad promise, provided that keeping it would be adverse to the public interest. He also says that he will (or should) break bad promises. This is *exactly* what an act-utilitarian would say.

In a letter to Williamson Durley dated October 1845, Lincoln rebuked Durley and other New York members of the abolitionist Liberty Party for refusing to vote for Henry Clay for President because Clay was a slaveowner.[3] In this letter, Lincoln endorsed a moral principle that many take to inconsistent with utilitarianism. Lincoln wrote:

[2] "Letter to Mrs. Orville Browning," April 1, 1838, Lincoln 1989, I, 38.

[3] The Liberty Party drew enough votes from Clay to alter the outcome of the 1844 US presidential election. If the 15,000 people who voted for the Liberty Party candidate in New York State had voted for Clay instead (almost all of them preferred the Whig, Clay, to the Democrat, Polk), Clay would have won the election.

> "We are not to do *evil* that good may come." This general proposition is doubtless correct. (Lincoln 1989, I, 111-2)

Durley and other members of the Liberty Party said that "we should not do *evil* that good may come" and took voting for Clay to be an evil action. Lincoln asked whether voting for Clay was an evil action:

> If by your votes you could have prevented the *extension [sic]*, &c., of slavery, would it not have been *good* and not *evil* so to have used your votes, even though it involved the casting of them for a slaveholder? By the *fruit* the tree is to be known. An *evil* tree can not bring forth *good* fruit. If the fruit of electing Mr. Clay would have been to prevent the extension of slavery, could the act of electing have been *evil*? (Lincoln 1989, I, 112)

Lincoln says that whether or not an action is evil depends its "fruits" (consequences). So, appearances to the contrary, his endorsement of the principle that we shouldn't do evil that good may come is perfectly consistent with utilitarianism. In fact, his discussion of this principle strongly supports the view that he was an act-utilitarian.

Act-Utilitarianism requires that one violate the law whenever doing so will produce better consequences than not. But this seems inconsistent with Lincoln's reverence for the law. In a very early speech from 1838, Lincoln said "Let every American ... swear ... never to violate in the least particular, the laws of the country; and never to tolerate their violation by others."[4]

But later, while he was President, Lincoln was willing to defy the law for utilitarian reasons. He defied Chief Justice Taney's order overruling Lincoln's suspension of the right of habeas corpus in the Merryman case in 1861. In this case, he put utilitarian considerations ahead of obeying the law. Arguably, by the end of his life, Lincoln gave priority to doing what has the best consequences over obeying the law. In the Merryman case, he was concerned to minimize violations of the law as opposed to maximizing welfare. He asked "are all the laws, *but one*, to go unexecuted, and the government itself go to pieces, lest that one be violated?"[5] In effect, he said that he violated one law so that many other laws would not be unenforced.

[4] Lincoln 1989, I, 32. Still, his reverence for the law seems to have had a utilitarian rationale - his fear that general lawlessness and mob rule would result were people able to break the law with impunity ("Address to Young Men's Lyceum of Springfield," in Lincoln, 1989, I, 28-36.) Further, he clearly did not think that all people at all times and places have an unconditional duty to obey the law, because he held that people oppressed by unjust governments have the right to revolution and he approved of the American Revolution, Lincoln, 1989, I, 167 and 32.

[5] "Address to Congress July 4, 1861," in Lincoln 1989, II, 253.

III Some Genuine Counter-evidence to the View That Lincoln Was an Act-Utilitarian

As we have seen, Lincoln's views about the morality of keeping promises changed considerably and were clearly consistent with utilitarianism by the end of his life. But, he apparently made a sharp distinction between ordinary promises and oaths, which we might describe as solemn promises.

Lincoln took very seriously his oath of office to "faithfully execute the Office of President of the United States, and ... preserve, protect, and defend the Constitution of the United States." He called it "an oath registered in Heaven."[6] He made a distinction between his official duty as President, which required him to defend the US Constitution and execute the laws of the United States, and his personal moral beliefs. He thought that slavery was morally wrong but that his public duty required him to follow and execute laws that protected the institution of slavery.[7] In his First Inaugural Address, Lincoln read a portion of the US Constitution which includes an explicit provision for the return of fugitive slaves. That provision reads "No person held to service or labor in one State under the laws thereof, escaping into another, shall in consequence of any law or regulation therein be discharged from such labor, but shall be delivered up on claim of the party to whom such labor or service may be due."[8] Lincoln promised to enforce this provision of the Constitution.

Lincoln seems to have taken his duty to abide by his oath of office to be an absolute or unconditional moral obligation.

Lincoln's frequent statements to the effect that we should follow God's will are another objection to the view that he was a utilitarian. He seems to have believed that the obligation to follow God's will, *when one can discern it*, is an absolute unconditional obligation.

In a speech in Worcester, Massachusetts in 1848, Lincoln gave priority to following God's will over utilitarian considerations. He said that "when divine or human law[9] does not clearly

[6] "First Inaugural Address," in Lincoln 1989, II, 224.

[7] "Letter to Horace Greeley," Lincoln 1989, II, 358.

[8] Article IV section 2 of the US Constitution.

[9] Whether or not Lincoln takes "human law" to mean the statutes of one's society is unclear.

point out what *is* our duty" we must discern what our duty is "by using our most intelligent judgment of the consequences."[10]

In a public response to a group pressing him to end slavery in September 1862, just nine days before he issued the Emancipation Proclamation, Lincoln said:

> If it is probable that God would reveal his will to others, on a point so connected with my duty, it might be supposed he would reveal it directly to me ... it is my earnest desire to know the will of Providence in this matter. *And if I can learn what it is I will do it!* These are not, however, the days of miracles, and I suppose it will be granted that I am not to expect a direct revelation. I must study the plain facts of the case, ascertain what is possible and learn what appears to be right. (Lincoln 1989, II, 361)

In his private notes "On Pro-slavery Theology," from 1858, Lincoln discusses the view that American slavery was in accordance with God's will. He endorses the idea that we should follow God's will, but says that there is difficulty in ascertaining it:

> Certainly there is no contending against the will of God; but still there is some difficulty in ascertaining, and applying it, to particular cases. (Lincoln 1989, I, 685)

Lincoln thought that the institution of slavery, which deprived enslaved people of the fruits of their labor, was contrary to God's will. In a 1859 speech he said:

[10] This speech criticized the anti-slavery Free Soil Party and its presidential candidate Martin Van Buren. Lincoln claimed that the Whigs and their candidate Zachary Taylor were just as much opposed to the extension of slavery as the Free Soil Party. He argued that supporting Van Buren would promote the election of the Democratic candidate, Cass, who supported the extension of slavery into new states and territories. Here is a larger portion of Lincoln's speech that includes the passages quoted above:

The "Free Soil" men in claiming that name indirectly attempted a deception, by implying the Whigs were *not* Free Soil men. In declaring that they would "do their duty and leave the consequences to God,' merely gave an excuse for taking a course that they were not able to maintain by a fair and full argument. To make this declaration did not show what their duty was. If it did we should have no use for judgment, we might as well be made without intellect, and when divine or human law does not clearly point out what *is* our duty, we have no means of finding out what it is by our most intelligent judgment of the consequences. If there were divine law, or human law for voting for Martin Van Buren, or if *a fair examination of the consequences* [my emphasis] and the first reasoning would show that voting for him would have the *best consequences* [my emphasis] and first reasoning would show that voting for him would bring about the ends they pretended to wish - then he [Lincoln] would give up the argument. But since there was no fixed law on the subject, and since the whole probable result of their action would be an assistance in electing Gen. Cass, he [Lincoln] must say that they were behind the Whigs in their advocacy of the freedom of the soil.

Basler 1953, Volume 2, 3-4. This passage is from a summary transcription of Lincoln's speech in a Boston paper, the *Daily Advertiser*, September 14, 1848. That accounts for the fact that this is not written in the first person voice and it talks about what "he" [Lincoln] said.

> I hold that if there is any one thing that can be proved to be the will of God by external nature around us, without reference to revelation, it is the proposition that whatever any one man earns with his hands and by the sweat of his brow, he shall enjoy in peace. I say that whereas God Almighty has given every man one mouth to be fed, and one pair of hands adapted to furnish food for that mouth, if anything be proved to be the will of Heaven, it is proved by this fact, that the mouth is to be fed with those hands, without being interfered with by any other man who has hands to labor with. I hold that if the Almighty had ever made a set of men that should do all the eating and none of the work, he would have made them with mouths only and not hands, and if he had ever made another class that he had intended should do all the work and none of the eating, he would have made them with all hands.[11]

Lincoln often quoted the passage in Genesis 3:19 in which God tells Adam and Eve "By the sweat of your face you shall eat bread until you return to the ground." Lincoln appealed to this passage to show that slavery (which allowed some people to eat bread from the toil of others) was contrary to God's will.

His "Meditation on the Divine Will" written in September 1862, shortly before he issued the Emancipation Proclamation, (Lincoln 1989, II, 359) also provides strong evidence that he sought to discern and follow God's will. Lincoln wrote:

> In the present civil war it is quite possible that God's purpose is something different from the purpose of either party - and yet human instrumentalities, working just as they do, are of the best adaptation to effect His purpose. I am almost ready to say this is probably true – that God wills this contest, and that it not end yet.... He could have either *saved* or *destroyed* the Union without a human contest.... Yet the contest proceeds. (Lincoln 1989, II, 359)

At the time he wrote this, both Lincoln and the Confederate State of America were seeking a quick and relatively bloodless victory in the American Civil War. The Confederates wanted to gain their independence and preserve and expand the institution of slavery. Lincoln's publically declared aim in fighting the war at this time was to preserve the union of the states and stop the spread of slavery. The upshot of this reflection is that Lincoln had come to believe that God willed that the war be long and terrible so that it would end American slavery. This is a central theme of his Second Inaugural Address. Lincoln's "Meditation on the Divine Will" supports the view that he was sincere several weeks later when he told his cabinet that he sought to discern and do God's will. During General Lee's first invasion of the North in the late summer of 1862, Lincoln made a solemn vow to God to issue the Eman-

[11] Lincoln 1989, II, 85. Also see Lincoln's Speech at Hartford March 5, 1860, in Basler 1953., IV, 9.

cipation Proclamation if the Union defeated Lee's invasion. He reported this vow to his Cabinet and told them that, because of the Union victory at the Battle of Antietam on September 17, 1862, "God had decided this question in favor of the slaves."[12]

IV Weighing the Evidence

How should we interpret this seemingly conflicting evidence? Act-utilitarianism can conflict with these other moral principles that Lincoln endorsed.

Since he apparently thought that, in practice, there were few, if any, serious conflicts between utilitarianism and these other moral principles, Lincoln might not have been interested in determining which principles were most fundamental or which took precedence in case of conflict. So, one possible interpretation of his moral views is this:

> 1. Lincoln endorsed a number of different moral principles including act-utilitarianism and had no opinion about which were most fundamental.

But this interpretation is difficult square with evidence that Lincoln attached great importance to following God's will and to his oath of office. Other reasonable interpretations of Lincoln's moral views are these:

> 2. Lincoln believed that God's will is the ultimate moral standard, but that it is seldom possible to discern God's will and that we should follow a version of act-utilitarianism, subject to the side constraint of keeping solemn oaths, when we cannot discern God's will.
>
> 3. Lincoln held that act-utilitarianism subject to the side constraint about keeping oaths is the true/correct moral principle and that God, when we can discern God's will, is the ultimate epistemic authority about morality.
>
> 4. Lincoln believed that: a. we should follow God's will, b. God is benevolent and desires human welfare, c. God desires that we keep solemn oaths, therefore, Lincoln believed that d. we should always do what will have the best consequences or best promote human welfare (subject to the side-constraint of keeping solemn oaths).

[12] This was reported by Lincoln's Secretary of the Navy Gideon Wells; see Fehrenbacher and Fehrenbacher 1996, 474.

According to interpretation 4, Lincoln endorsed both act-utilitarianism (with the side-constraint about keeping oaths) and the view that we should follow God's will and takes these principles of be fully consistent. According to this interpretation, Lincoln took the principle that we should follow God's will to be more fundamental, i.e., he thought that this constrained version of act-utilitarianism is true because God wills that we so act. I have no direct evidence that Lincoln held either b. or c.

1-4 are all possible interpretations of Lincoln's moral views. I don't have any decisive reasons for preferring any one of these interpretations, though I think that 2 is more likely to be the correct interpretation than 1, 3, or 4. 4, however, is the interpretation which makes Lincoln's moral views most coherent.

For my purposes in this paper, it is not necessary for me to defend any particular interpretation of Lincoln's moral beliefs. Regardless of what we say about that issue, it is clear that Lincoln was an act-utilitarian in practice, if not in theory, while he was President of the United States. In his major decisions and policies as President (his policies on the expansion and abolition of slavery, Southern secession, the suspension of the right of habeas corpus, the conduct of the American Civil War, the colonization of freed slaves, and the post-war status of African Americans), Lincoln always chose the actions which he thought would have the best consequences. I defend this claim at great length in my book *Lincoln's Ethics* and also argue that his actions and policies in the most important cases *did* have the best consequences. Further, since he very seldom thought that he could discern God's will, and since he rarely made choices about issues concerning which his oath of office applied, his actions were almost always consciously guided by utilitarian considerations. Lincoln knew that the powers of public opinion,[13] the US Congress, and the US Supreme Court were such that he was unlikely to succeed in any action or policies that were clearly contrary to the US Constitution and his oath of office. There is *not a single important case* during his presidency in which he took himself to be in a position in which by bringing about the best consequences he would be violating his oath of office or the will of God.[14] So, *in practice*, there was *no difference* between his moral views and those of an act-utilitarian. In his official capacity as

[13] In his first debate with Stephen Douglas in August, 1858, Lincoln talked about the importance of public opinion. He said "In this and like communities, public sentiment is everything. With public sentiment, nothing can fail; without it nothing can succeed. Consequently he who molds public sentiment, goes deeper than he who enacts statutes or pronounces decisions," Lincoln 1989, II, 524-5.

[14] It is important to stress that in the Merryman case when he defied the order of Taney's court and thus broke the law, he still took himself to be acting in accordance with his oath of office to "execute" the laws of the United States (see Carson 2015, 90-1). Recall that he asked "are all the laws, *but one*, to go unexecuted, and the government itself go to pieces, lest that one be violated?" In effect, he said that he violated one law so that many other laws would not be unenforced.

President of the United States, Lincoln was almost always trying to do what would have the best consequences and he never took this to be at odds with the other moral principles he endorsed.

Appendix

What Lincoln Read

Lincoln read J. S. Mill's *On Liberty* and *Principles of Political Economy* (1848 edition). He also read Hume's *Essays* and Francis Wayland's *Elements of Political Economy*.[15]

Mill on Lincoln and the American Civil War

Mill's 1862 essay on the American Civil War, "The Contest in America," is of considerable interest.

References

[13] Basler, Roy, et. al., eds. 1953. *The Collected Works of Abraham Lincoln.* New Brunswick: Rutgers University Press.

[14] Bray, Robert. 2010. *Reading With Lincoln.* Carbondale, Ill.: Southern Illinois University Press.

[15] Burlingame, Michael. 2008. *Abraham Lincoln: A Life.* Baltimore: Johns Hopkins University Press.

[16] Carson, Thomas. 2015. *Lincoln's Ethics.* New York: Cambridge University Press.

[17] Fehrenbacher, Donald, and Virginia Fehrenbacher, eds. 1996. *Recollected Words of Abraham Lincoln.* Stanford, California: Stanford University Press.

[18] Geulzo, Allen. 1999. *Abraham Lincoln: Redeemer President.* Grand Rapids, Michigan: William B. Eerdmans Publishing Company.

[19] Lincoln, Abraham. 1989. *Speeches and Writings.* New York: The Library of America.

[15] Bray 2010, 226-8; also see Guelzo 1999, 106-7.

Liberty as Resistance against Oppression and Epistemic Injustice in J. S. Mill

Gustavo H. Dalaqua, University of the State of Paraná, Brazil

Abstract

This chapter argues that J. S. Mill's philosophy advances a conception of liberty that entails resisting oppression and epistemic injustice. Whereas oppression refers to any act that deliberately curtails citizens' self-development, epistemic injustice denotes a specific type of oppression that harms people's capacity to know themselves and their desires. In *The Subjection of Women*, Mill elaborates a conception of liberty as non-subjection, which indicates that people lose their freedom when they suffer epistemic injustice. Since they were subjected to a system of education that shaped their psyche in such a way as to guarantee that their most ardent desire was to look attractive for members of the opposite sex, Victorian women were unable to discover and develop their potentialities, and thus were unfree. In a move reminiscent of republicanism, Mill maintains that the absence of freedom cannot be identified with interference tout court. Ultimately, any time lived in the absence of guarantees against arbitrary interference constitutes a time of non-freedom. In order to achieve freedom, people need to be protected from arbitrary interference so they can critically examine the customs that prevail in their society and experiment with different lifestyles. This intelligent following of custom, which can be identified as the ethical dimension of Millian liberty, allows each citizen to decide which experiment in living maximises the development of his or her character. The resistance against oppression and epistemic injustice that Mill deems indispensable for liberty also has a more political dimension, which can be observed in the proportional representation scheme proposed in *Considerations on Representative Government*. The public articulation of the plight of oppressed minorities in the representative assembly increases their social standing as citizens and, moreover, can produce alternative vocabularies and tactics that help them resist the oppressions perpetuated in civil society.

I

Resistance has made us what we are, and will yet make us what we are to be

Mill*, The Subjection of Women*

Though much has been written on Millian liberty, no scholar thus far has offered an explicit account of the entwinement of liberty with resistance in Mill's political philosophy. It is likely that what Iain McDaniel (2018) said of Benjamin Constant and Alexis de Tocqueville also explains the silence surrounding Mill's concept of resistance. Perhaps the reason scholars working on "resistance theory" nowadays tend to neglect Mill is because they do not expect

a nineteenth-century "liberal" philosopher to qualify as a "significant contributor" when it comes to understanding the importance of resistance for politics (McDaniel 2018, 433).[1] By exploring the connection between freedom and resistance in Mill's political thought, this chapter argues that Millian liberty entails resisting the oppression caused by epistemic injustice.

Since the publication of Miranda Fricker's *Epistemic Injustice: Power and the Ethics of Knowing*, philosophers have devoted increasing attention to the topic of "epistemic injustice", an expression used to denote any "wrong done to someone specifically in their capacity as a knower" (Fricker 2007, 1). Nevertheless, as Fricker herself acknowledges – and as Mill's works testify – epistemic injustice had been scrutinised by scholars before the concept 'epistemic injustice' was coined. In what follows, I contend that Mill's conception of liberty seeks to resist and overcome the oppression caused by epistemic injustice. The resistance Mill associates with freedom comprises two dimensions: ethical and political. In its ethical dimension, resistance against oppression caused by epistemic injustice involves what Mill calls "an intelligent following of custom" (CW XVIII, 263).[2] In its political dimension, it involves a proportional representation scheme that sustains a conflictive and polyphonic deliberative setting in the representative assembly, one in which the different social perspectives comprised in the demos are expressed and taken into account.

II

Published in 1869, *The Subjection of Women* is remarkable for advancing a conception of liberty as non-subjection (Urbinati 2002, ch. 5). According to Mill, women were unfree because they were subjected to male domination, which provoked epistemic injustice. "It is only a man here and there who has any tolerable knowledge of the character even of the women of his own family. I do not mean of their capabilities; these nobody knows, not even themselves, because most of them have never been called out" (CW XXI, 278). Nineteenth-century women experienced epistemic injustice because the oppressive milieu where they lived precluded them from knowing their potentialities (Zakaras 2009, 139). "All women are brought up from the very earliest years in the belief that their ideal of character is . . . not

[1] In Howard Caygill's (2013) *On Resistance* and José Medina's (2013) *The Epistemology of Resistance*, for instance, Mill is not cited. In the special issue 'Resistance in Intellectual History and Political Thought', published in 2018 by *History of European Ideas*, Mill's thinking on resistance is also ignored.

[2] Following common practice among Mill scholars, references to *The Collected Works of John Stuart Mill* are written as follows: CW VII, 313 for *Collected Works*, volume VII, page 313.

self-will and government by self-control, but submission and yielding to the control of others" (CW XXI, 271). Women not only lacked the opportunity to know and cultivate the capabilities that would develop their character to its utmost splendour, but also were taught never to explore and pursue such knowledge.[3]

Though Mill does not offer a precise definition of oppression, an attentive reading of *Subjection* reveals that oppression is present whenever citizens' capacity for developing themselves is deliberately dwarfed. Put differently, an individual is oppressed when she is deliberately impeded to freely cultivate her capacity for self-development. That can happen through violence, of course, but also through more subtle mechanisms – such as deformed desires and epistemic injustice. A woman is oppressed not only when she is subjected to physical force, but also when society shapes her psyche in such a way as to guarantee that her strongest desire is to look attractive for members of the opposite sex. Rather than simply curtailing behaviour, oppressive power can be *productive* and encourage certain lines of conduct by dint of the internalisation of oppressive norms. Oppression is perpetuated by *external* as well as *internal* forces.

Because it is less visible and involves the active participation of the oppressed subject, psychological oppression can be much harder to combat than physical oppression.[4] This is something Mill highlights in the introduction to *On Liberty*: in a way, psychological oppression is more difficult to confront than physical oppression because, by "penetrating much more deeply into the details of life and enslaving the soul itself", psychological oppression makes the formulation of resistant tactics more difficult (CW XVIII, 220). When oppression is transmitted solely on the basis of physical violence, there is only one way to resist, which is quite straightforward: just exert a contrary force. But when oppression is entrenched in one's desires, how is one to resist?

III

Mill's conception of liberty as non-subjection shows that being under the arbitrary will of somebody else, by itself, suffices to attest to the absence of freedom and the presence of

[3] The concept of character deployed by Mill is further clarified in the next section. On the centrality of the discourse on character in Victorian political thought, see Stefan Collini (1985).

[4] My understanding of the differences between psychological and physical oppression subscribes to Ann E. Cudd's (2006). I take epistemic injustice to be an example of psychological oppression.

oppression. In a move reminiscent of republicanism, Mill maintains that despotism, arbitrary subjection and tyranny – in short, the absence of freedom – cannot be identified with interference tout court.[5] Ultimately, any time lived in the absence of guarantees against arbitrary interference constitutes a time of non-freedom.

From that perspective, a woman living under the dominion of a magnanimous husband or father who never interferes with her conduct remains unfree. Magnanimousness describes the kind behaviour of someone who has the power to interfere with another's conduct in a whimsical manner, but who decides not to. The problem is that if the good will of the master subsides, magnanimousness disappears. When a woman living under the shadow of arbitrary subjection comes to terms with her predicament, she starts policing her words and deeds in such a way as to avoid arousing the master's anger – which, her greatest efforts notwithstanding, remains a very imperfect way of dodging actual interference, for nothing guarantees the master will not suddenly become cranky and decide, without any reason, to oppress her.

According to Mill, a society where arbitrary subjection is possible fosters sycophancy, servility and duplicity among its members (CW XXI, 279). Maintaining oppression over a long period of time is only possible with the active engagement of the oppressed. An arbitrary state of affairs can only reproduce itself systematically on the condition that people act in a way compatible with it. A regime that needs to resort to violence day in and day out in order to appease popular resistance is doomed to be short-lived. The capacity to shape citizens' desires and psyche in a way that co-opts them as active participants in their own oppression greatly facilitates the existence of an oppressive and arbitrary regime.

Mill's conception of liberty as non-subjection is linked to the power of formulating desires autonomously.[6] In the conclusion of *Subjection*, Mill affirms that the polities he is most supportive of

> are those which have most strongly asserted the freedom of action of the individual – the liberty of each to govern his conduct by his own feelings of duty, and by such laws and social restraints as his own conscience can subscribe to. (CW XXI, 336)

[5] On the republican features of Mill's conception of liberty as non-subjection, see Gustavo Hessmann Dalaqua (2018). Mill identifies himself as a 'republican' thinker in CW XXVI, 359.

[6] On the connection between liberty and autonomy in Mill, see Wendy Donner (2008), John Gray (2002) and Mauro Cardoso Simões (2008).

Freedom of action in this passage is identified with self-government, that is, with the capacity to regulate one's conduct by feelings and laws that somehow are one's own. Freedom of action is thus linked to what Mill had described as 'character' in *On Liberty*:

> A person whose desires and impulses are his own – are the expression of his own nature, as it has been developed and modified by his own culture – is said to have a character. One whose desires and impulses are not his own, has no character, no more than a steam-engine has a character. (CW XVIII, 264)

To guide one's conduct by desires and impulses of one's own – in other words, to have a character – does not entail immuring oneself from social intercourse. Pace Willaim Gairdner (2008, 11, 14), Millian liberty should not be conflated with atomism or individualism.[7] As the passage above suggests, the constitution of character arises out of the interaction between one's nature and one's culture. The thesis that the formation of character cannot do without social intercourse is further clarified when Mill associates freedom of action with "an intelligent following of custom, or even occasionally an intelligent deviation from custom" (CW XVIII, 263). The critical lifestyle Mill relates to freedom and character is not against custom per se, though it is at odds with "a blind and simply mechanical adhesion to it [i.e., custom]" (CW XVIII, 263).

The intelligent following of custom is a form of resistance against internalised oppressions that allows citizens to autonomously formulate their own desires. By being urged to critically examine social customs, a woman who was taught that her only desire should be to look charming to men can by and by realise there are other 'experiments of living' she can pursue besides that of an obedient and submissive wife (CW XVIII, 281). The intelligent following of custom and its concomitant engagement with different lifestyles incite the oppressed to resist epistemic injustice, because they bring to the fore the fact that the hegemonic narrative of how to live, act and desire is only one among several others. By following social customs intelligently, citizens can know what kind of lifestyle they might want to pursue.

The intelligent following of custom and its attendant engagement with different experiments in living constitute the ethical dimension of Millian resistance. Since both practices are connected with the formation of character, they qualify as ethical because, as Mill observes in *A System of Logic*, what he calls 'character' is nothing but a translation for the ancient term ethos (CW VIII, 869). As the next section highlights, the ethical and political

[7] As Catherine Audard (2009, 86-7) pointed out, it was precisely because Mill wanted to distance his philosophy from individualism that he started using the term 'individuality'.

dimensions of Millian resistance can be distinguished from one another, inasmuch as the latter focuses more on traditional political institutions such as the representative assembly.[8]

IV

In *Considerations on Representative Government*, Mill seeks to understand how 'collective resistance' can be preserved in the context of mass societies (CW XIX, 419). As he explains in chapter seven of this book, the great

> difficulty of democratic government has hitherto seemed to be how to provide, in a democratic society, what circumstances have provided hitherto in all the societies which have maintained themselves ahead of others – a social support, a *point d'appui*, for individual resistance to the tendencies of the ruling power. (CW XIX, 459)

In the Middle Ages, individuals were able to resist arbitrary power by organising themselves as members of a larger group that, as such, needed to have its voice taken into account by the government (CW XX, 292-93). This scenario changed with the advent of industrialisation and population growth. As Mill declared in *On Liberty*, "at present individuals are lost in the crowd" (CW XVIII, 268). With the spread of urbanisation and the weakening of membership in political groups, resistance became increasingly difficult.

Mill thinks the solution to such a predicament lies in proportional representation. According to him, elected politicians should represent social groups, not isolated individuals (CW XIX, 405). If representative government is to be truly democratic, it is imperative that the representative assembly expresses the social perspective of every political group comprised in the demos.[9] A proportional representation scheme respects that imperative because, unlike the first-past-the-post voting method, it does not allow only representatives who collect more than fifty percent of the votes to be elected. The winner-takes-all system leads to a falsified representative democracy in Mill's view because it offers no guarantee against the tyranny of the majority. Endorsing Pericles' view of democracy, Mill submitted that, rather than being identified with majoritarianism tout court, democracy should be described as

[8] This is not to deny that the ethical dimension of Millian resistance is of political relevance; the ethical and political dimensions are, indeed, mutually reinforcing. That does not mean, however, they cannot be differentiated.

[9] The association between representation and social perspective became prominent in contemporary studies on representation mainly due to Iris Marion Young (2000). The similarities between Young and Mill are interesting, yet to approach them here would lead us too far afield. For a good comparison between both writers, see Wendy Donner (2016).

the regime where the rule of the majority goes in tandem with the recognition and appreciation of human diversity (see CW XI, 319 and Thucydides 1982, 109ff). More than a political regime, representative democracy for Mill refers to a type of society where citizens' differences are a reason for celebration, not condemnation.

The reason proportional representation helps oppressed minorities resist epistemic injustice is twofold. For one thing, the mere fact of having the perspective of an oppressed minority expressed in Parliament increases its social status. It means the perspective of this oppressed minority should be taken into account by the government when laws are being made. The representative of the oppressed minority can then reveal to the wider public that many assumptions about the group she represents are inaccurate and demeaning. This revelation, along with her power to propose bills that tackle the epistemic injustice perpetuated against the group she represents, allows resistance to take place.

Moreover, minorities are more encouraged to resist the multifarious social sources of epistemic injustice that oppress them when they have someone expressing their perspective in the representative assembly. The public articulation of their plight by their representative in the face of political opponents – recall Mill's depiction of the representative assembly as an "arena where opposing forces should meet and fight out their battle" (CW XXV, 1106) – arms minorities with vocabularies and tactics that help them resist oppression. By doing so, it allows minorities to develop themselves freely.

Acknowledgement

This chapter was written while the author received financial support from São Paulo Research Foundation, FAPESP grant # 2015/22251-0.

References

[1] Audard, Catherine. 2009. *Qu'est-ce que le libéralisme? Éthique, politique, société*. Paris: Gallimard.

[2] Caygill, Howard. 2013. *On Resistance*. London: Bloomsbury.

[3] Collini, Stefan. 1985. "The Idea of Character in Victorian Political Thought." *Transactions of the Royal Historical Society* 35: 29-50. DOI: 10.2307/3679175.

[4] Dalaqua, Gustavo Hessmann. 2018. "John Stuart Mill's Republican Feminism." *Kalagatos* 15 (2): 14-33. DOI: 10.23845/kgt.v15i2.725.

[5] Donner, Wendy. 2008. “Autonomy, Tradition, and the Enforcement of Morality.” In *Mill’s* On Liberty: *A Critical Guide*, edited by C. L. Ten, 138-64. Cambridge: Cambridge University Press.

[6] _____. 2016. “Mill on Individuality.” In *A Companion to Mill*, edited by Christopher Macleod and Dale E. Miller, 425-39. Oxford: Wiley Blackwell.

[7] Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press.

[8] Gairdner, William. 2008. “Poetry and the Mystique of the Self in John Stuart Mill.” *Humanitas* XXI (1-2): 9-33.

[9] Gray, John. 2002. “Mill’s Conception of Happiness and the Theory of Individuality.” In *J. S. Mill’s* On Liberty *in Focus*, edited by John Gray and G. W. Smith, 231-57. London: Routledge.

[10] McDaniel, Iain. 2018. “Resistance in Intellectual History and Political Thought.” *History of European Ideas* 44 (4): 397-403. DOI: 10.1080/01916599.2018.1473955.

[11] Medina, José. 2013. *The Epistemology of Resistance*: *Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford: Oxford University Press.

[12] Mill, John Stuart. 1963-1991. *The Collected Works of John Stuart Mill*, 33 vols. Toronto: University of Toronto Press / London: Routledge and Kegan Paul.

[13] Simões, Mauro Cardoso. 2008. *John Stuart Mill & a liberdade*. Rio de Janeiro: Jorge Zahar.

[14] Thucydides. 1982. *História da guerra do Peloponeso.* Translated by Mário da Gama Kury. Brasília: Editora da Universidade de Brasília.

[15] Urbinati, Nadia. 2002. *Mill on Democracy*: *From the Athenian Polis to Representative Government*. Chicago: The University of Chicago Press.

[16] Zakaras, Alex. 2009. *Individuality and Mass Democracy: Mill, Emerson, and the Burdens of Citizenship.* Oxford: Oxford University Press.

[17] Young, Iris Marion. 2000. *Inclusion and Democracy*. Oxford: Oxford University Press.

Protagoras, Political Economy, and the Art of Politics: J.S. Mill in the 1830s

Stephen Engelmann, University of Illinois at Chicago, USA

Abstract

This essay reads Mill's early abstract of Plato's *Protagoras* with his work on political economy, social science, and the art of politics. *Protagoras* is notable for the sophist's defense of Athenian democracy in its division between expert advice and expert rule. And it is notable for Socrates's insistence on the identity between knowledge and virtue in a proto-utilitarian doctrine linking pleasure and the good, and advocating the adequate measurement of future pains and pleasures in relation to present ones. How does Mill incorporate moments of democracy and economic rationality into a broader art and science of government? My contention is that Mill's early work on political economy, where he develops the outlines of his logic of the moral sciences, amplifies resonances of the non-democratic elements in his thinking.

My reading contests Nadia Urbinati's powerful interpretation of Mill as agonistic democrat. Yes, Mill is an admirer of the "Sokratic" as opposed to the "Dogmatic" Plato. But this does not entail a defense of Protagoras's democracy so much as a defense of open and vigorous scientific inquiry. Urbinati too readily dismisses the importance of Mill's consistent analogizing of politics to medicine, which suggests less a contest over the direction of our common life than a contest over a cure for what ails us.

What is the best medicine? On the one hand, political economy according to Mill concerns only one aspect of the conduct of individual and social life. But on the other hand, his early writing on the subject sketches a view of the social that, combined with an admirably dynamic, progressive, and capacious approach to political economy, cedes a substantial role to economic rationality and to economic science. In this way the great liberal theorist unwittingly contributed to what William Davies calls our "disenchantment of politics by economics", and thus to a narrowing of the means and ends of individual and collective life.

Introduction

My work here is positioned between two projects. The first, a manuscript under revision, looks to John Stuart Mill and Charles Darwin as newly relevant for any consideration of the revival of naturalism in the contemporary social sciences and humanities. There I read Mill as a developmental naturalist thinker preoccupied with character—in particular with economic rationality and expansive sympathy—as the alpha and omega of improvement, or progress; and I read Darwin the same way, and both in counterpoint to Bentham, who was notoriously uninterested in character. The second project, a short book on economic rationality, looks to Mill among others as a classical thinker who demonstrates that economic reason is no mere instrumental reason, but instead a consequential ethics and quasi-politics. Both projects are aimed at showing how pervasive and how vexed what I along with others call the utilitarian art and science of government is in the nineteenth century and

today. As a political theorist, my interest is to read Mill neither as a promoter of nature or as a promoter of nurture, nor as a normative thinker or as the purveyor of a value-free social science (both of which are anachronistic oppositions), but as a supremely relevant biopolitical thinker interested primarily in the conduct of conduct. I hope that this perspective can shed a bit of light on difficult questions of Mill interpretation. First, in connection with *Principles of Political Economy*, is economic rationality for Mill a principle bearing only on that aspect of our conduct that concerns wealth, as he maintains, or is it a kind of civilizational marker and goal, at least when understood in broader terms as providence, or temperate foresight ([1848] 1965a and 1965b)? Which is also to say, what relation does political economy bear to Mill's broader ambition for the "Social Science" of his *System of Logic* (Mill [1843] 1974, 875-8)? Is it one small piece having only to do with the production and distribution of material wealth, as he insists, or does it, as long as it integrates the all-important principle of population, hold the key to human progress? Second, in connection with politics, do Mill's ambitions for an art and science of government lead him away from politics understood as a conversation about common ends, or even as an exercise in persuasion or interest advancement among equals, and a step or two towards scientific administration as a substitute? Mill does suggest in the *Logic* that the corresponding art to social science is the "art … of politics" (Mill [1843] 1974, 877), and it is the art of politics that plays a small but pivotal role in Plato's *Protagoras*.

A way in to these questions is provided by Nadia Urbinati's bracing revisionist interpretation of Mill as agonistic democrat (Urbinati 2002). Urbinati sets herself directly against a reading of Mill that I find quite compelling. On my reading, Mill skirts perilously close to embrace of a Platonic rule of experts in the slippage from analogy to identity that he promotes between politics and medicine. We see this analogy in a few places in the early work: Urbinati herself mentions the 1835 review of Tocqueville's *Democracy in America* (Urbinati 2002, 48). In the same year, Mill writes in his "Rationale of Representation" that the "parallel holds exactly between the legislator and the physician." The point here is that "the people themselves, whether of the high or the low classes, are, or might be, sufficiently qualified to judge, by the evidence which might be brought before them, of the merits of different physicians, whether for the body politic or natural." Yet, "it is utterly impossible that they should be competent judges of different modes of treatment. They can tell that they are ill; and that is as much as can rationally be expected from them. Intellects specially educated for the task are necessary to discover and apply the remedy" (Mill [1835] 1977, 40). Note, as Urbinati does, how this distinction anticipates Mill's discussion of the capacities and limits of representative assemblies in ch. V of *Considerations on Representative Government* (Mill [1861] 1977, 433; Urbinati 2002, 46-7). I could add a reminder that Mill in every edition of the *Logic* suggests that medicine is the art to which politics is "most nearly allied," and that

he criticizes past thinkers for being insufficient to the task of social inquiry because they attempted "to study the pathology and therapeutics of the social body, before they had laid the necessary foundation in its physiology; to cure disease without understanding the laws of health" (Mill [1843] 1974, 876-7). Regarding Plato, Urbinati acknowledges Mill's longtime infatuation with Platonic texts, but argues that this is really an infatuation with the "Sokratic" as opposed to the "Dogmatic" Plato (Urbinati 2002, 7-8, 49-50). Plato's Socrates provides us not with answers, but with the very model of how to conduct inquiry and to construct knowledge. And the encounters in Plato's dialogues model the equality that Mill prized in Athenian life, and that he worked to help bring about through his political reform efforts. Urbinati argues that Mill's political project, most bracingly communicated in his *Subjection of Women* ([1869] 1984, 259-340), is consonant with a long-running republican tradition that prizes the struggle against domination among an association of free equals (see also Skinner 1998, ix).

Urbinati even goes so far as to suggest that Mill's model is the Athenian politics portrayed in Plato's *Protagoras*, and justified by the Protagorean myth: a model that limits expertise to technical questions and excludes it from the general questions of practical politics (Urbinati 2002, 51-3). Although I think that Urbinati is correct, along with others, to identify Mill with the spirit of freedom as non-domination, I insist that Mill himself doesn't see any inconsistency between this spirit and a broadly therapeutic approach to politics that runs counter to Urbinati's more democratic reading.

I Protagoras

Protagoras was the first of the Platonic dialogues that Mill published in the 1830s, in translations that he initially rendered sometime after his breakdown in the 1820s. Mill writes in his introduction that "there are, probably, in this kingdom, not so many as a hundred persons who ever *have* read Plato, and not so many as twenty who ever *do*" (Mill [1834] 1978, 40). Even if exaggerated, the low estimate speaks to the sea change conducted by contemporary radicals, including Mill himself, in canonizing Plato into Anglophone letters; many since have read Plato largely because of their efforts.

Protagoras is an early dialogue between Socrates and Protagoras, a noted Sophist, or paid teacher of young men with political ambitions. When Socrates asks what expertise Protagoras has to offer these young men—what it is, exactly, that he teaches—Protagoras answers that he teaches *politikê technê* or civic and ethical know-how (Plato 2005, 16-7), what Mill

at one point translates quite literally as the "art of Polity" (Mill [1834] 1978, 49). The dialogue is ostensibly about what virtue is and whether it can be taught. Socrates suggests early on that the Athenians must think that the art Protagoras teaches cannot be taught, because in their deliberations they listen to experts and not to others when technical questions are before them with the ends already established (for example, how do we build a ship, or a plaza?). But when the questions before them consider the direction of the city (for example, should we even build a plaza?) anyone can rise and advise the city without distinction; everyone in the assembly is understood to have standing in these matters (Plato 2005, 28; Mill [1834] 1978, 48). Protagoras explains, by way of a myth, that there is no contradiction. According to the myth capacities for civic virtue allow people to live together in cities; these gifts were given, in addition to the illicit Promethean gift of fire, to humans out of pity for their natural weakness in relation to other animals. Protagoras suggests that the way *politikê technê* is both democratically distributed and learnable is akin to language. When children or adults say or do things wrongly instead of rightly they are corrected by parents or peers; we learn civic virtue in the same way that we learn speech, with some excelling in it or better able to teach others, but with a broadly democratic distribution of talents and skills (Plato 2005, 29-30; Mill [1834] 1978, 50-1).

Protagoras is of special interest also to the economic rationality at the heart of political economy because, as Mill notes in the first paragraph of his article on *Utilitarianism*, there "the youth Socrates listened to the old Protagoras, and asserted (if Plato's dialogue be grounded on a real conversation) the theory of utilitarianism against the popular morality of the so-called sophist" (Mill [1861] 1969, 205). Near the end of the dialogue Socrates argues against Protagoras and for the identity of the virtues with a proto-Benthamic statement of the "knowledge of measurement." According to Socrates here pleasure is good, all activity is ultimately pleasure-seeking, all pleasures and pains are commensurable, and when we act rationally and virtuously we adequately weigh the values of present and future pleasures and pains against one another (Plato 2005, 72-3; Mill [1834] 1978, 58-9). Much has been written by philosophers on whether this is really a Socratic doctrine (that's a minority position, of course), or is something deployed by him in order to introduce a firm distinction between appearance and reality in ethical matters, and to attack a popular morality invested in the idea of akrasia, or weakness of the will. On this latter point most commentators, including Mill, agree that the dialogue is serious here about the equation between virtue and knowledge; the idea of measurement is one way to set up this equation and to defend the bracing doctrine that no one ever does wrong except out of ignorance (Mill [1834] 1978, 61). In any case, the ending is, as Socrates himself points out, notably inconclusive, for example because he and Protagoras have changed places about whether virtue can be taught (Plato 2005, 79-80; Mill [1834] 1978, 60). And of course we can add

that taking this proto-utilitarian doctrine seriously has implications both for the status and scope of economic rationality—for the place and prominence of Mill's understanding of (im)providence—and for evaluating the Athenian position on democracy and expertise with which *Protagoras* begins.

In an excellent work of classics scholarship, Alexandra Lianeri accuses Mill of "effacing Socratic irony" in his translation of *Protagoras* (Lianeri 2007). She focuses first on the end of the dialogue, which Mill softens with summary language, missing how Socrates sharply ridicules, for its production of absurdities, the kind of philosophizing that he and Protagoras have engaged in (Mill [1834] 1978, 60; Plato, Lianeri 2007, 173). And she notes how Mill introduces a sharper distinction between theory and practice and between truth, persuasion, and power than the original allows, and how Mill can reduce the political to the social, as we see in his references to "social virtues" in the dialogue, foreshadowing the "social body" of the *Logic* (Lianeri 2007, 175-80; Mill [1834] 1978, 49; Mill [1843] 1974, 876). Her conclusions have serious implications for Urbinati's thesis. Lianeri's Socrates is political, in a far more fundamental respect than Mill's is.

II Political Economy and Politics

The reduction of the political to the social happens with some frequency in Mill's work, especially in these early writings. And indeed, Mill's 1836 "On the Definition of Political Economy" (Mill 1967, 309-39) performs one of the more dramatic of such reductions. There is a lot to be said about this remarkable essay; it is here, in my view, that Mill first formulates the distinctions and ambitions expressed for a comprehensive, causal social science in Book VI of his *System of Logic*, on the Logic of the Moral Sciences. On the one hand, "On the Definition of Political Economy" seems to make a lot of room for particular contingencies and for motives alternate to political-economic ones in the study of the social. On the other hand, its way of understanding scientific a priorism and its idea of the connection between the abstract and the concrete is such that that which falls outside of its political-economic models is understood as "disturbing cause" akin to "friction in mechanics" (Mill [1836] 1967, 330), and, if not fully assimilable to the science, can perhaps be put together with it, at least in principle, to build a fully comprehensive understanding of social science or what the essay calls "social economy, speculative politics, or the natural history of society" (Mill [1836] 1967, 320). Mill writes that "the science of social economy embraces every part of man's nature, in so far as influencing the conduct or condition of man in society." It may "be termed speculative politics, as being the scientific foundation of practical politics, or the art of government, of which the art of legislation is a part" (Mill [1836] 1967, 320-1).

In an admiring criticism of Jean-Baptiste Say Mill accuses Say of supposedly conflating the part that is political economy with the whole of social science. But "this large extension of the signification of [Political Economy] is countenanced by its etymology" (Mill [1836] 1967, 321). In the first edition of the essay Mill goes on to commit the following bald anachronism: "[Oikonomia politike], the economy of the [polis], or commonwealth, must originally have meant the whole of the laws or principles which determine the working of the social machine" (Mill [1836] 1967, 321n). Thus Say's mistake—to conflate political economy with all of social science—*is* a mistake, because political economy is actually "concerned with [man] solely as a being who desires to possess wealth;" it "does not treat of the whole of man's nature as modified by the social state, nor of the whole conduct of man in society" (Mill [1836] 1967, 321). But it is an understandable mistake, because the label "political economy" evokes an actually existing "social machine." And politics and the political art are nothing other than the art of superintending this social machine, as properly informed by a comprehensive science of it.

I would argue that, even though the language of 1836 is dropped, some of the spirit of this position is maintained through Mill's late *Considerations on Representative Government*. There Mill famously opens by rejecting the polar alternatives of two sorts of "political reasoners," those who think of politics as able to do anything and simply transform the world, and those who think of it as able to do nothing but follow the way of the world (Mill [1861] 1977, 374-5). Instead, just as medicine cannot cure just anything, it can cure some things, and it does this with proper knowledge of physiological causes and effects so that it can manipulate some causes to produce alternate effects. The question is how Mill's insistence in *Considerations* on the ultimate sovereignty of the people relates to his insistence on there being a truth, open to discovery, of what would most improve the people at any one time. And this has implications for the meaning of what Urbinati rightly notes is Mill's admiration for the Socratic as opposed to the dogmatic Plato. What Mill takes from Socrates, I am suggesting, is—as he emphasizes in his introduction to the early Plato abstracts referencing Schleiermacher—philosophical method: the elenchus or Socratic dialectic as a way to pursue truth (Mill [1834] 1978, 41). In this way the elenchus models not a game of common ends-seeking or persuasion or interest-advancement or other mode that we may think appropriate to politics, but an intellectual search among equals to best figure out the science to which we need to subordinate our art of government. And indeed, in the *Autobiography*, Mill praises the elenchus and notes its profound influence on his own and his father's work (Mill [1873] 1981, 25). On my reading then the problem with expert rule is, as it is in *On Liberty*, the fact that no one is infallible, and the fact that leaving one's affairs to others as opposed to vigorously engaging them stunts individual and collective growth (Mill [1859] 1977, 229-43 and passim); the problem is not in the underlying assumption that there is an

aggregate interest of the "social machine" and its advancement, and a science according to which that interest can be properly advanced.

What does this understanding of politics mean for political economy? If it is true that in our time the truth of politics remains, as the Clinton presidential campaign famously said, "the economy, stupid!" then we might find the claim of the political scientist Timothy Mitchell arresting: Mitchell claims that there really was no such thing as "the economy" until the early twentieth century. Which is to say, economy remains for the most part until this time a way of doing things, and not itself a thing (Mitchell 2005). Mitchell's contribution, informed by Science and Technology Studies, is to insist that we not think of social science as describing a world that exists prior to or outside of it, but that we recognize social science as one among many sociotechnical forces that helps to build the world of things it studies (Mitchell 2007); following the work of Michel Callon and others (1998), he insists that economics is performative, and that one of its surprisingly recent performances is "the economy" itself (he dates its reification from the 1930s) (Mitchell 2005, 126). A perfect illustration of what Mitchell is arguing is to be found in economist Paul Krugman's post-financial crisis attack on the U.S. macroeconomics profession. There Krugman lambastes mathematical macroeconomics for its attraction to beauty over truth, noting how this contributed to it completely missing the crisis. Instead, economics needs to recognize the essential "messiness" of the processes it tries to capture (2009). Mitchell would remind us that this focus on the problem of representation of a messy reality only serves, through familiar rhetorical conventions of literary realism, to reinforce and reinscribe "the economy" as a Millian social machine in need of expert superintendence. And this expert superintendence of a general interest works to elide the political commons.

It might seem that Krugman's Keynesian perspective is very different from the Austrian one; after all, Friedrich Hayek precisely denied that there was such a thing as an economy; this implies elements of organization and purpose that the great "catallaxy" of market society lacks, to its credit (1967, 173). But we should recall that Hayek's early work was an attack on the mathematical economics of *his* time for its role in central planning, and so his work conjured Krugman's messiness as a real and virtuous spontaneous order of dispersed knowledges activated and coordinated by the marvelous power of prices (Hayek 1945). And whenever Hayek got specific, his administrative orientation—and in particular an economic approach to law as utilitarian regulatory policy—could slip out. So we find in *Law, Legislation, and Liberty* the following: "The task of rules of just conduct can thus only be to tell people which expectations they can count on and which not. ... Which expectations ought to be protected must ... depend on how we can maximize the fulfillment of expectations as a whole" (Hayek 1973, 102-3). In Hayek as in Krugman, a social machine is conjured that is

always only imperfectly represented by its scientists, who can deliberate in Socratic fashion as to how best to administer it.

We can see the germ of this contemporary approach in an example from Mill's *Considerations on Representative Government*, even as it is likewise an example of Mill displaying the remarkable empathy of which he was so capable. After praising the governing classes of his time for their comparative attention to the interests of the poor, he reminds us of the costs of excluding whole sections of the governed from representation. "Yet does Parliament, or almost any of the members composing it, ever for an instant look at any question with the eyes of a working man? When a subject arises in which the labourers as such have an interest, is it regarded from any point of view but that of the employers of labour?" Mill continues, "I do not say that the working men's view of these questions is in general nearer to truth than the other: but it is sometimes quite as near; and in any case it ought to be respectfully listened to, instead of being, as it is, not merely turned away from, but ignored." (Mill [1861] 1977, 405) It is the truth finally of the proper representation of the social that is our goal, and it is that truth, rising above all differences in perspective and interest, that should govern once we ascertain, through inclusion, what it is. Mill's art of polity as the scientifically informed superintendence of the social in pursuit of an aggregate interest overshadows politics as a conversation or struggle over the meaning and direction of our common life.

This should give us pause, as Mill himself understood better than most the danger of any narrowing of the ends of life: consider much of the tenor of *On Liberty*, and consider his contemptuous description of the US as carried through all the pre-Civil War editions of *Political Economy*. "They have the six points of Chartism, and they have no poverty: and all that these advantages do for them is that the life of the whole of one sex is devoted to dollar-hunting, and of the other to breeding dollar-hunters" (Mill [1848] 1965, 754n). But my point is that Mill's earnestly and sincerely held outlook is undercut by his own participation in setting the preconditions for the "disenchantment of politics by economics" (Davies 2014, 1-34), a disenchantment that has in turn played a significant role in the formation of our present predicaments.

References

[1] Callon, Michel, ed. 1998. *The Laws of the Markets*. Oxford: Blackwell.

[2] Davies, William. 2014. *The Limits of Neoliberalism: Authority, Sovereignty, and the Logic of Competition*. London: Sage.

[3] Hayek, Friedrich. 1945. “The Use of Knowledge in Society.” *The American Economic Review* 35 (4): 519-30.

[4] _____. 1967. *Studies in Philosophy, Politics, and Economics.* Chicago: University of Chicago Press.

[5] _____. 1973. *Law, Legislation, and Liberty, Volume I: Rules and Order.* Chicago: University of Chicago Press.

[6] Krugman, Paul. 2009. “How Did Economists Get It So Wrong?” *The New York Times Magazine*, September 2, 2009. https://www.nytimes.com/2009/09/06/magazine/06Economic-t.html

[7] Mill, John Stuart. 1965a. *Principles of Political Economy, with Some of Their Applications to Social Philosophy, Books I-II.* Vol. 2 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[8] _____. 1965b. *Principles of Political Economy, with Some of Their Applications to Social Philosophy, Books III-V.* Vol. 3 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[9] _____. 1967. *Essays on Economics and Society*. Vol. 4 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[10] _____. 1969. *Essays on Ethics, Religion, and Society.* Vol. 10 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[11] _____. 1974. *A System of Logic Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, Books IV-VI and Appendices*. Vol 8 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[12] _____. 1977a. *Essays on Politics and Society I*. Vol. 18 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[13] _____. 1977b. *Essays on Politics and Society II*. Vol. 19 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[14] _____. 1981. *Autobiography and Literary Essays*. Vol. 1 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[15] _____. 1984. *Essays on Equality, Law, and Education*. Vol. 21 of *The Collected Works of John Stuart Mill,* edited by J. M. Robson. Toronto: University of Toronto Press.

[16] Mitchell, Timothy. 2005. “Economists and the Economy in the Twentieth Century.” In *The Politics of Method in the Human Sciences: Positivism and its Epistemological Others*, edited by George Steinmetz, 126-41. Durham: Duke University Press.

[17] _____. 2007. “The Properties of Markets.” In *Do Economists Make Markets? On the Performativity of Markets*, edited by Donald MacKenzie, Fabian Muniesa, and Lucia Siu, 244-75. Princeton: Princeton University Press.

[18] Lianeri, Alexandra. 2007. “Effacing Socratic Irony: Philosophy and Technê in John Stuart Mill’s Translation of the Protagoras.” In *Socrates in the Nineteenth and Twentieth Centuries*, edited by Michael Trapp, 167-86. Aldershot: Ashgate.

[19] Plato. 2005. *Protagoras and Meno*. Translated by Adam Beresford. London: Penguin.

[20] Skinner, Quentin. 1998. *Liberty Before Liberalism.* Cambridge: Cambridge University Press.

[21] Urbinati, Nadia. 2002. *Mill on Democracy: From the Athenian Polis to Representative Government.* Chicago: University of Chicago Press.

J.S. Mill on Rebellion, Revolution and Reform[1]

Don A. Habibi, University of North Carolina Wilmington, USA

Abstract

The chapter examines John Stuart Mill's nuanced opinions on rebellions, uprisings, and revolutions. Was he a 'revolutionary'? I claim that he was, but not in the Marxist sense of the term. The young Mill was raised to be a radical reformer for Jeremy Bentham's utilitarian movement, and he was politically active his entire life. He wrote on historic events such as the American and French Revolutions. He also lived through such events as the 1848 revolutions, the Maori Wars, the Sepoy Mutiny, the U.S. Civil War, and the Morant Bay uprising. From his essays, letters, and his record in Parliament, we see that he supported these insurgencies, at least initially, or sided with oppressed people. However, he advocated gradual change based on deliberation, debate, and reason over force. Mill's apparent contradictions are explained by examining his perspectives on these events through the lens of utility and progress. Avoiding unnecessary violence is a means to minimizing pain, and it takes time for free and open discussion to prepare and persuade people to accept change. For Mill, the aim of history is human improvement. This is what constitutes utility in the 'large sense' and the truest way to maximize happiness.

In the opening sentence of "Mill's Epiphanies," Elijah Milgram asserts that at sixteen, J.S. Mill was an inspired activist bringing utilitarianism to the world. "John Stuart Mill was raised to be the Lenin of the revolutionary movement that we remember as utilitarianism, and whose members at the time were called the 'Philosophic Radicals' … he never became the Lenin of utilitarianism" (Milgram 2017). Like Lenin, Mill was on fire for *his* cause. An energized young man on a mission, Mill was groomed to be Jeremy Bentham's worthy successor—the chosen leader of his utilitarian radical reform movement. Both Mill and Lenin revised and adjusted their received doctrine with notable success. But, as I shall explain, the analogy breaks down. They were different kinds of revolutionary activists. In this chapter, I will argue that Mill steered a wiser course than Karl Marx's Russian disciple, and show there is much to learn from the contrast.

The Philosophic Radicals were products of the Enlightenment—secular, scientific, and rational, and their agenda was indeed radical for its time. Their leader, Jeremy Bentham, had extensive plans for changing the British legal, penal, educational, economic, electoral and

[1] An earlier version of this paper was presented at the Colloque "John Stuart Mill et la Révolution" Université Paris 1 – Panthéon/Sorbonne, May 25, 2018.

governmental systems to make them fairer, meritocratic, rational, and efficient. He opposed the legal and taxation privileges of the aristocracy, packing juries, restrictions on usury, the established church, the monarchy, and colonialism, and supported universal suffrage, the secret ballot, humane prisons, a league of nations, and other progressive reforms. His revolutionary program aimed to redefine the ruling class and prepare the conditions for when the working class and rural poor have power.

Bentham and the Radicals were optimistic because they *believed* in their utilitarian worldview, which was informed by up-to-date theories of human nature, psychology, education, and governance. They knew that their agenda would take more than one generation to implement, and so Mill was dedicated and educated to fight for the cause. His father, James Mill, was a recognized expert in education and psychology, eager to test out his learning theory on his firstborn son. John Stuart received a most impressive education from his father, supported by Bentham his godfather, along with advanced tutoring from several Benthamites, among them David Ricardo, John Austin, Francis Place, and George Grote. The designated heir of the movement was home-schooled in isolation, and trained to be the exemplar of a rational utilitarian man. The experiment produced one of the most remarkable intellectual polymaths in history.[2] Although Mill rebelled against the confines of his education, he appreciated and made good use of his considerable intellectual talents throughout his life.

Bentham, Mill, Marx, and Lenin all foresaw that the rich will lose their power to the masses. They understood they lived during a transitional period in history, and that modern capitalism, industrialization, and urbanization, would eventually lead to majority rule. Bentham, Mill, and Marx lived in London where the Industrial Revolution was unfolding. However, the utilitarians foresaw that the oppressed masses were just as prone to abusing power as everyone else. There was an urgent need to educate and uplift the population destined for power. They championed the interests of the working class, but did not regard them as responsible and prudent. On the contrary, the masses were susceptible to serious deficiencies. Something must be done for the uneducated, functionally illiterate, creatures of habit who are easily manipulated and driven by narrow egoism, crass hedonism, and tribalism. Citizens needed to mature and understand their responsibility to promote the greatest happiness. Mill worried about chaos and the 'tyranny of the majority.' The people had to appreciate the function of free expression, tolerance, and collective decision making in a liberal democracy. It is good the people will have power, only if they wield it with care.

[2] See Mill's *Autobiography*, CW I, and Habibi 2001, Ch. 1.

Milgram is not wrong to call the teenage Mill a "revolutionary." Inspired by the moderate Girondists of the French Revolution, he fantasized participating in a revolt against the English ruling class (CW I, 64-67).[3] But such fervor was short lived. Mill's views matured. At age 20, he experienced his 'mental crisis,' which triggered his break with orthodox Benthamism and cooled his devotion to the movement somewhat. He declared his independence, but still held fast to the foundations of the Benthamite cause. Young Mill's enthusiasm further waned as he increasingly understood that revolutions were rare and usually failed (Goldstone 2014). He learned that power corrupts (CW XXI, 288-9, 95-96, 320-25), and that even idealistic visionaries, saints, and the poor lose their scruples when they attain power. Thus, political structures must be set up to prevent or minimize the damage caused by abusive power. He endorsed the French Revolutions of 1830, 1848, and 1870, but he was bitterly disappointed by their outcomes. Even well-justified good revolutions are corrupted by overzealous politicians, ideologues corrupted by power, ill-informed mobs, and foreign invaders.[4]

Mill saw the problem as a lack of foresight under conditions of uncertainty. With many actors, there are countless ways for things to go wrong. Making realistic assessments and plans requires clear thinking plus some good timing and luck. We do not know the future, and we often must react without the luxury of time. Revolutions are like war: in all likelihood, things will spin out of control in a torrent of unforeseen consequences. Mill understood that, as a general rule, organic, careful, incremental change was safer and surer than rapid or forceful revolutionary change. Of course there are exceptions and there are no guarantees. Nonetheless, to the extent it is possible, peace and order are better than chaos and war.

Mill and Lenin were both activists, but they operated with different time frames. Mill took the patient, long term view. He was a first rate historian, well versed in ancient as well as

[3] He writes that the 1789 revolution "might easily happen again; and the most transcendant glory I was capable of conceiving, was that of figuring, successful or unsuccessful, as a Girondist in an English Convention." The word 'convention' is defined in this context, as: "A meeting of Parliament without a summons from the sovereign." Oxford English Dictionary,
https://en.oxforddictionaries.com/definition/convention

[4] For example, the 1789 revolution was welcomed by Bentham and the Radicals (and it inspired the juvenile Mill). Unfortunately, the moderate Girondins were eliminated by the ruthless Jacobins, resulting in the Terror, the French Revolutionary Wars, and hundreds of thousands of deaths. This led to Napoleon crowning himself emperor and wars resulting in millions of deaths. When Mill first visited France, the Bourbons were back on the throne. The 1830 Revolution led to the Louis Phillippe monarchy. Late in Mill's life, things fell apart again for France with invasion and shocking defeat by Prussia.

modern history. He was also a cutting edge historiographer. He was convinced that historical development occurs in stages, and taking major shortcuts will fail. The requirements and possibilities for change depend on the circumstances: "the proper functions of a government are not a fixed thing, but different in different states of society; much more extensive in a backward than in an advanced state" (CW XIX, 383).[5] Changes work out better when properly managed. Without oversight there is chaos. Marx and Lenin believed that they understood the iron laws of history. Class warfare is a reality, and the proletariat will emerge victorious. Patience was not a virtue; thus, educating, organizing, and agitating to hasten the process is for the good. Mill was unaware of Marx's dialectical materialism and the historical determinism inferred from it. Mill relied on his own knowledge of history. This informs his trepidation with revolution. So much can go wrong that caution is advisable. Slow and steady wins the race if we can keep the peace, avoid blunders, and not lose control. Mill is pragmatic, not dramatic.

So how are we to decide with so much beyond our control? Every consequentialist must wrestle with this question of predicting the future. Mill's answer was that we still must do our best to anticipate and prepare for changes and problems. We are not absolved from making decisions and acting. There are some things that we *can* predict and some things over which we *do* have control. For instance, consider Mill's prescription for the shift to majority rule. Mill predicts the existing class system will eventually give way to sharing power. If we control this process, we can minimize some likely damage. He identified two general paths this might take:

> The two strongest tendencies of the world in these times are towards Democracy and Revolution; meaning by Democracy—social equality, under whatever form of government; and by Revolution—a general demolition of old institutions and opinions, without reference to its being effected peaceably or violently. (CW XX, 297)

Mill believed that increasing democracy *peacefully* (e.g., through universal suffrage and expanding freedoms) would bring civilized people closer to social equality at far less cost than violent revolution. He also had extensive ideas on forms of liberal democracy most conducive to social equality. Egalitarianism was vital to Mill's agenda. He committed to equal opportunity and equality under the law. He was a pioneer for gender equality and universal suffrage. To guide us down this path to democracy, gradual change is more likely to succeed. Taking time to build consensus is important. Mill reminds us throughout his writings that

[5] See also, "Centralisation" 1862, CW XIX, 590.

the minds and opinions of people need to be prepared for change to take hold and last.[6] The wiser course is to *avoid* the destruction, pain, and uncertainties of violence.

Another problem Mill anticipated stemmed from his understanding of human nature and history: he believed the natural state for human societies is to deteriorate. His answer was to control this negative dynamic by linking "Order and Progress." Progress is served by carefully conserving our fragile gains.

> Order … is a part and means of Progress itself. If a gain in one respect is purchased by a more than equivalent loss in the same or in any other, there is not Progress. Conduciveness to Progress, thus understood, includes the whole excellence of a government. (CW XIX, 377-88)

Forward progress is an illusion if one is also regressing. Because the stakes and the possibilities of backsliding are high, Mill's utilitarian calculus is based on damage control, and it is decidedly risk averse.

> the term Progress is the idea of moving onward, whereas the meaning of it here is quite as much the prevention of falling back … The natural tendency of men and their works was to degenerate, which tendency, however, by good institutions virtuously administered, it might be possible for an indefinite length of time to counteract … we ought not to forget, that there is an incessant and ever-flowing current of human affairs towards the worse, consisting of all the follies, all the vices, all the negligences, indolences, and supinenesses of mankind; which is only controlled and kept from sweeping all before it, by the exertions which some persons constantly, and others by fits, put forth in the direction of good and worthy objects. (CW XIX, 388)

Those persons who contributed their talents to human progress (including those 'virtuous administrators' who prevent setbacks) were Mill's heroes (and I recognize Mill as an exemplar of such heroism). He emphasized the crucial importance of the dedicated, energetic, industrious, informed, expert, innovative, nonconforming, eccentric, genius personality types who were the stimuli of progress. Mill believed in an elite class—a 'clerisy'—as those most likely to propel civilization forward. He recognized the centrality of individual actors directing and managing history—i.e., the 'great man' theory of history. He also understood that environmental conditions, economic forces, social dynamic tension, and the state of intellectual development are driving forces of history. The key to progress was reasoning and intellect: free and open discussion prepares people to benefit from positive change. Civil liberties promote diversity, variety, creativity, and improvement. The marketplace of ideas is the best forum for correcting errors and working out our differences. 'Systematic

[6] See, e.g., "Armand Carrel" 1837, CW XX, 200; "Reorganization of the Reform Party" 1839, CW VI, 482; "Coleridge" 1840, CW X,137f; *A System of Logic* 1843, CW VIII, 926.

antagonism' sharpens the public's understanding. It enables liberal societies to avoid the traps of authoritarianism, uniformity, and ossified customs. This is the dynamic progressive environment best suited for preparing responsible, rational, mature citizens for democratic governance.

Mill's optimism came partially from realizing that liberal Victorian England was at the cutting edge of innovation, progress and civilization. This is a social achievement that took generations to devise. Mill's pessimism stemmed from his knowledge that civilizations and centers of progress rise and fall. They are actually fragile. The historical trend was for an exceptional region or society to be open, energetic, dynamic, prosperous, and improving. This upward trajectory would begin to stagnate when complacency sets in, and then decline when beliefs become enshrined in custom or religion. Mill feared that Britain would fall from preeminence, if ever it lost its commitment to individual liberty and tolerance, limited government and the rule of law. To prevent this, he defended and reinforced an appreciation for the liberal value system, putting safeguards in place to keep the delicate experiment succeeding for future generations.

On the strategic level of ideas, Bentham, Mill, and the Philosophic Radicals were revolutionaries; however, in tactical terms they were reformers. It was generally more efficacious to work within the system to transform and improve it, rather than destroy and replace it. Thus, it was urgent to attend to the details of transitioning and finding optimal replacements for inadequate policies and practices. Bentham and Mill were far more circumspect in their tactics than Marx and Lenin. They thought through the consequences of their critiques and proposals. Mill denounced French political thinkers, such as Rousseau, Robespierre, and Proudhon for recklessly undermining social institutions without offering viable replacements. Mill criticized the continental tradition that relied on 'natural' rights claims for secular morals.

> This mode of thought reached its culmination in Rousseau, in whose hands it became as powerful an instrument for destroying the past, as it was impotent for directing the future. The complete victory which this philosophy gained, in speculation, over the old doctrines, was temporarily followed by an equally complete practical triumph, the French Revolution: when, having had, for the first time, a full opportunity of developing its tendencies, and showing what it could not do, it failed so conspicuously. (CW X, 299-300)[7]

[7] On Proudhon see: Mill's letter to Harriet Taylor, 31 March, 1849, CW XIV, 21.

Marx and Lenin learned from the French philosophers and Jacobins, and advocated violence as necessary and justified to overthrow capitalism.[8] Mass murder, forced labor, famine, and the gulag were the result. The commissars and centralized planning were far less efficient than the free market, as the communist parties in the Soviet Union, Eastern Europe, China, and Vietnam eventually figured out. For today's crop of 'new and improved' neo-Marxists, democratic socialists, and postmodernists who carry on the tradition of 'critical theory', the same problem remains—strident critiques without viable, practical alternatives (Nozick 2001, 55). Even if they eschew violent revolution as a means to achieving their visions, they still were dangerously destructive. Mill feared that when consequences are ignored, the cure would be worse than the disease.

In sharp contrast to destructive criticism, Mill praises Bentham for tending to the important task of offering viable alternatives to outdated and unfair programs.

> The age then is one of *destruction*! Disguise it as we will, it must be so characterized; miserable would be our lot were it not also an age of preparation for reconstructing. What has been the influence of Bentham upon his age?—it has been twofold—he has helped to destroy and also to rebuild. No one has done so much to forward … the work of destruction, as Mr. Bentham … if he ever annihilated a received opinion, he was sure of having something either good or bad to offer as a substitute for it; and in this he was most favourably distinguished from those French philosophers who preceded and even surpassed him, as destroyers of established institutions on the continent of Europe. (CW X, 501)

Bentham and Mill approached their activism with such consequentialist caution that they are best characterized as 'conservative revolutionaries.' They were committed to promoting utility by managing change. The French debacle reminded Mill to think seriously about progressing with the minimal amount of suffering. As he explains, in the aftermath of the bloody French experience, thoughtful people pondered the question of governance and how best to proceed.

> Other nations, and England more than any, are in the middle of *their* Revolution. The most energetic minds are still occupied in thinking, less of benefits to be attained, than of nuisances

[8] On some apologetic interpretations, the elder Marx understood the potential of universal suffrage, encouraging his activist followers in advanced countries to take advantage of civil liberties to subvert the capitalist system using peaceful means. See, Marx's Amsterdam speech to the International Working Men's Association (September 8, 1872), https://www.marxists.org/archive/marx/works/1872/09/08.htm. See also, Marx's pamphlet, *The Civil War in France*, "The Second Address" (September 9, 1870), https://www.marxists.org/archive/marx/works/1871/civil-war-france/ch02.htm. This is the next to last sentence in the Preface, https://www.marxists.org/archive/marx/works/1867-c1/p6.htm. See also, Richard W. Miller, *Analyzing Marx: Morality, Power and History* (Princeton University Press, 1984), 114-26.

> to be abated, and every question of things to be done, is entangled with questions of things which have first to be undone; or of things which must *not* be undone, lest worse should follow. (CW X, 298)

What Lenin and Mill thought about attaining power and governing were poles apart. Mill preferred to build consensus through reasoned debate rather than resort to force (Williams 1989). He understood that revolutionary changes are usually messy, disruptive, divisive, chaotic, and sometimes violent and irreversible. These are yet further reasons for responsible leaders *not* to rush in, but to proceed with forbearance.

None of this is to suggest that Lenin did not 'think through' Marx's ideas or make careful plans to achieve and expand power. The Bolsheviks' ready use of violence and their ruthless suppression of opponents, rivals, and even comrades came right out of their Marxist (and Russian) playbook. The enormous cost in terms of human lives and freedoms was simply the price for political victory. Lenin needed to move quickly to quit the war with Germany and fight the counter-revolutionary White Army plus eight foreign armies on multiple fronts. He had little time to make cautious decisions. Lenin was an expert synthesizer and communicator. His genius lay in his ability to adapt a mid-19th century German social philosophy to fit Russian culture and inspire enough people to accept his vision, eliminate opposition, and force the rest of society to comply. Lenin was able to make Marx's complex theories appeal to the hearts and minds of Slavic peasants. This is how in one generation the Bolsheviks transformed the poor, backward, feudal Russian Empire into a modern, industrialized Soviet super power (Northrop 1946, chap. 6). We must give Lenin his due. He succeeded on his own terms.

For all of Mill's influence as *the* public intellectual, he achieved nothing comparable to Lenin. He was not a forceful 'man of action.' He was not aiming to eliminate private property or to liquidate the nobility and bourgeoisie. He inspired no movements that instituted a new political system through force. From a Marxist standpoint, Mill was a timid, reactionary apologist for capitalism, standing in the way of history. He was evolutionary and not revolutionary. However, this reflects a narrow ideological perspective. In fairness to Mill, we must understand him on his own terms. After all, he *was* a revolutionary, in that he advocated a radical transformation of Britain and the world. *The* leading economist in his generation, Mill was attentive to socialist experiments and he remained both open-minded and critical. No pacifist, he formulated a coherent position on when political violence is legitimate (Williams 1989). But he also showed concern for when it is unnecessary, unlikely to succeed, and a formula for massive pain. He argued that dialogue and appeal to reason are more likely to avoid pain than the use of force.

Mill believed that political revolutions originate in moral revolutions (CW XX, 118). They often arise from just and good motives, and are legitimate reactions to injustice. Happy or privileged people do not revolt; rather, history is made by those who struggle against oppression. He sympathized with the oppressed, and publicly supported most rebellions on his watch. In addition to his participation in the July 1830 and the February 1848 revolutions in Paris, Mill took strong principled stands supporting the Canadian rebellions of 1837-8, the Morant Bay Rebellion in Jamaica (1865), and the Maori Wars in New Zealand (1845-72). He stood up for the Irish in their struggles with England. He strongly opposed the rebellious Confederate states in the U.S. Civil War. He opposed the Sepoy Rebellion in India (1857), albeit for less moral reasons. As I have argued, Mill consistently took principled positions based on his utilitarian passion for improvement and progress, which included minimizing pain (Habibi 2001, 2017). In his day, Mill was a vocal defender of many revolutionary causes. Today, his radical ideas are accepted and well within the mainstream of the liberal democracies and international norms. If my arguments have merit, then Mill's successes should serve to highlight, rather than obscure, his stature as a conservative revolutionary thinker. Mill's prescience taught us to build open, liberal societies as the best path for achieving a progressive and happier world.

References

[1] Goldstone, Jack A. 2014. *Revolutions: A Very Short Introduction*. Oxford: Oxford University Press.

[2] Habibi, Don. 2001. *John Stuart Mill and the Ethic of Human Growth*. Dordrecht: Kluwer.

[3] _____. 2017. "Mill on Colonialism." In *The Blackwell Companion to Mill*, edited by Christopher Macleod and Dale E. Miller, 518-32. Oxford: Blackwell.

[4] Milgram, Elijah. 2017. "Mill's Epiphanies." In *The Blackwell Companion to Mill*, edited by Christopher Macleod and Dale E. Miller, 12-29. Oxford: Blackwell.

[5] Mill, John Stuart. 1963-1991. The Collected Works of John Stuart Mill, 33 vols. Toronto: University of Toronto Press / London: Routledge and Kegan Paul.

[6] Northrop, F. S. C. 1946. *The Meeting of East and West*. New York: MacMillan.

[7] Nozick, Robert. 2001. *Invariances: The Structure of the Objective World*. Cambridge: Harvard University Press.

[8] Williams, Geraint. 1989. "J.S. Mill and Political Violence." *Utilitas* 1 (1): 102-11.

The Place of Good, Goodness and Goods within Consequentialist Frameworks

Martin Hähnel, Catholic University of Eichstätt-Ingolstadt, Germany

Abstract

Most of utilitarian theories are using the word “good” in the instrumental sense of “good for” and as an agent-neutral basis for the aggregation of different goods. Furthermore, if utilitarian approaches are apparently consequentialist and universal, they usually must expand their normative basis by adding the teleological component of absolute goodness. Thus, the universal and well-known principle of maximizing the goodness of consequences results from combining the instrumental and regional use of “good” with an almost cosmological outlook. Depending on the axiological structure this consequential goodness can also be defined as a goodness of outcome whereby goods are treated as certain states of affairs serving as commensurable measures of this goodness of outcome. From this it follows, that within consequentialist frameworks everything that has been, is or will be evaluated (i.e. the consequentialist goods) must be understood in terms of a state of affair (as a perfect bearer of instrumental and intrinsic value). In this paper I try to show that goods, in every respect, should not be identified with states of affairs. Against welfarism I claim that good life as an optimal combination of primary and secondary goods does not necessarily depend on thoughts about the aggregation or best distribution of these goods. Such goods, in my view, rather are species-relative qualities of an agent that characterize his moral flourishing. This kind of flourishing cannot be derived from an absolute or “best” goodness.

Introduction

What does it mean that a theory of goodness is or wants to be “consequentialist”? In order to give an appropriate answer to this question, we first have to look at what consequentialism properly means. As is well known the term was introduced by G.E.M. Anscombe (cf. Anscombe 1958) in order to characterize ethical approaches that typically emerge between Mill and Moore. To be "consequentialist” means nothing more than to evaluate actions as “right” or "good" solely with regard to the assessment of their consequences. For Anscombe it is the greatest temptation of consequentialism to establish a theory that allows intending certain consequences and effects of intentional acts.

But what characterizes consequentialism as a genuine *theory of the good* that differs from a standardized theory of the right? In principle, neither intrinsic features of actions (as in Kantian approaches) nor typical character-qualities of the agent (as in virtue ethical conceptions) play a crucial role for a consequentialist theory of the good and its justification. For consequentialist moral reasoning it is more decisive to justify that the best agent-neutral state of affairs – in the sense of an impartial maximization of positive outcomes – can be

expected and should be brought about. Thus, the good is something extrapolated to a person-indifferent state in the future whose entrance into the present has to be guaranteed in compliance with the premise that normative qualities can only depend on their consequences. From that it follows that the emerging form of the good generally has a resultative form or character.

Insofar as these consequentialist moral theories evaluate decisions, actions, and motives according to what is consequentially good or bad, they presuppose that there exist certain facts which are, in the end, intrinsically good. However, this goodness can only be intrinsic, as already mentioned, if it has a resultative (not a final!) form. A so-called "goodness of outcome" can be thematic in the context of a monistic conception as simple pleasure (as in classical utilitarianism), as an objective concept of well-being, as desert (Feldman 2012), or as regards the equality and inequality of individuals and their rights (Scanlon 1977; Sen 1979). On the other hand, we can adopt a pluralist account of resultative goodness that is open to the integration of the aforementioned elements into its own explanatory paradigm.

But if consequentialist approaches are so heterogeneous and can hardly be distinguished from approaches that vehemently insist on being non-consequentialist, what is their peculiarity? According to a common thesis, it might be possible that all consequentialist accounts share the same axiology that is flexible enough to suggest that moral assessments, which are exclusively based on the evaluation und promotion of actions with positive outcomes, are consistent with common-sense judgements (Portmore 2011) and a universal (and theological: Camosy 2012) understanding of the common good.

If we take a closer look at this subject matter, we can state the following: In all consequentialist approaches, it is not necessary to qualify an action for its own sake or to speak about agency in terms of ascribing to a person a certain moral property, e.g. responsibility. The intrinsic goodness of an agent does not really matter because what is good or bad is only a question of capturing and measuring something as a certain state of affairs. But can such states of affairs be "good" or “bad” at all?

Weyma Lübbe has shown that consequentialists are inclined to talk about certain states of affairs because they cannot indicate "where an action ceases and where its consequences begin.” (Lübbe 2016, 326). Thus, the death of Fritz, who was shot by Franz (no matter how to describe the process of killing) can be qualified as a certain state of affairs. However, this evaluation contradicts our basic intuitions as we continue to believe that the death of Fritz a) belongs to one act, namely the voluntary activation of the pistol trigger by Franz, and that b) the death of Fritz is the obvious consequence of the lethal entry of the bullet into his organism.

Of course, we do not usually assume that the death of a person is a state of affairs one normally has to strive for. However, there are several problems with regard to the determination of a best state of affairs: How can a certain state of affairs be good as such, especially when it has to be brought about first? A classical criticism particularly refers to the argument that there is no absolute good (cf. the criticisms of Kraut 2012; Thomson 1997), which is completely detached from any material conditions and subjective perspectives. Most consequentialists elude this criticism by asserting that "goodness-for" can provide a moral justification, even though we accept the existence of an absolutely good thing from which we can derive the relative goodness and agent-dependent reasons. Although some of the consequentialists insist on the existence of an absolute good we should not confuse this account with the medieval idea of a *summum bonum* that is mostly identified with God. The fact that there is a similarity between the good of the consequentialists and the *summun bonum* has something to do with the idea that the majority of consequentialists, from Sidgwick to Singer (cf. chapter I), identifies the objective good with the ‚good from the standpoint of the universe'. The good at which we can or should look from a universal standpoint guarantees the impartiality and impersonality of our moral judgements. From a practical and partial point of view, however, this absolute ethical standpoint must remain blind to the difference between foreseeable and intended consequences (remember Anscombe's caveat) because it constantly constrains agents to do the good always and everywhere (regardless of whether or not an agent is able to do that). With regard to the scope of our duties this strategy legitimately evokes the objection of being morally over-demanding. With respect to the nature of our motivation to act, consequentialism tends to be sub-demanding. Hence, the consequentialist, unlike the deontologist or virtue ethicist, can assert that the good consequences of a bad action improve the general state of the world or that the bad action does not contribute to worsening this universal condition. Breaking into a pharmacy for stealing a drug that could save somebody's life does not make the burglary good. Similarly, the legal purchase of a remedy, which the patient takes on medical advice but which leads to unexpected multi-organ failure, cannot be described as bad just because the purchase of the drug will cause the death of the patient.

It therefore belongs to consequentialist approaches to deny their preliminaries as well as the consequences that their own theory – to the extent that the underlying axiology may not be aware for their users. In particular, it was Friedrich Nietzsche who first uncovered the particular intention of the Anglo-Saxon utilitarian and consequentialists, whose ideal of general welfare he emphatically rejected. By exposing their inability to question the premises of morality, consequentialism shows that immoralism is a serious option outside the morality while their defenders still believe that there is no possibility of doubting about the truth of moral judgments.

For a real defense of consequentialist theories of the good, there is no need for a separate appeal to a particular strategy of ‚consequentialisation', especially if this strategy does not affect one's own axiology at all. It suffices, as the next example will show, to do the following: A classical consequentialist theory of the good states that it is senseless or restrictive to use "good" attributively, that is to say, in the sense of "good for someone" or "good in one way or another". Rather, consequentialists hold to the predicative use of "good" because the formation of an average amount of good actions is very important for the process of aggregation. Consequently, in order to solve the problem by limiting the good through its attributive use, consequentialists assume that the predicative use of "good" does not ultimately limit the attributive use. The attributive use rather complements the predicative one, which means that every attributive use, if it implies no restriction, is at the same time predicative. For example, according to consequentialist opinion, it can be said that "good for the consequences" or "good for the world" corresponds to the attributive model of "good in a certain respect" (cf. Sinnott-Armstrong 2003). In the end, consequentialists are convinced that the attributive use of "good" is not a real threat for their theory, respectively their axiology.

I Main Representatives of a Consequentialist Theory of the Good

Most representatives of a consequentialist theory of the good are also representatives of so-called noncognitivism. Noncognitivism is the collective term for a metaethical position according to which moral statements are not objective truths in the sense of moral qualities or facts to which concepts such as "the good" could refer at all, but merely considered and evaluated as personal statements. To put it bluntly, the normative judgment "The murder of Fritz is bad" means nothing more than 'Yuck! Murder!' As a serious result of this 'boo-hurray-theory of ethics' goodness is the expression of a subjective consent or disapproval of a morally neutral state of affairs (such as murder). Most consequentialists of our times hardly have distanced themselves from this radically subjective foundation of morality, which is of course not free from indecency. In the sequel, the four most important protagonists of this moral theory are briefly introduced: Henry Sidgwick, Richard M. Hare, Peter Singer and Derek Parfit.

Without a doubt, Henry Sidgwick is one of the great fathers of the other authors just mentioned, and thus one of the decisive champions of a consequentialist turn in 20th-century

moral philosophy. He was the first who tried to do ethics from an absolute valuation standpoint, opposing both the Kantian and the ancient paradigms of ethics. At first, it was important for him to nullify the claims of the classical divine command theory which Elisabeth Anscombe deems to be unfounded in order to establish modern morality. Sidgwick thus asserted that a theist must not have a objective standard of goodness that is independent of God – something that Singer later revisits while making it explicit in his own way. Basically, Sidgwick sees himself as an intuitionist. For him, ethics are based solely on the self-evident premise of rational benevolence. He claims, among other things, that given the "point of view of the universe", one's own good cannot be worth more than the good of the other. In doing so, he tries to counteract a (quite possible) egoistic disinterest in this objective perspective by focusing on a universal benevolence that is supposed to enable agents to combine their own interests with the interests of others: "[T]he good of anyone individual is of no more importance from the point of view … of the Universe, than the good of any other" (Sidgwick 1981, 382). However, Sidgwick fails to develop a universal approach from a subjective non-cognitive moral theory of self-interest, which actually brings self-interests together with foreign interests.

For this reason, following Sidgwick, we should mention the moral theory of Richard Mervyn Hare, who assumes that self-interests and the moral judgments are logically intertwined because our self-interests are universalizable in the guise of internalisable rules. The so-called prescriptivism states that moral judgments do not reflect how the world is (in a certain sense), but preserve their meaning through the properties of prescriptiveness and, as already mentioned, through its universalizability. According to the quality of prescriptiveness, moral judgments always imply imperatives. However, since you can only really accept imperative sentences according to Hare, even if you act accordingly, you can only accept a certain moral judgment if you act accordingly under the same circumstances. For example, if I sincerely make the moral judgment, I should do ϕ right now, then I am also determined to do ϕ. In this respect, moral judgments reflect volitive attitudes. According to the property of universalizability, moral judgments (similar to the deontological position of Kant) are generalizable: for example, if I pass judgment, Person P should now do ϕ, then I agree that any person who is in similar circumstances as P, should do ϕ.

Finally, it is Peter Singer who wants to merge Sidgwick's insights into the impartial ethical standpoint and Hare's reflections on prescriptivism into a universal theory of consequentialist altruism. Goodness-for and absolute goodness are thereby 'extended' to the good from the impersonal point of view of the universe (say the proponents of this approach) or 'reduced' (the critics would say). Singer currently introduces his reflections on an expansive and absolute moral standpoint to a greater audience with the help of proclaiming an "effective altruism", which decidedly integrates economic considerations into its program. This

suggests a new synthesis of individual consequentialist and decidedly (world-) community-oriented designs. While for former representatives of the idea of a free global market, such as Friedrich von Hayek, altruism is something that is fundamentally irrelevant for the market, because it only plays an effective role at close range (among family members and relatives, etc.), effective altruists assume a 'responsibility for the world', which obliges all people, no matter if they are near or far from the needy, to give as much of their income as they can. This attitude is not anti-profit, because the more you earn, the more you can give to others. Although effective altruism has certainly lost any belief in the self-regulation or the natural evolution of the market, the fact that effective altruism is not anti-profit does not mean that it is already prosperity-oriented. Its central focus is no longer on the unlimited proliferation of goods, but on the universal elimination of evils. Goods may be increased only if they remain bound to the purpose of the elimination of evils. Other purposes, e.g. aesthetic, may not contribute to the distribution of goods. Singer cleverly avoids the problem of equitable distribution of aggregated goods by preventing a leveling down insofar as people do not have to forego their individual needs, i.e. they do not have to make personal sacrifices when donating something in order to bring about a good state of affairs, e.g. a fair distribution of food.

As the last representative of a highly elaborate and broadly understood consequentialism, which has partially obscured its traces to the theoretical origin, is the consequentalism of Derek Parfit. Parfit does not conceive of the best state of the world as the objective quantity to be achieved by using only the right means effectively but he defines the "best" as something that is contained in the meaning of impartial reasons. Thus, Parfit believes that he is able to ward off a utilitarian interpretation of the goodness of outcome or resultative goodness. Furthermore, he criticizes the view that goodness is only in the future, i.e. as a consequence of actions that are intended to bring about an good state of affairs (Parfit 2006, 233). Because Parfit asks in his work for the aggregation of the well-being of many people, he inevitably participates in the "debate about the best version of consequentialism" (Parfit 2017, 31). Although it is repeatedly claimed that the importance of Parfit's approach cannot be limited to consequentialist theories, the suspicion is vivid that Parfit only refers to non-consequentialist models (e.g. Kant) in order to refine his comprehension of it. His refined consequentialism, for example, is expressed in his so-called Triple Theory, in which he transforms Kantian, contractualist and consequentialist theory into a superordinate model. Similar to Hare, Parfit attempts to reconcile Kantianism, which forms the basis of contractualism, with consequentialism. Although Parfit, especially in his early work *Reasons and Persons*, contemplates act-consequentialism, he tends to accept in his later work *On what matters* rule-consequentialist principles. In so doing, he interprets Kant's Categorical Imperative insofar as generalizability is no longer the result of a rationality-driven test for the

consistency of subjective maxims, but the result of a balance between impartial and partial reasons, with the greater generalization potential being due to impartial reasons, which ensure a better course of things than partial reasons. However, it is questionable how Parfit deals with categorical prohibitions such as the killing of innocent people against the background of his theory. Another criticism of Parfit is that it is generally impossible for agents to find out which principles can bring about the best states of affairs ensuring that we can act morally right (cf. Scruton 2011). Parfit counters this criticism by withdrawing, as many British moral philosophers did before him, into a commonsense position, claiming that we have always been in unanimous decision-making (such as politics and society) and we are still in agreement – so what?.

II Consequentialism and the Good Life

We have already seen that in most cases consequentialist theories of the good also form the basis for a particular theory of the good life. So-called welfare theories are not to be confused with eudaemonistic theories, because – in short – it is not an issue for consequentialists to 'make happy people, but to make people happy'. Consequentialist goodness therefore is independent of how a human being is or should be; this special type of goodness is decisively a goodness of objectified or objectifiable states of affairs, which are solely determined by the subjective well-being of an individual or entire groups. This form of well-being can consist in the preservation and promotion of pleasure or the absence and avoidance of suffering, in the fulfillment of preferences or in the updating of objective values or normative ideals. Of course, the list is not complete yet. However, it can generally be said that most of current welfare theories are characterized by an entanglement of actuality (such as pleasure feeling) with ideal states (freedom from suffering, wishfulness) as well as between selfishness and altruism.

In this context, the question of the sizing of the resultative good seems quite interesting. In recent years, new theoretical models that are based on consequentialist moral reasoning have emerged here: aggregationism and prioritarianism. In this regard, the welfare-economic model of aggregationism claims that one can sum up the good without having to go into utility comparisons leading to particular distribution decisions. In contrast, prioritarianism considers that benefit comparisons play an important role in determining the welfare of individuals and particular groups. Prioritarianism thus describes a constructivist theory of consequentialism in which two states are technically created (the state of the poorer and the state of the better off), which are judged and weighted by an impartial observer in order

to arrive at a just distribution of an absolute good (that could be wealth in this case). Prioritarianism seeks to correct, in particular, the ethical blindness of aggregationism in distributional contexts.

This short description is primarily intended to clarify that both aggregative and prioritized approaches have central properties of consequentialist theories of good: the aggregate and, if possible, fair distribution of benefits remains the result of an assessment of the outcomes. In both cases, the consequences have been quantified first; while in the first case they are merely aggregated, in the second case they are both aggregated and distributed. It is also important to know that successful aggregations and effectively distributed benefits represent an impersonal value for the theories themselves, because the positive benefits (this time being understood as personal value) for the worse or equals and the negative for the better off are the same.

III Conclusion

But which use of "good" seems to come closest to our intuitions? It is not surprising that we frequently use the word “good” in each normative framework. However, there are fundamental differences with regard to the used operator and its role or value for constituting moral judgements:

Operator	**Value/Role**	**Corresponding Ethical Paradigm**
“good for X”	instrumental, aggregative	virtue ethics, consequentialism
“good of X”	constitutive, functional	virtue ethics, deontology
“good at X”	performative	virtue ethics, consequentialism
“good as X”	exemplary	(Neo-Aristotelian) virtue ethics
“the good X”	attributive	virtue ethics, deontology

Within consequentialist frameworks everything that has been, is or will be evaluated as good must be understood in terms of a state of affairs, the perfect bearer of instrumental and intrinsic value. Hence, consequential goods are good states of affairs, which serve as commensurable units or measures of a so-called ‘goodness of outcome’. From that is fol-

lows that modern consequentialist evaluations of good states of affairs lead to an aggregation of goods; and an aggregation of goods leads to the prioritization (ǂ fair distribution) of some (= better) goods.

Finally, I would like to provide an outlook for another definition of what is good without bringing about any good states of affairs and without confusing our everyday intuitions: a Neo-Aristotelian understanding of goodness. What is it about? According to current Neo-Aristotelianism, the species-relativity of "good" (Foot 2001; Thompson 2008) secures the agent-relativity of moral judgements in the following non-consequentialist way:

1. "Good *for* me" can only be good for me if I belong to the species *homo sapies* (inhuman things are not good for me) that imposes certain norms on me.
2. "Good *for* all" is only good if all members of the same species fulfill 1, everyone for herself/ himself.
3. "Good *for* its own sake" is only good if 1 and 2 are not identical.

Now consequentialists could say that doing an action which is ruled out by a constraint is bad-relative-to-the-agent. But by virtue of the species-dependence of goodness and badness constraints cannot be accommodated by consequentialism (there is no fitting attitude, because the fitting relation is defined by the species and its necessities). One should finally say: The species itself defines what fits the good or not!

In this paper I have tried to show that ‚good', ‚goodness' or ‚goods' should not be identified with calculable states of affairs. Against consequentialist welfarism I claim that the good life as an optimal combination of primary and secondary goods does not depend on thoughts about the aggregation or best distribution of goods. Goods, in my view, are species-relative qualities of an agent that characterize his moral flourishing while undermining every fitting attitude analysis. This kind of flourishing cannot be derived from an absolute or "best" goodness.

References

[1] Anscombe, G. E. M. 1958. "Modern Moral Philosophy." *Philosophy* 33: 1-19.

[2] Camosy, C. C. 2012. *Peter Singer and Christian Ethics. Beyond Polarization.* Cambridge: Cambridge University Press.

[3] Feldman, F. 2012. *Utilitarianism, Hedonism, and Desert*. Cambridge: Cambridge University Press.

[4] Foot, 2001. *Natural Goodness*. Oxford: Clarendon Press.

[5] Kraut, R. 2012. *Against Absolute Goodness*. Oxford: Oxford University Press.

[6] Lübbe, W. 2016. "Handlungen, Handlungskonsequenzen und das Vollständigkeitsaxiom – Ein handlungstheoretischer Kommentar zum entscheidungstheoretischen Konsequentialismus." *Zeitschrift für Philosophische Forschung* 70 (3): 325-41.

[7] Parfit, D. 2006. *Climbing the mountain* (unpublished).

[8] _____. 2017. *Personen, Normativität, Moral. Ausgewählte Aufsätze*. Frankfurt a.M.: Suhrkamp.

[9] Portmore, D. W. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press.

[10] Scanlon, T. M. 1977. "Rights, Goals, and Fairness." *Erkenntnis* 11 (1): 81-95.

[11] Sen, A. 1979. "Utilitarianism and Welfare." *Journal of Philosophy* 76 (9): 463-89.

[12] _____. 1982. "Rights and Agency." Philosophy & Public Affairs 11 (1): 3-39.

[13] Scruton, R. 2011. "Parfit the Perfectionist." *Journal of the Royal Institute of Philosophy* 89: 621-34.

[14] Sidgwick, H. 1981. *The Methods of Ethics*. Indianapolis: Hackett Publishing.

[15] Sinnott-Armstrong, W. 2003. "For Goodness' Sake." *Southern Journal of Philosophy* 41: 83-91.

[16] Thompson, M. 2008. *Life and Action. Elementary Structures of Practice and Practical Thought.* Cambridge, MA: Harvard University Press.

[17] Thomson, J. J. 1997. "The Right and the Good." *The Journal of Philosophy* 94 (6): 273-98.

On Parfit's Wide Dual Person-Affecting Principle

Jonas Harney, Saarland University Saarbrücken, Germany

Abstract

Parfit (2017) proposed a novel principle for outcome betterness in different people and different number choices. It is claimed to solve the Non-Identity Problem while avoiding the Repugnant Conclusion, and it shall do so in person-affecting rather than in impersonal terms. According to this Wide Dual Person-Affecting Principle, one of two outcomes would be (i) in one way better if this outcome would together benefit people more, and (ii) in another way better if this outcome would benefit each person more. I argue that a plausible construal of this principle has two features that make it vulnerable to objections. First, the most plausible interpretation of the second part of the principle turns out to incorporate an average function. Although this helps to avoid the Repugnant Conclusion, it implausibly implies that it can be better to add further people with less bad but still miserable lives to populations consisting only of lives full of suffering. Second, the principle is not based on a comparative but on an intrinsic notion of benefit. This allows to solve the Non-Identity Problem, yet it accounts only for a weak sense of person-affecting rather than for the more substantive person-affecting intuition that it is morally significant that particular people are made better (or worse) off. Eventually, I highlight what we can, nevertheless, learn from Parfit's idea of combining different ways in which outcomes might be better.

Introduction

Derek Parfit concluded the fourth part of his *Reasons and Persons,* one of the most influential works in population ethics, with a challenge to the philosophical community: We shall find *Theory X* – the best theory of beneficence that solves the Non-Identity Problem and avoids the Repugnant Conclusion, among other problems of population ethics. Parfit predicted that Theory X will take an impersonal form insofar as it "will not appeal to what is good or bad for those people whom our acts affect" (Parfit 1984, 378). While some population ethicists follow Parfit's direction, many others reject impersonal views as implausible. They claim that the part of morality that is concerned with people's wellbeing needs to be spelled out in person-affecting rather than in impersonal terms. In his posthumously published paper "Future People, the Non-Identity Problem, and Person-Affecting Principle", Parfit gives in to this person-affecting intuition. He, now, claims it "to be a bad mistake" (Parfit 2017, 123) to have been advocating an impersonal view and abandons it in favour of the

> *Wide Dual Person-Affecting Principle*: One of two outcomes would be in one way better if this outcome would together benefit people more, and in another way better if this outcome would benefit each person more. (Parfit 2017, 154)

In this paper, I examine the Wide Dual Person-Affecting Principle, or Dual Principle for short. Unfortunately, Parfit is not very clear in said paper since he could not finish it before he unexpectedly died. Given that Parfit's formulation of the principle is therefore rather vague, I start by clarifying the Dual Principle and highlighting the principle's aims and features in section I. I then argue that the principle is deficient in two ways. First, in its most plausible construal, the principle turns out to be a hybrid of total and average functions, which makes it vulnerable to objections against the latter (section II). Second, the Dual Principle accounts only for a weak person-affecting view since it fails to capture the more substantial intuition that making particular people better (or worse) off is morally significant (section III). I close by discussing how the Dual Principle might be improvable and what we still can learn from it.

I The Wide Dual Person-Affecting Principle

The Dual Principle is supposed to provide a definition for the overall betterness of one of two outcomes. It holds that one of two outcomes is overall better in virtue of (1) together benefiting people more, and in virtue of (2) benefiting each person more. Unfortunately, the Dual Principle lacks analytical clarity for a proper understanding regarding both the *two ways* of betterness and the *overall* betterness they are supposed to constitute.

Note that the two ways of betterness are supposed to *conjointly* make one of two outcomes overall better and, thus, need to be weighed against each other in case of conflict. Since Parfit did not provide specification for the weight of the two ways, the Dual Principle as such does not suffice for overall betterness yet. For simplicity, I assume that the two ways have equal weight, and I will point to alternative ratios only if it makes a difference for the overall betterness. Furthermore, it still remains unclear how to understand the two parts in themselves. I will provide a plausible interpretation in the next section. But let me mention the three aims the Dual Principle is supposed to achieve beforehand.

First, the Dual Principle is claimed to be a person-affecting rather than an impersonal principle. While person-affecting principles are based on what is good (or bad) for people, impersonal principles do not refer to the relative property of being "good for someone", but

to the absolute property of being "good simpliciter". The Dual Principle bases the comparison of outcomes on the extent to which people benefit. And given that benefits always refer to what is good (or bad) for people, the Dual Principle is person-affecting rather than impersonal.

Second, although person-affecting, the Dual Principle solves the Non-Identity Problem because it does not use a *comparative* but an *intrinsic* notion of benefits. In the intrinsic sense, people are benefited not if (and if then because) they are better off than they would have been otherwise, but if (and if then because) they are caused into a state that is intrinsically good for them. Therefore, people are intrinsically benefited by being caused to exist leading a life worth living even if they are not better off than otherwise since they would never have exist at all. It follows that an outcome with a very happy person is better than an outcome with a numerically distinct and less happy person even though it is not better for any of the two persons. It is better, according to the Dual Principle, because it intrinsically benefits people together more.

Third, the Dual Principle is claimed to avoid the Repugnant Conclusion. This is due to the principle's claim that one of two outcomes is overall better in virtue of two things: the outcome maximizes benefits to *people together*, and it maximizes benefits to *each person*. Take a typical comparison between the two outcomes A and Z.

A: One million people exist with very high wellbeing at level 100.

Z: One billion people exist with very low but still positive wellbeing at level 1.

The principle's first part implies that outcome Z is better than outcome A. This is so because the Z-people together are benefited by one billion units of wellbeing, while the A-people together are benefited only by one hundred million units of wellbeing. Thus, Z benefits people together more than A. This very implication is the Repugnant Conclusion. The Dual Principle's second part, by contrast, avoids the implication. For the people with very low wellbeing would each be benefited considerably less by being caused to exist than the people with very high wellbeing would have been benefited. Given the assumption that the two parts have equal weight, the Dual Principle implies A to be better than Z because, although Z is ten times better than A according to the first part, A is a hundred times better according to the second part. Thus, the Dual Principle seems to avoid the Repugnant Conclusion.[1]

[1] It does so only for some cases as long as we assume equal weight for the two parts of the principle, because for any population Z' that is larger than ten billion people it would be true that Z' is better than A. However, other (potentially variable) ratios might avoid the Repugnant Conclusion in all cases.

We have seen that the Dual Principle is a person-affecting principle that solves the Non-Identity Problem in virtue of the intrinsic notion of benefits and avoids the Repugnant Conclusion in virtue of its dual character. In the next section, however, I argue that the most plausible construal of the Dual Principle is vulnerable to an objection against average principles.

II The Principle's Duality

Parfit initially introduces the two parts of the Dual Principle separately.

> *Collective Principle*: One of two outcomes would be in one way better if it benefits people *together* more, by giving people a greater total sum of benefits. (Parfit 2017, 153)
>
> *Individual Principle:* One of two outcomes would be in one way better if it benefits *each person* more. (Ibid.)

Consider the Collective Principle first. It claims that an outcome is better if it benefits people together more. This just means that the outcome would be better if the *sum of intrinsic benefits were greater*, as Parfit's addition "by giving people a greater total sum of benefits" indicates. Thus, the Collective Principle incorporates a total function in the same way as Parfit's

> *Impersonal Total Principle*: If other things are equal, the best outcome is the one in which there would be the greatest quantity of whatever makes life worth living. (Parfit 1984, 287)

The difference between these principles just is that the former one is formulated in person-affecting terms, while the latter refers to impersonal goodness. But both use a total function in order to determine outcome betterness.

Now, consider the Individual Principle. Since it, unfortunately, lacks analytical clarity, we need a plausible construal as a starting point. First, the antecedent in the Individual Principle can be interpreted differently. I suppose that what "benefits each person more" could mean that either

> (1) each person is benefited more than *she* would have been benefited otherwise, or
>
> (2) each person is benefited more than *another person* would have been benefited otherwise.

One might be inclined to think that (1) is the option Parfit had in mind. It seems intuitively more plausible than (2) since the principle is called "individual". However, this must be wrong. If (1) were the correct construal, the antecedent would be true only if people were also benefited in the comparative sense because a person would be better off by being intrinsically benefited more than *she herself* would be intrinsically benefited otherwise. But if this were the case, the Dual Principle could not avoid the Repugnant Conclusion. For it would then imply that outcome Z (one billion people having a wellbeing level of only 1) is better than outcome A (one million numerically distinct people having a wellbeing level of 100) if no one exists in both A and Z. This is so because, since A would not be better for anyone, the antecedent of the Individual Principle would not be true. Thus, the Individual Principle could not mitigate the Collective Principle's assessment. Given that the Dual Principle was meant to avoid the Repugnant Conclusion, (1) cannot be the correct interpretation for Individual Principle.

Second, (2) can be interpreted differently again since "another person" can refer to either

(a) a *particular* other person, or
(b) *each* other person, or
(c) the *average* per person.

These different interpretations do not make a difference if we hold on to Parfit's restriction of the Dual Principle that there is perfect equality within each outcome (cf. Parfit 2017, 152). However, this restriction does not hold for most cases we are concerned with. And since this would severely limit the scope of the Dual Principle, we are well-advised to decide between (a), (b), and (c). Though one would not expect so on first sight, it will turn out that (c) is the most plausible construal as soon as we drop the equality restriction.

Consider (a). If it were correct, the Individual Principle would hold that each person must be benefited more than another *particular* person. This construal immediately raises the question who that particular person is supposed to be. Since there might exist completely different people in two outcomes, this remains quite unclear. Furthermore, if we compared each particular person's benefit only with *one particular* other person's benefit, the Individual Principle would miss different number cases. For there would be some people in the higher populated outcome without a counterpart then. But if we compared some people with *multiple* other people, any actual comparison would be rendered arbitrary. Thus, (a) is unconvincing.

Consider (b). The Individual Principle would then state that each person of one outcome must be benefited more than *each* other person in the compared outcome. Obviously, this

fairly strong condition would hold only for very few cases. Thus, while yielding the correct result concerning the Repugnant Conclusion, namely that A is better than Z, this construal would fail a tiny variation of that case:

> A: One million people exist with very high wellbeing at level 100.
>
> Z*: One billion people exist with very low but still positive wellbeing at level 1, *and one person at wellbeing level 101*.

The people in A are not benefited more than *each* other person in Z* since one person in Z* has more wellbeing than each of the people in A. Therefore, A would not be better than Z* if (b) were correct. Rather, Z* would be better since only the Collective Principle would apply to this case. But the additional person in Z* should not turn around the overall assessment. Therefore, (b) must be wrong. It would make the Individual Principle irrelevant in many cases.

If (a), (b), and (c) indeed exhaust the space of plausible interpretations of "another person", then we are left with (c). For, as I have shown, if (a) or (b) were correct, the Individual Principle would be either arbitrary or irrelevant in most cases. By contrast, if (c) is correct, there is no need for mapping counterpart relations between particular people, and we still get the correct result concerning A and Z*. For the proportion of average intrinsic benefits between A and Z* is nearly the same as between A and Z, which reflects that, although there is a difference between Z and Z*, it is not a huge one.

Third, note that (c) might mean that each person needs to be benefited more in the better outcome than the average benefit per person in the other outcome. However, this would give rise to similar objections as it is the case for (b). Therefore, (c) plausibly means just that the average benefits in one outcome are higher than the average benefits in the other one.

If my arguments are sound, the Individual Principle is best understood as the condition that the better outcome benefits people more *on average*, and the Collective Principle as the condition that the better outcome benefits people more *in sum*. Hence, the Dual Principle boils down to a hybrid of well-known total and average principles (which take intrinsic benefits as value units here).

The problem is that the average character of the Individual Principle makes the Dual Principle vulnerable to a common objection against average principles. Suppose that there are the two outcomes

> B: Ten people exist each having a miserable live at wellbeing level –100.

C: Ten people exist at wellbeing level –100, and one person exists at wellbeing level –10.

The average benefits for the people in C are higher than in B, because C intrinsically burdens people on average less than B. Thus, according to the Individual Principle, C is better than B although there is one additional person with a less bad but *still miserable* life. This is highly implausible. Furthermore, the Collective Principle fails to mitigate this result. For, according to the Collective Principle, C is just around one hundredths times worse than B, because the sum of intrinsic burdens in C is just slightly higher than the sum of intrinsic burdens in B. But, according to the Individual Principle, C is around one tenth times better than B. Thus, assuming that the proportions of betterness and worseness have equal weight, C is even all things considered better than B.

One might be inclined to reply that a different weighing ratio that gives more weight to the Collective Principle can avoid the objection. However, even though this is true for the presented case, there will be some hypothetical situations in which the Dual Principle still yields such an implausible implication even with a weightier Collective Principle. Furthermore, and even more problematic, if the Collective Principle had more weight, the Dual Principle would be more prone to the Repugnant Conclusion again. Thus, as long as we have not found more sophisticated weighing functions solving this problem[2], the Dual Principle is prone to either the Repugnant Conclusion or the stated implausible implication due to the average function the second part incorporates.

We have seen that Dual Principle has implausible implications given the average function of the Individual Principle. Thus, while the duality of the Dual Principle helps to avoid the Repugnant Conclusion, it creates another, potentially even worse problem. In the next section, I argue that the Dual Principle captures the person-affecting intuition only in a weak sense and thus fails to achieve one of its main aims.

III The Principle's Person-Affecting Character

The Dual Principle is claimed to be a *person-affecting* principle in virtue of it being spelled out in terms of *benefits for people*. Furthermore, it is based on an *intrinsic* notion of benefits securing that the principle is applicable to different people choices and able to solve the

[2] Matthew Clark argued for such a solution in his talk "The Continuous Weak Superiority View" at the 15th Conference of the International Society for Utilitarian Studies 2018, Karlsruhe.

Non-Identity Problem. The question, however, is whether the Dual Principle captures the person-affecting intuition if benefits are used in the intrinsic sense. My claim is that, although the Dual Principle is person-affecting in a weak sense, it fails to capture a more substantial person-affecting idea.

Take a principle about outcome betterness to be person-affecting if and only if any assessments made on behalf of the principle supervenes on personal goodness or personal betterness, that is, facts about how particular people absolutely or comparatively fare. This is in line with the person-affecting claim that, as Nils Holtug puts it, "the part of morality that concerns individual welfare should be cashed out in terms of what is good and what is bad (or what is better and what is worse) for individuals" (Holtug 2004, 129). Given this understanding, we can distinguish a weak and a strong sense of person-affecting. According to the weak sense, assessments need to supervene on personal *goodness* only. According to the strong sense, by contrast, assessments need to supervene on personal *betterness* too.[3]

The Dual Principle is spelled out in terms of intrinsic benefits for people, that is, how well off particular people are in a certain outcome. It thus complies with person-affecting-ness in the *weak* sense. The overall outcome betterness is derived from facts about how *well (or badly) off* people are. However, it fails to account for the *stronger* sense because it does not account for people's being *better (or worse) off*. For illustration, consider a case in which there are the two outcomes

D:	Ali has wellbeing 10	Bel has wellbeing 5	Cam doesn't exist
E:	Ali doesn't exist	Bel has wellbeing 10	Cam has wellbeing 5

According to the Dual Principle, D and E are equally good. For both outcomes benefit people by 15 units of wellbeing in sum and by 7.5 units of wellbeing on average. However, there is a clear sense in which D is worse than E: D is worse than E for Bel. Note that this assessment is not attenuated by the fact that Cam has only 5 units of wellbeing in E while Ali has 10 units of wellbeing in D. For Ali and Cam are different people. Hence, Cam could not have been better off in D. Rather, he would not have existed at all. Thus, while the Dual Principle considers what is *good* and what is *bad* for individuals, it fails to properly include what is *better* and what is *worse* for them. The principle is person-affecting only in the weak but not in the strong sense.

I wonder why this should be so. Given that the Dual Principle is already composed of two ways in which outcomes can be better, there does not seem to be, on principle, a reason

[3] This implies what is commonly known as person-affecting restriction, according to which one of two outcomes can be better only if it is better for someone.

against adding (or combining it with) other views. And I think that we would do better by taking the strong sense of person-affecting-ness seriously. My claim in favour of this option is that it morally matters not only how much people are *intrinsically* benefited, but also how much they are *comparatively* benefited. We care not only about how well or badly off people are, but also about how much particular people would gain or lose. If this is so, we should consider a third way in which outcomes are better that Parfit discusses but rejects.

> *Weak Narrow Principle*: One of two outcomes would be in one way better if this outcome would be better for people. (Cf. Parfit 2017, 129)

This principle expresses the more substantial person-affecting idea to take into account what is *better* or *worse* for people. Even though narrow person-affecting principles fail to solve the Non-Identity Problem (which is why Parfit rejects it), the Weak Narrow Principle does not prevent a solution either since it is neither a necessary nor a sufficient condition for outcome betterness. Hence, we can still combine it with another principle that provides a solution to the Non-Identity Problem such as the Collective Principle.

IV Conclusion

In this paper, I argued that the most plausible construal of Parfit's Dual Principle is vulnerable to two objections. First, it implausibly implies that adding less bad but still miserable lives to a population consisting only of lives full of suffering makes things better. This implication could be mitigated but only at the cost of making the principle more prone to the Repugnant Conclusion again. Second, the Dual Principle fails to capture the strong sense of person-affecting according to which it is morally significant that particular people are comparatively benefited, that is, that they are made better off. We should rather consider a third way in which outcomes are better that captures the strong person-affecting sense.

Where does that leave us? Although Parfit's particular principle has serious shortcomings, we may still learn from the basic idea: a plausible principle of outcome betterness that solves the problems of population ethics might need to be composed of different parts capturing different morally significant factors. With the Collective Principle, the Individual Principle and the Weak Narrow Principle we have three possible components on the table. I think it is worthwhile to examine possible combinations in more detail. I have discussed the problems for combining just the former. Investigating on the other possible combinations may help to make further progress in solving the problems of population ethics.

References

[1] Boonin, David. 2008. "How to Solve the Non-Identity Problem." *Public Affairs Quarterly* 22 (2): 129–59.

[2] Heyd, David. 2014. "Parfit on the Non-Identity Problem, Again." *The Law & Ethics of Human Rights* 8 (1): 1–20.

[3] Holtug, Nils. 2004. "Person-Affecting Moralities." In *The Repugnant Conclusion*, edited by Torbjörn Tännsjö and Jesper Ryberg, 129–61. Dordrecht: Kluwer Academic Publishers.

[4] Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

[5] _____. 2017. "Future People, the Non-Identity Problem, and Person-Affecting Principles." *Philosophy & Public Affairs* 45 (2): 118–57.

[6] Roberts, Melinda A. 2007. "The Non-Identity Fallacy: Harm, Probability and Another Look at Parfit's Depletion Example." *Utilitas* 19 (3): 267–311.

Giving Hedonism a Second (and Proper) Chance[1]

Moritz Hildt, University of Tübingen, Germany

Abstract

Classical Utilitarianism attributed hedonism a prominent place within its theoretical framework – as the "theory of life" on which utilitarianism is grounded (Mill) –, and thus gave hedonism its first chance (in modern times). Today, however, many regard hedonism as fundamentally flawed. As a result, hedonism has more or less dropped out of the picture in the current debates on well-being and the good life. My basic argument will be that the critics of hedonism fail to show the necessity of a wholesale rejection, and that hedonism, thus, deserves a second, and proper chance. After spelling out the argumentative structure of the main charge brought forward against hedonism nowadays – the charge of its (alleged) systematic insufficiency –, I will discuss this charge in its two most prominent forms: Robert Nozick's thought-experiment of the "experience machine", and Thomas Carlyle's polemics that utilitarianism is a philosophy worthy only of swine. I will argue that there is good reason to doubt that Nozick's argument even touches the question of the plausibility (or implausibility) of hedonism, and thus, under scrutiny, appears to be – despite its popularity – a rather easy case for hedonism. Mill's answer to Carlyle, his qualitative hedonism and its well-known problems, however, does indeed present more trouble. My claim here will be that hedonism, as a theory of well-being, should stress its conceptual independence from any moral theory, and especially from utilitarianism, since some of the deepest problems Mill's hedonism encounters are linked to specifically utilitarian concerns. I will conclude by summarizing the main implications my discussion has for contemporary hedonism, and, in doing so, sketch a version of hedonism as a theory of well-being that builds upon the very aspect that many today regard as its major deficit: I argue that hedonism's systematic simplicity can be a major asset.

Introduction

Although hedonism is present in the philosophical debates about well-being and the good life ever since Antiquity, it has very few proponents today. Many seem to think that hedonism – if it ever was plausible – has had its fair chance in modern times within Classical Utilitarianism, where both Jeremy Bentham and John Stuart Mill regarded hedonism as the "theory of life" (Mill 1998, II.2) on which they grounded their moral theory. Today, many

[1] The arguments of this paper are part of a larger research project concerned with re-establishing hedonism as a theory of well-being. This independent research project is generously funded by the Fritz Thyssen Stiftung, to whom I am deeply grateful for providing this scholarship. I have also profited from the comments of the audience at the ISUS 2018 conference in Karlsruhe, Germany, most notably from a conversation with David Lanius. Last but not least, I am very grateful to Jonathan Riley, for a thorough conversation and discussion about Mill and hedonism in New Orleans in April 2018.

regard hedonism as fundamentally flawed. Even most Utilitarians seem to have left it behind for good, shifting their attention away from pleasure and pain, and more towards concepts of preference or desire-satisfaction.

I want to call into question the quick dismissal of hedonism, which has become common practice in much of today's discussion about plausible accounts of well-being. My main claim in this paper is that the critics not only fail to show the necessity of a wholesale rejection of hedonism, but that the very feature they most frequently attack might in fact turn out to be one of hedonism's major assets: its systematic simplicity.

In what follows, I will first spell out the main charge brought forward against hedonism today, which I will call the charge of its alleged systematic insufficiency (*Section I*). I will then discuss this charge in its two most prominent modern forms: Robert Nozick's famous thought-experiment of the "experience machine" (*Section II*), and Thomas Carlyle's polemics against Classical Utilitarianism, which Mill famously discusses in Chapter 2 of *Utilitarianism,* and against which he develops his qualitative hedonism (*Section III*). In both cases, I will try to show how the arguments put forward eventually fail to establish conclusive reasons for the rejection of hedonism. In the last section, I will summarize the implications my findings have for contemporary hedonism, and, on that basis, I will, although quite tentatively, sketch a version of hedonism which takes its systematic simplicity not as a problem, but, rather, as one of its major assets (*Section IV*).

I The Charge of Systematic Insufficiency

It is no overstatement that hedonism can be called the most attacked school of ethics. Ever since Antiquity, when the ethics of hedonism was laid out most prominently by Epicurus and his followers, the theory has experienced a vast array of attacks and blows, ranging from rhetoric polemics and public ridicule to more serious systematic objections. Throughout the most part of the history of ethical thought, hedonism was attacked either for its mere instrumental, and thus, it was argued, deficient account of reason, or for the morally dubious implications of a theory that so openly seems to embrace radical egotism.

Today, however, a third line of criticism seems to be dominant in the debate – so dominant, in fact, that the other two get rarely even mentioned. This contemporary line of attack views hedonism as, at its base, too simple: It cannot account for all the various things that con-

tribute to a person's well-being, which go well beyond the experience of pleasure and absence of pain, or so it is argued. This is the charge of hedonism's alleged systematic insufficiency.[2]

In order to argue for that point, it has become common practice nowadays to merely refer, often only in an half-sentence, to the argument which has become the standard objection against hedonism, and which itself is a version of the charge of systematic insufficiency.[3] This standard objection is Robert Nozick's so-called "experience machine."

II The "Experience Machine"-Objection: Aversion to Hedonism, or to Radical Alienation?

As already noted, Robert Nozick's thought-experiment of the "experience machine" is frequently evoked in the current debate in order to dismiss hedonism as a plausible answer to the question of what makes a human life good.[4] Under scrutiny, however, it is highly doubtable that Nozick's argument has anything whatsoever to do with the plausibility (or implausibility) of hedonism, or so I will argue.

Let's briefly review the argument. Nozick's basic claim is that we "care about things in addition to how our lives feel to us from the inside" (Nozick 1989, 104). In order to argue for this claim, Nozick asks us to imagine a quite extraordinary machine – a machine, which, once we've plugged in, would supply us with all kinds of pleasurable feelings. The pleasures we feel, though, would not correspond to any real experiences in the world, but would be created by chemically induced nerve-stimuli. After having set out the scenery, Nozick now asks us if we would plug into the machine. He adds that it is not a momentary decision, but one

[2] For a more detailed analysis of this charge, see Hildt 2018.

[3] For a paradigmatic, and pioneering, example of this practice, see Griffin 1986, 9. Concerning the overall situation of hedonism in the contemporary debates, Roger Crisp pointedly states that "these days hedonism receives little philosophical attention, and students are warned off it early on in their studies, often with a reference to Nozick" (Crisp 2006, 99).

[4] Cf. Matthew Silverstein's assessment of the situation: "Many of the most prominent philosophers of value – including James Griffin, David Brink, Stephen Darwall, and L.W. Sumner – take this thought experiment to be the definite response to hedonism and, more broadly, to all mental state theories of well-being" (Silverstein 2000, 282). We should add to this, though, that Sumner, despite his confidence in the experience-machine, rejects Nozick's reality-condition as implausible (cf. Sumner 1999, 157).

you make for the rest of your life. Nozick assumes, and probably rightly so, that the vast majority of his readers will answer: No! The crucial question now becomes: Why?

Nozick has a ready answer. According to him, the reason is that we sense that there is more to a good human life than mere pleasurable feelings. Therefore, Nozick concludes, hedonism is refuted: "there is more to life than feeling happy" (1989, 106).[5] In this concluding statement, it becomes clear that Nozick's argument is a version of the charge of systematic implausibility (cf. Secion I above).

For the present context, I want to leave aside the possibility that the majority may simply be wrong – the normative relevance of majority-arguments is a hugely problematic issue –, and I also won't go into the more general question of the relevance of empirical findings for normative questions. Instead, I want to draw attention to an alternative explanation of why we do not want to plug into Nozick's machine.

It is noteworthy that the existence Nozick pictures once we've successfully plugged in, appears to be radically different from our ordinary life as we know it: Instead of our everyday decision-making and acting, we will lie still, our body almost lifeless contained in some sort of nutrient solution. This surely is a case of radical alienation.

Radical alienation is nothing we cope with easily. This point is made vivid from Plato's allegory of the cave up to the famous *Matrix*-movies of Lilly and Lana Wachowski (formerly the Wachowski Brothers). If we add to this the thoroughly negative, for many downright repulsive, connotations that the image of a body plugged into a machine evokes in our culture – one immediately thinks of hospitals, life-prolonging procedures, and the like –, it seems that a deep desire to avoid this kind of situation, this form of radical alienation, might alone suffice to refrain from being plugged into Nozick's machine.

This is a different story from the one Nozick tells us. Most notably for our present context, it is an explanation for our wish to keep our distance from the "experience machine" which does not even touch upon the plausibility, or implausibility, of hedonism, and the question wether hedonism can provide an adequate theory of human well-being.

One could object that the explanation I am offering is more psychological than philosophical. This may be true. But, granted it is a possible and not altogether implausible explanation of the discomfort people feel when asked if they would plug in, there is a major consequence: It calls into question the alleged unambiguity of Nozick's explanation. Thus, it has

[5] Nozick's additional aspects of a good human life include a desire for authenticity, a desire of a stable connection to reality, and the desire that our beliefs about the world be true and accurate (cf. ibid.).

the power to mitigate the whole thought-experiment: It might turn out that the “experience machine” does not even concern the question of the plausibility of hedonism.

III The “Philosophy of Swine”-Objection, and the Independence of Hedonism

If my argument above is sound, Nozick’s “experience machine” does not prove to be a fundamental problem for hedonism. At the very least, it should be clear that what Nozick offers is not tantamount to justify a wholesale rejection of hedonism.

The underlying objection in Nozick’s argument, as we have seen, is the charge of systematic insufficiency: Hedonism is accused of being too simple to account for the complexity of human life. This objection goes back to an argument John Stuart Mill famously deals with in *Utilitarianism* (1998, II.3–II.10): Thomas Carlyle’s polemics that identifying human happiness with the experience of pleasure is equivalent to degrading humans to swine. The upshot is that hedonistic happiness is claimed to be not only systematically insufficient, but also unworthy.

To this charge, Mill presented his well-known solution: According to him, pleasures can differ not only in quantity and duration (like Bentham would have it), but also in quality. Thus, Mill argues, it is no problem for hedonism to assume that some kinds of pleasure are more desirable and valuable than others (1998, II.4).

While Mill appears to be content with his answer, his so-called qualitative hedonism, the vast majority of his interpreters today think that he runs into a dilemma: Either Mill’s hedonism reduces, upon closer inspection, to mere quantiative hedonism, or it introduces a value different from the feeling of pleasure, and thus leaves the basic premise of hedonism.[6] While this situation might seem to be first and foremost an interpretative issue – a matter of understanding the structure of Mill’s argument –, it has, however, direct consequences for the main question of this paper, namely, whether hedonism deserves a second chance:

[6] For the first horn of the dilemma, cf. in particular Mill’s discussion of the competent judges where it seems that the more elevated pleasures are the ones which, when enjoyed, produce a greater amount of pleasure (1998, II.5 and 6). For the second horn, see Mill’s description of the higher pleasures as “more noble” (1998, II.4 and 6), which reads as if Mill conceived it as a separate and genuine, distinctly qualitative value for measuring and comparing pleasures – which would be in conflict to his earlier statement according to which pleasure and the absence of pain are the only things good in themselves (1998, II.2). Crisp 1997, 32 provides a concise and helpful analysis of this dilemma.

If the introduction of the notion of a "quality" of pleasures produces the complications just mentioned, it might seem that hedonism, in order to deal with Carlyle's charge, needs to confine its notion of pleasure to mere quantitative matters.[7] Given this situation, the "Philosophy of Swine"-Objection, although much older than Nozick's argument, appears to be far more troublesome for contemporary hedonism.

If hedonism does not want to reduce its notion of pleasure to the dimension of quantity alone – and indeed I think it should not –, I suggest that contemporary hedonists should examine closely which of the problems Mill's theory runs into relate to hedonism *as such*, and which relate to hedonism *as a basis* ("theory of life") *for utilitarianism*, i.e. which follow from utilitarianism's specific demands. As a consequence, hedonism, as a theory of well-being, should claim its independence from any moral theory, and, given the current context, especially from utilitarianism. In this respect, I want to suggest one general, and two specific reflections.

The general reflection concerns the claim often found in today's discussions of hedonism that, if Mill's attempt fails, hedonism has to "fall back" to its quantitative dimension. It is worth noting that this claim assumes an anachronistic default-position: From Epicurus onwards, almost all hedonists (except, of course, Bentham) put forward theories which encompassed both quantitative *and qualitativ*e aspects of pleasure – and saw no particular problem in their compatibility.

The two more specific reflections both stem from the same observation of philosophical context: While the question of how to choose between different pleasures is, of course, a major question for all types of hedonism, the special urgency and thoroughness this question receives in Bentham and Mill is arguably due to their specifically moral, i.e. utilitarian context: Measuring pleasures, and intra-personally comparing them, are aspects which are of major concern for utilitarian thinkers: they arise from the need to determine what's the morally right action. They are not, in and by themselves, demands hedonism as a mere, and thus morally-neutral, theory of well-being needs to fulfill, at least not with the same urgency and thoroughness.

Instead of trying to achieve precise calculability, I want – this is the first specific reflection – to suggest that hedonism should do the exact opposite, and stress the vast plurality of pleasures: As a theory of well-being, hedonism should be concerned not so much with the exact calculability and comparability of its pleasures, but rather embrace, and stress, the many

[7] And this is, indeed, frequently assumed in the contemporary debate. Cf. vividly in White 2006, 41–2, 54.

ways in which we experience pleasure, and be open to the vast array of actions and things that induce and produce this sensation.[8]

A second specific reflection directly concerns the notion of "quality", which appears to be the source of so much trouble. Instead of understanding quality as a value-judgment – this is the way most interpreters seem to understand Mill, and he makes it easy enough for them to read him this way –, hedonism can make sense of quality in a different, and less problematic way: "Quality" can not only refer to a value-judgment, but also to a specific feature of the object in question. If I say "hiking has a distinct quality", I might judge hiking to be more valuable than other comparable activities. This is the standard understanding of "quality". I might, however, mean something quite different: namely, that hiking, as an activity, has certain specific attributes, like, for example, the pleasantly tired feet in the evening, the slowness of movement, etc. The central point in this respect is that quality in this second sense does not presuppose any judgment of value – and it were value-judgments which got Mill's qualitative hedonism into trouble in the first place.

IV Towards a Contemporary Hedonism – a Tentative Sketch

So where does all this leave us? I set out to make a case for hedonism, and to argue that it deserves to be taken seriously in the current debates on well-being. With regard to the argument that is commonly used today in order to show hedonism's alleged systematic insufficiency – Robert Nozick's "experience machine" –, I tried to show why hedonists need not be overtly troubled with this thought-experiment: There are good reasons to assume that Nozick's argument might not even touch upon the question of hedonism plausibility or implausibility.

The charge of systematic insufficiency, however, goes deeper, and does indeed pose problems which need to be answered. This has become clear in the discussion of the "Philosophy of Swine"-Objection and Mill's problematic, maybe even failed, attempt to answer it. Here, I argued that hedonism, as a theory of well-being, should claim its independence from any moral theory, and, in particular, from utilitarianism: Although it was utilitarianism which supplied hedonism with its first chance in modern times, it is time for hedonism to detach

[8] The plurality of pleasures is actually a point stressed by Mill himself, when he talks about the various and distinct pleasures that arise from the use of our different faculties, additionally to the pleasures of sensation: "the pleasures of the intellect, of the feelings and imagination, and of the moral sentiments" (1998, II.4).

itself from utilitarian concerns, if the theory aims to be a convincing account of well-being still worthy of discussion.

In connection to this last point, I have already argued that hedonism should endorse the plurality of pleasures, and that it can, and should, make use of the notions of both quantity *and* quality, understanding the latter as a statement about specific features, and not as a judgment of value.

Keeping with the programmatic tone of the paper, in what remains I will briefly present three key questions which I think every contemporary hedonism, as a theory of well-being, needs to address. I will also, albeit quite tentatively, suggest a possible direction an answer could take. What unites my suggestions here is the basic intuition that the very thing most people today find problematic with hedonism – its systematic simplicity – might turn out to be one of its major assets.

IV.a What Do We Talk about When We Talk about Pleasure?

Within the current debate, the few remaining hedonists have put forward two main candidates to answer this question: Roger Crisp argues that we should understand pleasure as a sensation, and that while "enjoyable experiences do indeed differ in all sorts of ways", they have one thing in common: "they all feel enjoyable" (2006, 110). Against this model, Fred Feldman has developed what he calls "attitudinal hedonism", according to which the pleasures relevant to human well-being are not sensations as such. Instead, he is concerned with those which are being directed in a certain way: with "attitudinal pleasures" we take in the existence of certain "state of affairs" (Feldman 2002, 611).

While there are reasons for and against both of these candidates, I want to suggest that if hedonism wants to make good use of its systematic simplicity, it should endorse an understanding of pleasure along the lines of Crisp's model, and thus conceptualize pleasure as a sensation. This model of pleasure not only keeps closer to our everyday-notion of pleasure,

but it also steers clear of the quite considerable systematic difficulties Feldman's "attitudinal hedonism" runs into.[9]

IV.b What Is the Basis of Explanation?

According to hedonism, pleasure, and the absence of pain, are the only determinants of human well-being. This basic premise is, of course (and has been throughout the history of ethical thought), as simple as it is provoking. One obvious, though far from easy, question that hedonists need to address is: What, then, is pleasure? And at the very bottom of this question lies the issue of the basis of explanation: How deep can we go in explaining what pleasure is, and does?

Here, I think that one promising route for contemporary hedonism could run along the lines of T. M. Scanlon's methodological approach in *What We Owe to Each Other*. Scanlon (1998, 17) famously starts his book by announcing that he will take "the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it". Hedonists could try to do something similar with regard to their notion of pleasure. Such a claim of the primitive nature of pleasure would also be in line with understanding pleasure as a sensation, and it would also fit with the idea of systematic simplicity as an asset.[10]

IV.c What about False, and Morally Corrupt Pleasures?

Where people in the current debates get involved with hedonism, two problems appear to many to be especially troublesome for hedonistic theories of well-being: The case of "false" pleasures – where we experience pleasure because of an erroneous assessment of the world around us, like in Shelley Kagan's famous example of the "deceived businessman" (cf.

[9] My main worries with Feldman's account include: the question whether Feldman's hedonism is indeed an alternative to a sensationalist account of pleasure, or rather an unsubstantiated selection among pleasures which are to count in matters of our well-being; the question of the methodological soundness of Feldman's argument to counter possible objections by including his answer into the theory itself (and thus moving from *Intrinsic Attitudinal Hedonism* to *Veridical Intrinsic Attitudinal Hedonism* to *Desert Adjusted*, viz. *Double Desert Adjusted Intrinsic Attitudinal Hedonism*); and the more general question if Feldman's account is more accurately described as a type of desire-fulfillment theory, which would be tantamount to leaving hedonism behind altogether.

[10] Interestingly enough, hedonism here gets seconded by a philosopher whom few recognize to be a hedonist: Immanuel Kant. In his *Metaphysics of Morals*, Kant sets out to explain his concept of pleasure, and then states that: "pleasure and displeasure cannot be explained more clearly in themselves; instead, one can only specify what results they have in certain circumstances, so as to make them recognizable in practice" (1997, 6:212).

Kagan 1997, ch. 2) – and the case of "morally corrupt" pleasures, where a person takes pleasure from morally base actions so that engaging in these actions, although they would be rejected by any moral standard, seems to add to this person's well-being.[11]

Although many nowadays seem to assume that these two kinds of "problematic" pleasures pose systematic difficulties for hedonism as a theory of well-being, I want to suggest that they might as well not. If hedonism stays true to its systematic simplicity, it can, and should, claim that a pleasure is a genuine pleasure, despite and independent of the question whether its circumstances include erroneous believes about the world or the motives of other people. Pleasures of the latter kind may have unfortunate consequences, but that does not diminish their felt intensity when they are experienced, and thus, neither adds, nor subtracts, their impact on our well-being.

The case with morally corrupt pleasures is quite similar. As I have argued above, hedonism is not a theory of morality, but of well-being. As such, it accounts for the things that make a life feel good for the person who is experiencing it. How we deal with persons who behave morally wrong, and in ways which are not compatible with our values, is a different matter; it is a genuine moral question.

As I have said in the outset of this last section, much more will need to be said about contemporary hedonism, and about how to address these three questions. What I have tried to do in scraping on the surface of each of them, so to speak, was first and foremost to make the case that hedonism is nowhere near of being out of the picture. It deserves a second, and proper chance.

References

[1] Crisp, Roger. 1997. *Mill on Utilitarianism*. Abingdon and New York: Routledge.

[2] _____. 2006. *Reasons and the Good*. Oxford: Clarendon Press.

[3] Feldman, Fred. 2002. "The Good Life. A Defense of Attitudinal Hedonism." *Philosophy and Phenomenological Research* 65 (3): 604-28.

[4] Griffin, James. 1986. *Well-Being. Its Meaning, Measurement and Moral Importance.* Oxford: Clarendon Press.

[11] For an instructive discussion on this point, see Feldman 2002, 620–2.

[5] Hildt, Moritz. 2018. “Das vermeintliche Ungenügen des Hedonismus.” *Zeitschrift für Ethik und Moralphilosophie* 1 (1): 75–89.

[6] Kagan, Shelley. 1997. *Normative Ethics*. Boulder, CO: Westview Press.

[7] Kant, Immanuel. 1797/1997. “The Metaphysics of Morals.” In *Practical Philosophy*, edited by Mary J. Gregor. Cambridge and New York: Cambridge University Press: 353–604.

[8] Mill, John Stuart. 1871/1998. *Utilitarianism*. Edited by Roger Crisp. Oxford and New York: Oxford University Press.

[9] Nozick, Robert. 1989. *The Examined Life. Philosophical Meditations.* New York: Touchstone.

[10] Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

[11] Sumner, Leonard Wayne. 1999. *Welfare, Happiness and Ethics*. Oxford and New York: Oxford University Press.

[12] Silverstein, Matthew. 2000. “In Defense of Happiness. A Response to the Experience Machine.” *Social Theory and Practice* 26 (2): 279-300.

[13] White, Nicholas. 2006. *A Brief History of Happiness.* Malden, MA: Blackwell.

[14] Wilson, Catherine. 2015. *Epicureanism. A very short introduction*, Oxford and New York: Oxford University Press.

Brandt's Rule Utilitarianism and the Future. Replies to the Demandingness Objection

Stefan Hofmann, University of Tübingen, Germany

Abstract

For some philosophers rule consequentialism (RC) is the most plausible form of ethical theory: RC can rebut many of the well known objections raised against classical act utilitarianism. Yet, RC has its own weaknesses. One powerful objection against RC is that even this form of consequentialism might get too demanding: If we think of the future of a "broken world", RC might loose its moderate credentials (Mulgan 2015). This paper unfolds in three steps: First, I will distinguish five versions of the so-called demandingness objection against rule consequentialism (section I). Then the rule consequentialist conception of ethics proposed by Richard Brandt (Brandt 1979, 1992, 1996) is presented (section II). Brandt has argued for an influential version of *ideal acceptance rule utilitarianism*. Some amendments are needed to adjust Brandt's ethics to a global scheme, but this can be granted. Section III proposes options open to Brandt in order to respond to the demandingness objection regarding the future of a broken world. Brandt's version of rule utilitarianism includes many resources that help to answer the objection (e.g. a concept of supererogatory acts, Brandt's idea of "fully rational persons" etc.). Section IV will discuss and evaluate Brandt's options. The paper argues that the most convincing responses open to Brandt implement deontological propositions into his moral theory. The demandingness objection in view of a broken future effectively urges to review the commitment to the rule utilitarian account.

Introduction

For some moral philosophers rule consequentialism is the most plausible form of ethical theory. Rule consequentialism seems to embody many of the attractive elements of consequentialism. Yet, it can rebut many of the well-known objections raised against classical act utilitarianism.

One powerful objection against rule consequentialism is that even this form of consequentialism might get too demanding. At the beginning of the 21st century, the prospects of the future are uncertain. Climate change, environmental pollution or nuclear disaster might lead to a "broken world", where resources do not suffice to supply for the needs of all. In so far as rule consequentialism tries to maximize happiness over time, the prospect of a broken world would significantly alter the rules for the present. Thus, rule consequentialism loses its moderate credentials (Mulgan 2015).

This paper reconsiders the influential version of rule utilitarianism presented by Richard Brandt (1979, 1992, 1996) in order to look for possible replies to the above-mentioned objection. The argument unfolds in five steps: In section I, I distinguish five versions of the so-called demandingness objection in order to focus on one specific version. Section II presents the rule utilitarian conception of ethics proposed by Richard Brandt. Section III proposes the options open to Brandt in order to reject the demandingness objection regarding the future of a broken world. In section IV, I will discuss and evaluate the different options open to Brandt. Finally, Section V summarizes the argument. The paper concludes that the most convincing responses open to Brandt implement deontological propositions into his moral theory. The demandingness objection in view of a broken future effectively urges to review the commitment to the rule utilitarian account.

I The Demandingness Objection and the Prospects of the Future

The objection that utilitarianism represents an implausible account of normative ethics since it entails obligations that are incredibly demanding for the agent is familiar. It was first raised against act utilitarianism, but since 1990 it has also been raised against rule consequentialism (Hooker 1990, 1991, 2000; Mulgan 2001, 2015; Arneson 2005; Hills 2010; Tobia 2013). In fact, there are at least five versions of this objection:

1. the demandingness objection under "full compliance" in our present world: Rule consequentialism may be deemed too demanding, since the afflictions of the present world might call for rules that demand too much even though all agents take their share in the necessary effort.

2. the demandingness objection under "partial compliance" in our present world: Rule consequentialism might get too demanding, since "partial compliance" with its rules is what might be expected in actual societies. At least, if the utilitarian code contains a "prevent disaster rule", partial compliance inevitable seems to have this effect (Hooker 1990; Carson 1991; Hooker 1991; Hooker 2000).

3. the demandingness objection in view of possible worlds: Rule consequentialism could lead to overly demanding rules, if we consider possible worlds where circumstances are much worse than in our own world: If billions of people are starving because of apocalyptic disasters, rule consequentialist considerations might lead to very demanding rules for everyone (Mulgan 2001; Arneson 2005; Portmore 2009).

4. the demandingness objection in view of a broken future: It might be the prospect of a broken future that calls for excessively demanding rules (Mulgan 2015).

5. the objection may be raised, if we claim that animals are to be included: Some authors argue that rule consequentialism would get too demanding, if we included animals among the beneficiaries for whom the rules are to be made (Hills 2010).

In what follows, I will confine myself to the objection concerning the prospect of a broken future. For our discussion, consider the following scenario: In 2050 the prospects for the preservation of our natural environment have decreased significantly. After a nuclear war in the late 2030s, scarcity of water and many other important resources is to be expected within decades. Scientific research has made tremendous progress, but, unfortunately, no major break-through is to be expected. Clean water and other resources will run out in the 2080s, if they are not proportioned rigorously.

Let us see how the scenario influences the maximizing moral code of the year 2050. Since utilitarian consequentialism tries to maximize happiness over time, the prospects for the future will in any case make a huge difference for the determination of the moral rules for 2050. If the scenario fittingly describes the prospects of 2050, the rules to be propagated then will be highly demanding: Water will be proportioned rigorously. In an extreme situation, even human rights may be abandoned. As Timothy Mulgan rightly says: "if it is not possible for everyone to survive, then there is nothing that can meaningfully be guaranteed to everyone." (Mulgan 2001, 109). In order to secure the same chances for the presumed population of the 2090s, the utilitarian rules of 2050 would require strictest saving of clean water. They would even reduce the presumed right to life to mere chances to survive (cf. Mulgan 2015, 110). If prospects are really bad, a person of the 2050s, let us call him Frank, may be morally allowed to drink 0.3 liters of clean water the day. Beyond that quantity of clean water he might be required to resort to heavily contaminated water supplies that are still available.[1] This conclusion, however, seems implausibly demanding. It does not seem plausible that the contingencies of the distant future of the 2090s can have that much weight as to suspend the right to life in 2050. To be sure, there is much reason to argue for rigid saving of water supplies in our scenario. But it seems highly counterintuitive to assume that the contingent developments of the distant future may override the basic human rights of actually living people like Frank.

[1] In 2018 the World Health Organization estimated that by 2025 half of the population on earth will live with water stress. Already in 2015 two billion people had to use contaminated drinking water (cf. WHO 2018).

II The Rule Utilitarian Moral Theory of Richard Brandt

Richard Brandt faced the first version of the demandingness objection, the demandingness objection under "full compliance" in our present world, in the 1990s. He argued that a yearly donation of 600,- $ by all inhabitants of the wealthy nations on earth would suffice to provide for a life above the poverty line for all who are in need. Brandt's argument for this conclusion draws on empirical informations supplied by the *World Development Report* by the World Bank (Brandt 1996, 229-36). However, since this argument does not treat the prospects of future developments, it may be set aside. What is needed for our purpose is a short overview of Brandt's moral theory as a whole. This theory and his main argument for rule utilitarianism is presented in his main opus *A Theory of the Good and the Right* (1979). In what follows, I will shortly introduce Brandt's main theses on the foundations of normative ethics as presented in his *Theory*.

According to rule utilitarian ethics an action is morally wrong, if it is prohibited by a set of rules which prove to be happiness maximizing rules. As most classic utilitarians, Brandt claims that happiness is the only intrinsic value. The deontic status of an action is therefore to be considered as a function of its contribution to happiness. Rule utilitarianism differs from act utilitarianism, however, in the way it applies the consequentialist principle: According to the rule utilitarian it is not the individual action that ought to maximize happiness. Rule utilitarians propose to go *indirect*. Brandt claims: Moral rules are to be judged by utilitarian criteria, individual actions are to be judged on the basis of these rules.

Brandt proposes to speak of a "pluralistic welfare-maximizing moral system" (Brandt 1979, 286), since the moral code proposed by his rule utilitarianism contains more than just one rule. The argument which inaugurates this kind of utilitarianism is quite complex. Since it draws heavily on empirical psychology, it has been suggested that Brandt's argument should be classified as "foundationalism in science" (Mitchell 1996, 330; cf. Timmons, 1987). And there is, of course, much reason to this claim. Even Brandt himself agreed to this classification (cf. Brandt 1995). However, his argument also contains contractualist considerations: Brandt proposes to think of a hypothetical choice of a moral system for one's own society. After reflecting on how to define the concept of a fully rational person, i.e. a fully informed person, he suggests that the deontic status of an action should be determined by the code which fully rational persons would choose for their own society: "I suggest, however, that we assign 'is morally wrong' the descriptive meaning 'would be prohibited by any moral code which all fully rational persons would tend to support, in preference to all others or to

none at all, for the society of the agent, if they expected to spend a life-time in that society'." (Brandt 1979, 194; cf. Kavka 1993)

Obviously, one crucial question for this argument is how to define the concept of fully rational persons. For these are the ones who choose the moral code. Brandt proceeds in two steps. He first defines the concept of a fully rational action. Then, the concept of a fully rational person is clarified by appeal to this concept of rational actions. Brandt considers an action to be fully "rational", if the desires and aversions which lead to that action have been fully criticized by facts and logic and if the cognitive inputs present at the time of decision have been optimal as well (Brandt 1979, 11). The fully rational person, correspondingly, is a person "in whom the mechanisms underlying desire, pleasure, and action have been fully suffused by relevant available information" (Brandt 1979, 88).

There are two further characteristics of Brandt's Code that should be mentioned. First, the version of rule utilitarianism which Brandt tries to defend is defined as an *ideal acceptance rule utilitarianism*: It is an "ideal" rule utilitarianism due to the fact that it advocates ideal rules in contrast to the actual rules of a society. It is an example of "acceptance rule utilitarianism", since it is supposed to count not only the consequences of compliance with its rules but also the costs and benefits that are involved in the teaching and the internalization of that code in society (Brandt 1979, 198). Second, when speaking about the "society of the agent", Brandt was mostly thinking of the societies of countries or nations. But in the 1990s he suggested that a world-wide code would be adequate for global issues (Brandt 1992, 192; Brandt 1996, 227-36). We may at this point simply assume that Brandt's theory of justification of moral codes might equally well be used to justify a global moral code.

For our current interest in replies to the demandingness objection Brandt's view about what fully rational persons would choose is fundamental. The way Brandt proceeds is important because it introduces two levels of argumentation for the justification of specific moral rules: Brandt argues for a meta-normative rationale for the choice of a welfare-maximizing moral code before starting the investigation in order to find out the content of the specific rules. He first argues at the level of a meta-normative choice scenario which may be abbreviated as L_{cs} (for $L_{\text{choice scenario}}$). Only after that does he start to reflect about the specific content of the rules of his code. This second level of reasoning may be called the "level of rule determination", abbreviated as L_{rd}. For it is at this level, that Brandt is concerned with empirical investigations and the utilitarian calculus that determine the content of particular rules.

With respect to questions of demandingness two features of Brandt's social moral code seem to be highly relevant: Brandt's positive evaluation of personal freedom to pursue

one's own projects, and his concept of supererogatory acts. Each of these general features is established by some reasoning. For the discussion of possible replies to the demandingness objection it is, however, important to see at which level of reasoning these features are introduced:

1. Brandt argues in a consequentialist way that agents should be free to pursue their own projects: The opportunity to pursue one's own projects generates a sense of freedom and pleasure (Brandt 1992, 86). The transmission of a code that restricts freedom, on the contrary, would cause a lot of frustration and a huge loss in happiness. This argument is conducted at the level of consequentialist rule determination (L_{rd}).

2. The concept of supererogation is part of Brandt's definition of a moral code (Brandt 1979, 168, 172 and 201). But in addition, Brandt adduces consequentialist considerations for the acceptance of this concept. He claims that it is beneficial for a society to acknowledge the idea of supererogation so that some very beneficent but also costly actions are recommended even though they are not required by the rules of its moral code. Thus, people will be encouraged to perform demanding or even heroic deeds since these are praised although they are not required. The concept of supererogation allows to omit the costs of making those deeds obligatory for everyone (cf. Brandt 1979, 289).

III Brandtian Replies to the Demandingness Objection in View of a Broken Future

Brandt's moral theory obviously offers at least five ways to react to the demandingness objection in view of a broken future exemplified in the scenario of Section I:

1. Brandt argues that agents should be free to pursue their own projects since this freedom probably maximizes happiness in society. Thus, the expected gain in overall happiness generated by freedom of action could justify Frank in drinking more than 0.3 liters. This argument is brought forward on the level of consequentialist rule determination L_{rd}.

2. Brandt's rule utilitarianism justifies the concept of supererogatory acts. It is one option for Brandt to claim that his moral code would highly recommend to spare water even though Frank is not morally required to do so. This point is again made

on L_{rd}. But maybe Brandt would also hold that the concept of supererogation is part of the definition of a moral code, thus making an additional point by defining the framework of his choice scenario.

3. Brandt's choice scenario suggests the choice of a social moral code for a society in which the agents – and possibly their children (cf. Brandt 1996, 240) – will live a lifetime. The significance of the distant future may therefore be disputed at the level of the apparently meta-normative choice scenario L_{cs}. This means that the framework of Brandt's choice scenario allows to discount the future.

4. It is open to Brandt to argue that "fully rational persons" would choose to make significant aggregate contributions to the happiness of future people, but not unreasonably high contributions. A similar line of argument is put forward by Brad Hooker in his more coherentist argument for rule consequentialism (Hooker 2000, 150 and 166). In the case of Richard Brandt, this reasoning would be established in terms of the meta-normative choice scenario L_{cs}.

5. Brandt could of course also bite the bullet and claim that our common-sense intuitions are not reliable in the case at hand and contend that the conclusions of our scenario are correct: This means, Frank of the 2050s is morally not allowed to drink more than 0.3 liters of clean water the day. This reply is very much in touch with Brandt's low esteem for reference to moral intuitions. Yet, we may put it aside. It would lead too far to discuss the epistemic value of strong moral intuitions. And, as it seems, despite his official disregard for intuitions Brandt's moral philosophy tries to approximate utilitarianism to common-sense morality in many ways.

IV The Evaluation of Brandt's Options

One remark seems to be in place before our discussion of the five options mentioned above. The argumentation of Brandt's main opus (Brandt 1979) does not completely succeed in its project of justifying a thoroughgoing utilitarian morality. Brandt himself acknowledged that (Brandt 1993). The concept of a code for one's own society might be understood in a way that restricts morality to a particular society, to the exclusion of certain demands from outside. The idea of a society in which oneself and one's children are supposed to spend their life would therefore provide the option to discount the value of the distant future. Yet, Brandt himself did not pursue these tracks to restrict the demandingness of his theory. On

the contrary, as Brandt's works clearly show, he considered it to be a weakness of his argument, that he was not able to establish a completely universalistic morality (Brandt 1993, 246). His intention was explicitly to include the needs of future generations (Brandt 1996, 133 and 142). The idea of discounting the future by reference to Brandt's concept of the society of the agent (and his children) may, therefore, be put aside. This kind of argument would surely reject the demandingness objection. It abandons, however, one of the classical utilitarian tenets; and Brandt himself refused to take this line.[2]

With that in mind let us turn to the discussion of the options introduced in section III. Options one and two may be treated together: Here, Brandt could adduce consequentialist arguments. For option two conceptual considerations might be added. The virtue of an argument via consequentialist considerations is obvious: This defense avoids any non-utilitarian assumptions. However, the reasoning of these replies cannot reject the demandingness objection exemplified in our scenario. If the freedom to pursue one's own projects is exclusively justified by consequentialist considerations, the reasons for its introduction will surely be outweighed by the even more urging consequentialist reasons which flow from the dire need of others as described in our scenario. This applies to Brandt's acceptance rule utilitarianism as well as to other versions of rule utilitarianism (cf. Carson 1997, 92). A similar reasoning may be put forward concerning the concept of supererogatory acts of option two: Frank's health projects must surely be set aside, if their impact in the utilitarian calculation is measured against the lives of future people in the 2090s. The alternative open to Brandt would be an argument at the level of his choice scenario (L_{cs}). Yet, if this contractualist argument is supposed to establish normative conclusions which could not be established at the level of consequentialist rule determination (L_{rd}), Brandt's code would appear to contain genuine elements of deontological morality. Consequentialism purports to derive the deontic status of actions in terms of consequences alone. Deontological theories, on the contrary, deny what consequentialism affirms: They reject the idea that the right is solely to be determined as a function of what realizes the best balance of good over evil (cf. Frankena 1973, 14-15; Timmons 2013, 111). Thus, if Brandt uses his choice scenario in order to introduce normative considerations that cannot be established by consequentialist reasoning, he in fact falls back to deontology. What first seemed to be a meta-normative kind

[2] One could, of course, abandon the temporal impartiality of classical utilitarianism and argue for some kind of sophisticated rule consequentialism that incorporates a discount rate. However, discounting the future would drop a central point of utilitarian thinking. Cf. Cowen and Parfit 1992; Gesang 2011; Birnbacher 2013, 189-196 and 413-417; Kaczmarek 2017, Mulgan 2017. I owe this point concerning Brandt's code and the option to discount the future to a comment by Thomas L. Carson.

of contractualist choice scenario designed to establish a thoroughgoing consequentialist moral theory would turn out to be itself a decisive means for normative conclusions.

The third and forth replies seem to provide a more solid ground to reject the demandingness objection. Yet, both of them use the contractualist reasoning of Brandt's meta-normative choice scenario L_{cs}. As mentioned above, this means to leave the grounds of pure consequentialism. If there are valid reasons why some consequences do not count as much as others, the resulting normative theory will be a hybrid theory. Brandt's choice scenario which at first seemed to be a purely hypothetical means to justify utilitarianism (and which was therefore described as a "meta-normative" foundation of his moral philosophy) would be transformed into a decisive part of his normative theory.

One could of course point to the concept of "fully rational persons" in order to specify reply four: One could argue that "fully rational persons" would choose to make significant aggregate contributions to the happiness of future people but not unreasonably high contributions. Yet, it seems that would presuppose a normative concept of practical rationality which again leads to "dangling" deontological propositions (Kagan 1989, 14). Brandt's utilitarianism would then hold to an agent-centred prerogative and to corresponding rules which it cannot support on its own ground. If the rule utilitarian account is to be defended, no result of consequentialist reasoning may be classified as unreasonable. Brandt himself, of course, rejects the idea of a substantially normative concept of practical rationality (Brandt 1989).

V Conclusion

It has been shown that Brandt's moral theory provides many resources that help to answer the demandingness objection concerning the prospects of a broken future world. Yet, Brandt's most convincing replies lead to an acceptance of deontological propositions into the consequentialist account. If our present considered moral judgements are to be accepted, the prospect of a broken future constitutes a tough challenge for Brandt's form of rule utilitarianism. Brandt's contractualist argument seems to recommend restrictions against overly demanding rules. The acceptance of these restrictions, however, introduces "dangling" deontological propositions into his otherwise rule utilitarian account of normative ethics. This means, the demandingness objection successfully urges to accept deontological constraints.

References

[1] Arneson, Richard. 2015. "Sophisticated Rule Consequentialism. Some Simple Objections." *Philosophical Issues* 15 (1, Spring): 235–51.

[2] Brandt, Richard B. 1979. *A Theory of the Good and the Right*. Oxford: Oxford University Press.

[3] _____. 1989. "Practical Rationality. A Response." *Philosophy and Phenomenological Research* 50 (1, Spring): 125–30.

[4] _____. 1992. *Morality, Utilitarianism, and Rights.* Cambridge: Cambridge University Press.

[5] _____. 1993. "Comments" In *Rationality, Rules, and Utility. New Essays on the Moral Philosophy of Richard B. Brandt*, edited by Brad Hooker, 207–48. Boulder: Westview Press.

[6] _____. 1995. "Foundationalism for Moral Theory." In *On the Relevance of Metaethics. New Essays on Metaethics*, edited by Jocelyne Couture and Kai Nielsen, 51–65. Calgary: University of Calgary Press.

[7] _____. 1996. *Facts, Values, and Morality.* Cambridge: Cambridge University Press.

[8] Carson, Thomas L. 1991. "A Note on Hooker's 'Rule-Consequentialism'." *Mind* 100 (1): 117–21.

[9] _____. 1997. "Brandt on Utilitarianism and the Foundations of Ethics." Review of *Morality, Utilitarianism and Rights*, by Richard B. Brandt. *Business Ethics Quarterly* 7 (1): 87–100.

[10] Cowen, Tyler, and Derek Parfit. 1992. "Against the Social Discount Rate." In *Justice Between Age Groups and Generations*, edited by Peter Laslett and James S. Fishkin, 144–61. New Haven: Yale University Press.

[11] Gesang, Bernward. 2011. *Klimaethik*. Berlin: Suhrkamp.

[12] Frankena, William. 1972. *Ethics*. Englewood Cliffs: Prentice Hall.

[13] Hills, Alison. 2010. "Utilitarianism, Contractualism and Demandingness." *Philosophical Quarterly* 60 (239): 225–42.

[14] Hooker, Brad. 1990. "Rule-Consequentialism." *Mind* 99 (393): 67–77.

[15] _____. 1991. "Rule-Consequentialism and Demandingness: A Reply to Carson." In: *Mind* 100 (2): 269–76.

[16] _____. 2000. *Ideal Code, Real World. A Rule-consequentialist Theory of Morality*. Oxford: Clarendon Press.

[17] Kaczmarek, Patrick. 2017. “How Much is Rule-Consequentialism Really Willing to Give Up to Save the Future of Humanity?” In: *Utilitas* 29 (2): 239–49.

[18] Kavka, Gregory. 1993. “The Problem of Group Egoism.” In *Rationality, Rules, and Utility. New Essays on the Moral Philosophy of Richard B. Brandt*, edited by Brad Hooker, 149–63. Boulder: Westview Press.

[19] Mitchell, Haney. 1996. “An Annotated Bibliography on Moral Epistemology.” In *Moral Knowledge? New Readings in Moral Epistemology*, edited by Walter Sinnott-Armstrong, and Mark Timmons, 326–37. New York: Oxford University Press.

[20] Mulgan, Timothy. 2001. *The Demands of Consequentialism*, Oxford: Clarendon Press.

[21] _____. 2015. “Utilitarianism for a Broken World.” *Utilitas* 27 (1): 92–114.

[22] _____. 2017. “How should utilitarians think about the future?” *Canadian Journal of Philosophy* 42 (2-3): 290–312.

[23] Portmore, Douglas. 2009. “Rule-Consequentialism and Irrelevant Others.” *Utilitas* 21 (3): 368–76.

[24] Timmons, Mark. 1987. “Foundationalism and the Structure of Ethical Justification.” *Ethics* 97 (3): 595-609.

[25] Tobia, Kevin. 2013. “Rule Consequentialism and the Problem of Partial Acceptance.” *Ethical Theory and Moral Practice* 16 (3): 643–52.

[26] WHO. 2018. “Drinking-water.” Accessed December 5, 2018. http://www.who.int/news-room/fact-sheets/detail/drinking-water.

Bentham's Theories of the Rule of Law and the Universal Interest

Michihiro Kaino, Doshisha University Kyoto, Japan

Abstract

It has been very rare in the Bentham studies that Bentham had an idea of the rule of law. But according to Professor Postema, Bentham was arguing for 'the reflexive dimension of the rule of law' meaning that those in power are also held accountable under the law and are subject to it. In his forthcoming book, Postema focuses on Bentham's theory of constitutional constraints on the sovereign, which Bentham called '*leges in principim*'. Postema's point that Bentham, who tried to maximize the accountability of officials, was analyzing the conditions of rule of law or law's ruling in a community is convincing. However, it is difficult to assume that the majority, who are motivated by self-interests, would exercise the moral sanctions of Public Opinion Tribunal when the interests of minority, which have relatively little effect on those of majority, are violated by some legislators or officials. And Bentham himself seems to argue that it is difficult to rely on individuals as they are interested more in their own or particular interests than in the general interests of community or the universal interest. So, it would be argued that the system with judicial review is better as it protects the rights of minority better. However, I want to argue that Bentham was in a sense a precursor of those modern theorists who try to design some architecture for deliberative democracy. For example, in 'A Table of the Springs of Action', Bentham provides a description of 'deontologists' who are expected to lead ordinary citizens, supplying such motives to them that will promote the happiness of society, or the universal interest.

Introduction

I would like to focus on two chapters of Professor Postema's forthcoming book, *Utility, Publicity, and Law*, which are 'The Soul of Justice: Bentham on Publicity, Law and the Rule of Law' and 'Interests: Universal and Particular'. Postema's point that Bentham was analyzing 'the conditions of *law's ruling* in a political community' (Postema forthcoming, ch. 12) is new and convincing. And I think Postema's point would help to put Bentham's theory of law in the English tradition of the rule of law.

However, it is difficult to assume that the majority, who are motivated by self-interests, would exercise the moral sanctions of public opinion tribunal when the interests of minority, which have relatively little effect on those of majority, are violated by some legislations. I would like to argue that Bentham was in a sense a precursor of those modern theorists who try to design some architecture for deliberative democracy and also that this aspect of Bentham and the paternalistic nature of Bentham's theory would strengthen his theories of the rule of law and the universal interest.

I Bentham and the English Tradition of Rule of Law

In 'Utility and Command', Postema notes that Bentham's legal theory departs 'significantly from other "command theories" of law' in that 'he held that whether a sanction is attached to a legal directive is a contingent matter' (Ibid., ch. 8). And Postema focuses on Bentham's theory of constitutional constraints on the sovereign, which Bentham called *leges in principim*. According to Postema, Bentham departed from 'the simpler Hobbes-Austin model of commands' in that 'whether a sanction is attached to a legal directive is a contingent matter' (Ibid.). And when Postema analyzes Bentham's *leges in principim*, he emphasizes that 'Bentham was even willing to say explicitly that the rules and standards issuing from firm expectations of public opinion themselves constitute a kind of *law*' (Ibid., ch. 12).

On the other hand, the following statement of Bentham in *Of the Limits of the Penal Branch of Jurisprudence*, which is quoted by Postema, seems to show that Bentham faithfully followed the English tradition of the rule of law, which has been sustained by the distinction between legal and political sovereign.

> The mandate of the sovereign, be it what it will, can not be illegal: it may be impolitic; it may even be unconstitutional: but it can not be illegal. It may be unconstitutional, for instance by being repugnant to any privileges that may have been conceded to the people whom it affects: but it would be perverting language and confounding ideas to call it *illegal* (Bentham 2010, 10-11).

In *A Comment on the Commentaries*, Bentham looks to a 16th century English example when, in the reign of Henry the 8th, '(t)he Legislature made over its whole power to the King alone' and '(t)he King's Proclamations were enacted in general terms without reserve to have the force of Laws'. According to Bentham, This is an example when '(t)he constitution was actually destroyed'. And he also adds that 'I will take up arms whosoever the Legislature pass an Act, giving the force of Statutes in all cases and for this country that I write in, England, to the Sovereign's Proclamations', and that 'Legislature would act consistently and legally in setting a price upon my head' (Bentham 1977, 56-57).

In *Securities Against Misrule*, Bentham treated the Petition of Right as one of the 'legislative arrangements that have been established or have been endeavoured to be established for the security of the governed against the governors' (Bentham 1990, 23). As Postema notes, *leges in principim* – constitutional constraints on the sovereign – 'impose legal duties, nevertheless, by virtue of their enforcement by the social or "moral" sanction

of public opinion' (Postema forthcoming, ch. 8). And as Professor Lieberman shows, Bentham thought that appeals to Magna Carta and the Bill of Rights 'helped focus public opinion on the abuses of the current political order, by taking advantage of well-established and well-publicized standards for the critical evaluation of public power' (Lieberman forthcoming).

Postema shows an excellent theory of rule of law and attributed it to Bentham:

> The rule of law is not robust in a community – law does not effectively rule there – if some of those who wield political power and hold others accountable to the law are not themselves accountable under law. … (H)e [Bentham] analyzed the background conditions and engineered the supporting institutions needed for a comprehensive and effective architecture of accountability (Postema forthcoming, ch. 12).

Postema call this 'the *reflexive* dimension of the rule of law' in that '(t)hose in power as well as those subject to that power must be subject to the law' (Ibid.).

And as we saw above, Bentham relied on *leges in principim* – constitutional constraints on the sovereign –, which is enforced by the social or moral sanction of public opinion to make the sovereign be subject to the law. And it is possible to say that Dicey in the 19th followed Bentham and emphasized the reflexive dimension of the rule of law. Dicey argues that there is an external limit to the power of sovereign which 'consists in the possibility or certainty that his subjects, or a large number of them, will disobey or resist his laws' and that 'widespread resistance would result from legislation which, though legally valid, is in fact beyond the stretch of Parliamentary power' (Dicey 1982, 30, 32).

II The Universal Interests and the Minorities

It would be argued that the system with judicial review is better as it protects the rights of minority better. It is difficult to assume that the majority, who are motivated by self-interests, would exercise the moral sanctions of public opinion tribunal when the interests of minority, which have relatively little effect on those of majority, are violated by some legislations.

It is usually argued that Bentham's theory of utility is not based on a simple aggregation of pleasures of the people of a society as suggested by Rawlsian interpretation of utilitarianism. As Professor Schofield shows, Bentham's argument is that legislation or policies will,

through necessity, seek universal interests, because it is difficult to pursue particular or sinister interests within a representative democracy. Bentham may be too optimistic about democracy, but, as Schofield argues, the fact that Bentham was an individualist cannot be denied. Thus, Bentham's universal interests incorporated each member's security of person and of property (Schofield 2014).

However, as Postema notes, 'Bentham counsels us not to expect willing and spontaneous sacrifice of personal or particular interest to the universal interest' (Postema forthcoming, ch. 6). So, even when it is in the universal interest to exercise the moral sanctions of public opinion when the interests of minority are violated by some legislations, the majority would not take that course as it may not be based on their personal or particular interest. As to the difference between the particular interest and the universal interest, Bentham writes:

> The individuals who compose the particular interest always are, or at least may be – and have to thank themselves and one another if they are not – a compact harmonizing body – a chain of iron: the individuals making the universal interest are on every such occasion an unorganized, uncombined body – a rope of sand. (Bentham 1962, 96)

With regard to the universal interest, Postema writes,

> The argument for understanding the compositional principle in terms of the universal interest rests on the claim that the way to respect this fundamental value, understood as equal for every person, is to focus moral attention primarily on those interests broadly compatible with the interests of others in the community, and especially those interests all share or can come to share. (Postema forthcoming, ch. 6)

Certainly, Bentham's legislators are supposed to pursue the common interests of all rather than the interests of majority. And we should also note the importance of the public opinion tribunal and the publicity in Bentham's thought. According to Postema, although 'particular passions, narrowly focused interests can still influence public opinion', Bentham thought that '(t)hrough participation in debates at the local level, members of the community [would] come to recognize the public dimensions of their concern to secure themselves against depredation or oppression and their individual part in that universal interest, and at the same time [would] come to understand the difficulty of enlisting the cooperation of others to advance their private, "sinister" interest' (Postema forthcoming, ch. 13). However, Bentham was a realist and suspicious of the competence of the public and the public opinion tribunal. For example, although Bentham often argued that consensual homosexual acts are harmless, he also proposed that consensual homosexual acts should be punished by banishment instead of by hanging, in the face of strong prejudice against homosexual acts

and the expected unpopularity of decriminalised homosexual behaviour in nineteenth-century England. Concerning women's franchise, although Bentham's position was that the exclusion of women was based on prejudice, he 'was prepared to surrender to that prejudice, and wait the arrival of more enlightened attitudes' (Quinn 2014, 79).

So, we should, I suppose, look to Bentham's paternalistic aspects, such as his theory on indirect legislation and his discussion about the 'deontologists'.

Some of Bentham's indirect legislation try to make people find 'the true interests' by some indirect means. For example, as Quinn shows, Bentham suggested that 'if you were able to demonstrate that, for instance, widely admired figures … had been in the habit of engaging in consensual homosexual acts, and successfully disseminated that demonstration, you might hope that hostility to homosexuality would gradually abate' (Ibid.).

Thaler and Sunstein's libertarian paternalism or nudging justifies the influence on peoples' choice, if people are provided with free choice and that influence make them chose the better options. According to them, for example:

> When social influences have caused people to have false or biased beliefs, then some nudging may help. […] If many people do something or think something, their actions and their thoughts convey information about what might be best for you to do or think. (Thaler and Sunstein 2009, 58)

So, their examples include the 'public forum' of parks and streets. It is interesting to see that Bentham's also included 'the freedom of press and public discussion' in his strategy of indirect legislation. In 'Place and Time', Bentham writes that 'as a means of obviating dissatisfaction, indirect legislation should be preferred to direct: gentle means, to violent: example, instruction, and exhortation should precede or follow, or, if possible, stand in the place of law' (Bentham 2011, 174).

In addition, Bentham also argues that rulers not only follow public opinion but also 'lead' it (Bentham 1983, 36). And in 'A Table of the Springs of Action', Bentham provides a description of 'deontologists' who are expected to lead ordinary citizens, supplying such motives to them that will promote the happiness of society. To be more specific, Bentham expected deontologists, for example, to articulate their opinions and engage others on the same side. According to Bentham:

> What then is business of the Deontologist? In every instance to bring out of their obscurity, out of the neglect in which they have been hitherto in so large a portion been buried, the points of coincidence to the extent of which extra-regarding interest is connected and has by the hands of nature been identified with self-regarding interest (Bentham 1983, 193).

So, these deontologists would show the people the universal interest which corresponds their particular interests.

Acknowledgement

I would like to thank Professor Xiaobo Zhai for organizing a panel of 'On Postema's Two Books on Bentham's Legal Philosophy' in ISUS 2018 Conference at Karlsruhe, which focused on Professor Postema's two forthcoming books of *Bentham and the Common Law Tradition, Second Edition with Postscript* and *Utility, Publicity, and Law: Bentham's Moral and Legal Philosophy*. And I am most grateful to Professor Postema for letting us read the manuscripts of his two excellent books which are forthcoming. I should also add that, as to the roles of Deontologists, I drew on Dr. Kazuya Takashima's arguments in his book written in Japanese.

References

[1] Bentham, Jeremy. 1962. *The Works of Jeremy Bentham*, vol. 3, edited by J. Bowring. New York: Russell and Russell.

[2] Bentham, Jeremy. 1977. *A Comment on the Commentaries and A Fragment on Government*, edited by J. H. Burns, and H. L. A. Hart. Oxford: Clarendon Press.

[3] _____. 1983. *Constitutional Code*, vol. 1, edited by F. Rosen, and J. H. Burns. Oxford: Clarendon Press.

[4] _____. 1983. *Deontology together with A Table of the Springs of Action and Article on Utilitarianism*, edited by A. Goldworth. Oxford: Clarendon Press.

[5] _____. 1990. *Securities against Misrule and Other Constitutional Writings for Tripoli and Greece*, edited by P. Schofield. Oxford: Clarendon Press.

[6] _____. 2010. *Of the Limits of the Penal Branch of Jurisprudence*. edited by P. Schofield. Oxford: Clarendon Press.

[7] _____. 2011. *Selected Writings: Jeremy Bentham*, edited by S. Engelmann. New Haven: Yale University Press.

[8] Dicey, Albert Venn. 1982. *Introduction to the Study of the Law of the Constitution*, 8th edition. Indianapolis: Liberty Fund.

[9] Lieberman, David. forthcoming. "Declaring Rights: Bentham and the Rights of Man." In *Natural Law and Politics*, edited by R. Whatmore. Cambridge: Cambridge University Press.

[10] Postema, Gerald. forthcoming. *Utility, Publicity, and Law: Essays on Bentham's Moral and Legal Philosophy*. Oxford: Oxford University Press.

[11] Quinn, Michael. 2014. "Popular Prejudices, Real Pains: What is the legislator to do when the people err in assigning mischief." In *Bentham's Theory of Law and Public Opinion*, edited by X. Zhai, and M. Quinn. Cambridge: Cambridge University Press.

[12] Schofield, Philip. 2014. "A Defence of Jeremy Bentham's Critique of Natural Rights." In *Bentham's Theory of Law and Public Opinion*, edited by X. Zhai, and M. Quinn. Cambridge: Cambridge University Press.

[13] Takashima, Kazuya. 2016. *Bentham on Language: Utilitarianism and Pragmatism in the Thought of Jeremy Bentham* [in Japanese]. Tokyo: Keio University Press.

[14] Thaler, Richard and Cass Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth and Happiness*. London: Penguin.

Utilitarianism and the English Poor Law Reform

Emily Lanman, Notre Dame University Australia, Australia

Abstract

The Industrial Revolution transformed all aspects of society in England and Wales throughout the first half of the nineteenth-century, with one major aspect being mass migration to the new industrial centres. With this mass migration and increased industrialisation came unprecedented levels of poverty which the systems in place were not equipped to handle. This resulted in the system of poor relief, which had stood relatively unchanged since 1601, needing to adapt to the needs of the changing society through the Poor Law Reform of 1832-1837, which was heavily influenced by Bentham's Utilitarianism. The topic of this paper addresses an essential period in the history of welfare in England and Wales where a longstanding system of poor relief was radically transformed through the creation of a national system of poor relief for the first time. Despite an expansive body of literature surrounding poverty in the nineteenth-century, there is a specific gap surrounding the philosophical influences and the extent of their influence over the Poor Law Reform. This is also represented in the literature surrounding Utilitarianism and the reform, as historians generally do not agree to what degree there was influence. This is largely due to the conflation between Bentham and his theory. This paper specifically looks at the influence of Utilitarianism on the 1832 Royal Commission, the report it produced and its passage through parliament to the passing of the 1834 Poor Law Amendment Act, with discussion with the implementation following the enactment. The premise of the argument will be that Utilitarian ideas were central to the reform across all aspects. This study utilises reports, debates, legislation and relevant primary documentation to construct a narrative of the influence of Bentham's Utilitarianism on the English Poor Law Reform in England and Wales between 1832-1837.

The nineteenth century was a period of rapid and unprecedented change for England and across all aspects of society, which were also intensifying poverty in an increasingly disruptive manner leading to the Poor Law Reform in England and Wales beginning in 1832 (Gregg 1965, 46; Checkland and Checkland 1974, 29). The Poor Law Reform was initiated to reform the 1601 Poor Law, later known as the Old Poor Law, and through this created a national system of poor relief for the first time in English history. This newly reformed system utilised the workhouse, an institution that provided a last resort for people who could no longer financially support themselves, as its main form of poor relief. This was not the first time the workhouse had been used as a method of relief, as they had been introduced in the seventeenth century, however, workhouses operated under the Old Poor Law had been smaller, more domestic institutions (Newman 2013b, 123). It should be noted that admittance into the workhouse was done so on a voluntary basis and was meant to differentiate between the deserving and undeserving poor, as those who were not truly destitute and simply just work-shy would not sacrifice their freedom to the workhouse. This, it was

thought would ultimately lead to the restoration of social order (Stokes 2001, 711; Newman 2013a, 360). Central to the discussion regarding the influence of Utilitarianism is the specific aspects of the theory, namely the pleasures and pains as outlined in *An Introduction to Morals and Legislation*, as well as the principle of utility (Bentham and Harrison 1960, 155, 125). From this subsequent investigation it becomes apparent that there was a Utilitarian underpinning throughout the reform. To establish this, the historical context that results in the nineteenth century reform will be covered before looking at how Utilitarianism was able to influence the implementation of the reform.

Poverty was not a new issue to the nineteenth-century, legislation combatting pauperism dates back to 1388 with the passing of the first act to deal with vagrancy, other legislation was passed with similar themes towards the end of the fifteenth century and into the sixteenth (Blomfield et al. 1974, 7, 73; Gilliom 2001, 21). Methods for controlling vagrants were often extreme: while a 1547 vagrancy act proposing enslavement of vagrants was deemed impractical, a 1572 act instigated punishment of whipping and boring a hole through the right ear for the first two offences, with the death penalty for the third. However, the need to collect funds for the relief of the poor was recognised from the 1550s, though legislation to raise taxes was opposed, in favour of weekly parish collections under the supervision of the clergy. This would be further developed from 1572, with a further scheme of compulsory rates implemented by parishes to relieve the poor, the sick and the aged (Guy 1988, 326, 220-1). This would culminate with the highly significant 1601 act, this legislation aimed to relieve the old and infirmed, to train children in trades and provide work for the unemployed, for this each parish was made responsible for its own poor which helped reduce costs and prevent undesirable people moving freely (Beckett 1988, 389; Royle 1987, 172). This would be further solidified by the 1662 Act of Settlement which allowed the landless poor to be expelled back to their parish of birth if it was thought that they could become a burden on the parish in which they were settled (Keynon 1969, 944). Settlement in a parish away from an individual's birthplace became dependent on gaining employment for a year, undertaking an apprenticeship, becoming a ratepayer, and (for women) through marriage (Marshall 1956, 186).

Whilst the Old Poor Law would stand until 1834, there were two key attempts at reforming the system. The first of these was the 1782 Gilbert's Act, this shifted away from traditional notions of poor relief being solely a parish issue by encouraging parishes to unite and build workhouses to relieve vulnerable community members, such as the old, sick and insane (Fowler 2014, 28). However, it must be noted that this amalgamation of parishes was not mandatory, and primarily the pauper's local parish was still the main administrative unit for poor relief (Driver 1993, 43). The act also provided a wage supplement to low-paid workers, this in affect facilitated employers to poorly pay workers, shifting the burden of support

onto the general tax base, ultimately leading to the downfall of the system (Fowler 2014, 28). The second attempt came in 1795 in the form of the Speenhamland system. This aimed to provide relief to the unemployed and to supplement employed labourer's wages when the price of bread exceeded a shilling (Watson 1960, 527; Arnstein 1971, 46-7). However, of the two attempts at reform discussed the Speenhamland System can be argued to be the least successful because, whilst at the surface level it seems beneficial, it removed any incentive from landlords and employers to increase wages as they knew that their workers would be entitled to a supplemented wage, and thus lead to the demoralisation of labourers (Arnstein 1971, 6, 47; Hobsbawm 2013, 202).

To understand why by the 1830s there was an intense desire to reform the Old Poor Law, the society that existed at the beginning of the nineteenth-century needs to be explored. One of the most commonly cited reasons for the reform was the increase in costs, from two million pounds in 1784 to nearly six million pounds in 1815 (Beckett 1988, 390; Dyson 2013, 422). This saw a massive shift in the social structure, which is highlighted by the statistics on agricultural employment which show in 1801, thirty-six percent of the population were employed in agriculture which would drop to twenty-two percent by 1851 (Rapport 2005, 83). However, other factors that led to the demise of the Old Poor Law were the demoralisation that came as a result of lower wages, the burden placed on ratepayers, and the higher birth rates which are attributed to the law (Blaug 1974, 123). By 1800 there was a fear of poverty developing amongst the more affluent classes because of social conditions, which in turn led to depletion in charitable outputs (Beckett 1988, 389). This contributed to the years between 1813 and 1837 being described as the blackest years of English farming (Blaug 1974, 123).

To understand how Utilitarianism influenced the Poor Law Reform as a whole between 1832- 1837, it is important to understand the commission. The 1832 Royal Commission into the Poor Laws was the first of its kind, and its success would lead to the model's future utilisation (Finer 1970, 39, 42; Derry 1992, 212). It has been argued that the format itself embodies a Benthamite philosophy of identifying a problem and directing an expert committee to advise on its resolution (Arnstein 1971, 44; Finer 1970, 39; Derry 1992, 212). The commission was appointed to examine the implementation of the Poor Laws whilst the government was preoccupied with the Reform Bill, which aimed to increase political representation (Checkland and Checkland 1974, 29; Dunkley 1981, 124; Royle and Walvin 1982, 158). Alongside this examination of the laws, the commissioners were also instructed to suggest their recommendations for their amendment of the law (Llewellyn 1972, 100). This came to the attention of parliament as the majority of people were dissatisfied with the implementation of the Poor Laws, particularly the landowners who thought their financial obligations

to the poor rates were too high (Finer 1970, 42; Derry 1992, 212). In turn, this led to increased pressure for reform, but ideologically Utilitarianism drove the reform (Royle 1987, 191). In terms of the problem with the Poor Law, the commission and parliament had opposing views: the commission saw the moral and social degradation as being significant alongside the administrative and financial issues, whereas the government only saw the political advantages of its reform (Bowley 2003, 284). The commission was made up of nine individuals, including three clerical representatives presided over by the Bishop of London, Charles Bloomfield, and arguably the most notable members of the commission were Nassau Senior and Edwin Chadwick (Checkland and Checkland 1974, 29). Both Senior and Chadwick can be shown to have been influenced by Utilitarianism, with Chadwick studying under Bentham for an extended amount of time – thus the involvement of these two individuals ensured the prominence of the theory in the commission and the report (Finer 1970, 35; Brundage 1988, 20; Finlayson 1969, 72). Bentham also helped shape the ideology of Nassau Senior who also ideologically shaped the commission (Royle 1987, 191).

The report produced from this commission embodies the Utilitarian ideology they embraced. Utilitarian manifests itself in this document primarily through the promotion of the greatest good for the greatest number. This penetrates the core of the report with the belief that anyone should have access to relief stating, “To refuse relief … is repugnant to the common sentiments of mankind” (Blomfield et al 1974, 334). Through this the principle of utility is promoted through the maximisation of happiness for both the individual and the wider community (Bentham and Harrison 1960, 127). This would continue through the outline for the depauperizing of the able-bodied in the wider community as it would elevate the general condition of the mass of the society. However, limiting conditions would be placed on relief given it was “the country at large, at whose expense he is to be relieved” (Blomfield et al 1974, 335, 337, 375). This would protect the financial interests of the ratepayers in the wider community, promoting the Bentham’s pleasure of wealth. Through these principles, the promotion of the overall happiness of the community would occur, thus demonstrating the principles of Utilitarianism (Bentham and Harrison 1960, 127).

Before examining the individual cases of Hansard for evidence of Utilitarian ideas, a general understanding of the Hansard must be gained. The debates necessary for discussion range from the 21st February to the 13th August 1834 to reflect the reform’s passage through parliament to the passing of the bill. The influence of Utilitarianism is predominantly found in the discussion of the implementation, rather than the machinery of the Act itself. The Hansard has been examined for evidence of direct and indirect references to Utilitarian thought as Bentham said, “A man may be said to be a partizan of the principle of utility, when the approbation or disapprobation he annexes to any action … is determined by and propor-

tioned to the tendency … to have to augment or to diminish the happiness of the community." (Bentham and Harrison 1960, 127) This means that whilst an action may not be created specifically with Utilitarian beliefs in consideration, it is possible for the theory to be represented in the values by which they are guided. Utilitarianism was also able to permeate through parliament through Joseph Hume and John Arthur Roebuck who were known subscribers to the philosophy (Angas Weaver 1987, 1).

A common issue routinely raised in the Hansard of 1834 was the separation of the family unit, as it was proposed that inmates would be segregated by gender, and in some cases by age[1]. This was regularly called into question by those opposed to the bill as being unnecessarily cruel however, this was refuted through statements by the likes of Lord Althorp who stated that separation was necessary "to ensure the proper regulation of workhouses".[2] This is an idea that is steeped in Utilitarian influence as it draws on the pleasure versus pain principle by using the separation of the family unit as a deterrence to paupers claiming relief from the parish (Bentham and Harrison 1960, 155).

In the discussions surrounding the mechanics of the bill, food and luxury products namely beer and tobacco were discussed in terms of their denial.[3] It was generally agreed that workhouse inmates "… should not be so well fed."[4] This relates to the "pleasure of taste or palate; including whatever experienced in satisfying the appetites of hunger and thirst", which would be denied in the workhouse as bland food was to be provided in the workhouse as a disincentive to staying (Bentham and Harrison 1960, 156; Miller 2013, 945-6). The denial of a paupers "… accustomed enjoyments – no beer, no tobacco …" was raised, which further demonstrates the Utilitarian influence in its denial of the pauper the "pleasure of intoxication", which is one of Bentham's pleasures of sense.[5] This demonstrates how the Utilitarian pleasures infiltrate the debates in parliament.

The topic of bastardy was discussed with reference to whom should have to predominately support the child.[6] Whilst it was agreed that the mother should retain partial responsibility of the child, it was also maintained that the father should also be held, "responsible to the parish for the maintenance of his illegitimate child for otherwise the changes on the parishes in large manufacturing towns and districts would be much increased"[7]. This promotes

[1] HC Parliamentary Debates, 17 April 1834, vol.22, c. 896.
[2] HC Parliamentary Debates, 9 June 1834, c.24, c.338.
[3] HC Parliamentary Debates, 1 July 1834, vol.24, c.1035.
[4] HC Parliamentary Debates, 23 May 1834, vol.23, c.1304.
[5] HC Parliamentary Debates, 1 July 1834, vol.24, c.1035; Bentham and Harrison 1960, 156.
[6] HC Parliamentary Debates, 18 June 1834, c.525, 527, 535.
[7] HC Parliamentary Debates, 18 June 1834, c.525, 527, 535.

the principle of utility through the "pleasure of wealth", as by making the parents of the child responsible for its upkeep, instead of the of the parish, the financial interests of the wider community, and specifically the rate-payers would be protected (Bentham and Harrison 1960, 127, 156).

Given the prominence of religion in society, it is not surprising that it was a major topic of debate throughout the Hansard. It was proposed that the commissioners should not be permitted to, "… oblige the inmates of a workhouse to attend any religious service that they did not conscientiously believe in, or to oblige the children in a workhouse to be educated in any faith that their parents did not approve of …".[8] This principle was further discussed in the sitting at the end of June, where it was proposed that children whose parents perish within the workhouse would not be educated in a faith that their parents did not agree with[9]. This positively represents the "pleasure of piety" as presented by Bentham's Utilitarianism which stated, "… the belief of a man's being in the acquisition or in the possession of the good-will or favour of the Supreme Being …" is a pleasure which people will seek to maximise (Bentham and Harrison 1960, 157). Concerning the religious education, Lord Althorp once again reiterates the "… greatest importance that persons of every religious denomination should have religious instruction from their own pastors …", even if they were not ministers of the established church.[10] This ensures that paupers are still entitled to the "pleasure of piety" as they have the opportunity to acquire …" the good-will of favour of the Supreme Being …" (Bentham and Harrison 1960, 157).

The 1834 Poor Law Amendment Act was produced through the contributions of the Royal Commission and parliamentary debate over its report and drafting of legislation, thus the final form of the Act is also evidence of Utilitarianism's influence on the Poor Law Reform. Historians regard the Act as radical piece of legislation that highlights the triumph of the newly emerging liberalism in parliament; this is also considered to be significant as it coincides with the rise of middle class as an influential entity (Dentith 2009, 79; Salvadori 1972, 2; Edsall 1971, 1). Despite its radical nature, the Act is vague in content: it does not make an explicit plan for reform through its one hundred and ten sections, but rather provides guidelines predominately focusing on the operations of the commissioners and guardians (Midwinter 1969, 7; Public General Act 1834, s.1-18). Once again, similar traits appear in the Act as seen in the Hansard, including the role of religion the upkeep of illegitimate children, outdoor relief and the prohibition of alcohol within the institution (Public General Act 1834, s.57, 27, 91). These guidelines structured the daily operations of the workhouses and

[8] HC Parliamentary Debates, 21 June 1834, c.719.

[9] HC Parliamentary Debates, 27 June 1834, c.926.

[10] HC Parliamentary Debates, 11 August 1834, c.1225.

relief of the poor being left in the hands of local guardians (Public General Act 1834, s.38). In a circular sent to parishes in November of 1834 by Edwin Chadwick, the aim of the new law was laid down clearly for the overseers, stating that it was not passed, "… for the purpose of abolishing the necessary relief to the indigent, but for preventing various illegal and injurious practises, which had by degrees grown up in the administration of such relief" (Chadwick 1834b, n.p). The first eighteen sections of the Act relate specifically the commissioners and aspects of their operations, including their appointment, who is eligible to sit and so forth (Public General Act 1834, s.1-18).

Following the enactment, Utilitarian influence can be found in the implementation in the workhouse. Food, it has been argued, was seen by the central authority as an integral part of workhouse discipline (Crowther 1983, 213-4). Food was doled out in the workhouse from six predetermined diets that local commissions could pick from; these were graded according to the age, sex and status of the inmate, the able-bodied receiving only the plainest fair (Roberts 1963, 103; Chadwick 1835, n.p.; Crowther 1983, 214). Bentham supported determined diets stating: "The dietary should not be fixed to a single mess: but a list of messes …", further stating that the diets should "not be left to the local authority" (Bentham and Quinn 2010, 140). These diets all consisted of bread and gruel for breakfast, and bread and cheese for supper, the main meal of the day differs between the six, but being made up of meat, bread, cheese, soup or meat pudding (Chadwick 1835, n.p). This demonstrates Utilitarian principles as it removes, "The pleasures of taste or palate; outlining whatever pleasures are experienced in satisfying the appetites of hunger and thirst" through the monotonous fair they would be given (Bentham and Harrison 1960, 160). For the majority of paupers, the distaste concerning the food came from the lack of luxuries such as beer, rather than a lack of nourishment; in some cases, access to salt was denied, further stripping their diet of everything that was familiar or acceptable. Once again removing the "pleasure of taste" from the inmates (Crowther 1983, 218; Bentham and Harrison 1960, 156).

The evidence of Utilitarian influence also manifests within the principles surrounding outdoor relief, which the Royal Commission report had sought to prevent, but which nevertheless was allowed by the Act[11]. This provision of outdoor relief promoted the greatest good as it allows the Board of Guardians to still provide outdoor relief at their discretion, thus promoting the principle of utility as it allows for the Guardians to act in a way that promotes the greater interest of the community (Public General Act 1834, s.23; Bentham and Harrison 1960, 126-7). It also highlights how the pauper was able to retain some dignity through the "pleasure of a good name" as they would not have to submit themselves to the workhouse

[11] HC Parliamentary Debates, April 17 1834, 883, 889; Public General Act 1834, s.23.

and thus enter a lower social ranking (Bentham and Harrison 1960, 157; Newman 2013a, 366). This notion relates to the idea permeating through wider society of the deserving and undeserving poor, as only the deserving poor were said to be willing to give up their freedom to the workhouse (Stokes 2001, 711; Newman 2013a, 366). The idea of the deserving and undeserving poor is particularly relevant, as by submitting themselves to the workhouse, a pauper would reduce himself to a lower social standing (Newman 2013a, 365). The evidence of Utilitarianism also presents itself in the circular sent to the overseers of the poor in November of 1834, to advise how to make the transition to the new system as smooth as possible. For example, point 4 discusses the financial allowances allotted to labourers with a number of children, stating, "… you should not suddenly or altogether discontinue these allowances, but you should make them in kind, rather than in money" (Chadwick 1834b, n.p). This endorses the principle of utility as it promotes the pleasure of the individuals in the community; specifically, this was done through the "pleasure of wealth", as it does not strip the paupers of their allowances immediately but rather weans them off the parish fund (Bentham and Harrison 1960, 127, 156). This then results in the preservation of the "pleasure of a good name", as it means they are not forced to enter the workhouse immediately to obtain their relief (Bentham and Harrison 1960, 157). So, it is evident through the principles of outdoor relief under the 1834 Poor Law Amendment Act that Utilitarianism manifested itself within the operation of the New Poor Law, as well as through the workhouse system.

Religion was a significant aspect of life within the workhouse, such as regular prayers, including before meals (Poor Law Commissioners 1836b, n.p). However, in the Utilitarian spirit no inmate would be forced to conform to, or practise a religion they did not believe in, as highlighted in the 1834 Poor Law Amendment Act and circulars sent out to parishes after enactment (Public General Act 1834, 19; Poor Law Commissioners 1836b; 1836a, n.p). Bentham discusses how "being in the acquisition or in the possession of the good-will or favour of the Supreme Being" under the pleasures of piety (Bentham and Harrison 1960, 157). This notion is further expanded to the religious instruction of pauper children in the event of the death of the parents, because pauper children would continue to be educated and allowed to practise the faith of their parents, even in the event of their demise (Public General Act 1834, s.19). Once again this relates back to the "pleasure of piety" as it allows for the inmate, and their children, to gain a knowledge and good will of their Supreme Being (Bentham and Harrison 1960, 157).

So, it can be determined that Utilitarianism was a driving force behind the English Poor Law Reform of 1832-1837. This can be observed through the prominence of Utilitarian thinkers in the Royal Commission and the subsequent report that was produced. These ideas then

continue through the parliamentary debates, culminating with the 1834 Poor Law Amendment Act. Through this Act, the Utilitarian principles can be seen through the workhouse system it reformed. This is shown through the system that Utilitarianism facilitated the creation of the materially dealt with the structural poverty created by the industrial revolution, while providing an ideological narrative that justified blaming paupers for their own poverty, and the regulating of their behaviour to minimise social disruption (Finlayson 1969, 72; Derry 1992, 212; Newman 2013b, 123). The influence of Utilitarianism can be established through the analysis of the primary documents relating to the reform, namely reports produced for the British government, Hansard, legislation and archival material pertaining to the operations of the workhouse and the 1834 Poor Law Amendment Act.

References

[1] Angas Weaver, Stewart. 1987. *John Fielden and the Politics of Popular Radicalism*, *1832-1847*. Oxford: Oxford University Press.

[2] Arnstein, Walter L. 1971. *Britain Yesterday and Today: 1830 to Present*. Lexington, MA: D.C. Heath and Company.

[3] Beckett, John V. 1988*. The Aristocracy in England 1660-1914*. Oxford: Basil Blackwell.

[4] Bentham, Jeremy and Wilfrid Harrison. 1960. "An Introduction to the Principles of Morals and Legislation." In *A Fragment on Government and an Introduction to the Principles of Morals Legislation*, 113–410. Oxford: Basil Blackwell.

[5] Bentham, Jeremy, and Michael Quinn. 2010. "Pauper Management Improved." In *Writings on the Poor Laws*, Vol. II, 1–478. Oxford: Clarendon Press.

[6] Blaug, Mark. 1974. "The Myth of the Old Poor Law and the Making of the New." In *Essays in Social History*, edited by Michael W. Flinn, and Thomas C. Smout, 123–54. Oxford: Clarendon Press.

[7] Blomfield, Charles J., John B. Summer, William Sturges Bourne, Nassau W. Senior, Henry Bishop, Henry Gawler, Walter Coulson, James Traill, and Edwin Chadwick. 1974. "Reports from Commissioners: Twenty-Two Volumes." In *The Poor Law Report of 1834*, edited by Sydney G. Checkland, and Edith O. A. Checkland, 67–499. Harmondsworth: Penguin Books.

[8] Bowley, Marian. 2003. *Nassau Senior and Classical Economics*. Milton Park: Routledge.

[9] Brundage, Anthony. 1988. *England's "Prussian Minister": Edwin Chadwick and the Politics of Government Growth, 1832-1854.* University Park, PA: The Pennsylvania State University Press.

[10] Chadwick, Edwin. 1834a. *To the Churchwardens and Overseers and the Other Officers Charged with the Relief of the Poor*. Poor Law Commission: Minutes Books, Vol.1. The National Archives, Kew, England. MH 1/1.

[11] _____. 1834b. *To the Overseers of the Poor*. Poor Law Commission: Minutes Books. Vol.1. The National Archives, Kew, England. MH 1/1.

[12] _____. 1835. *Poor Law Commission Office. To Union and Parish Officers.* The National Archives, Kew, England. MH 10/7.

[13] Checkland, Sydney G., and Edith O. A. Checkland. 1974. *The Poor Law Report of 1834.* Harmondsworth: Penguin Books.

[14] Crowther, Margaret A. 1983. *The Workhouse System 1834-1929: The History of an English Social Institution*. London: Methuen & Co.

[15] Dentith, Simon. 2009. "'The Shadows of the Workhouse': The Afterlife of a Victorian Institution." *Literature Interpretation Theory* 20 (1-2): 79–91. DOI: 10.1080/10436920802690448.

[16] Derry, John W. 1992. *Charles, Earl Grey: Aristocratic Reformer*. Oxford: Blackwell Publishers.

[17] Driver, Felix. 1993. *Power and Pauperism: The Workhouse System, 1834-1884.* Cambridge: Press Syndicate of the University of Cambridge.

[18] Dunkley, Peter. 1981. "Whigs and Paupers: The Reform of the English Poor Laws, 1830-1834." *The Journal of British Studies* 20 (2): 124–49. DOI: 10.1086/385776.

[19] Dyson, Richard. 2013. "The Extent and Nature of Pauperism in Five Oxfordshire Parishes, 1786–1832." *Continuity and Change* 28 (3): 421–49. DOI: 10.1017/s0268416013000374.

[20] Edsall, Nicholas C. 1971. *The Anti-Poor Law Movement 1834-44.* Manchester: Manchester University Press.

[21] Finer, Samuel E. 1970. *The Life and Times of Sir Edwin Chadwick*. London: Methuen & Co.

[22] Finlayson, Geoffrey B. A. M. 1969. *England in the Eighteen Thirties: Decade of Reform*. London: Edward Arnold.

[23] Fowler, Simon. 2014. *Workhouse: The People, the Places, the Life behind Doors*. Barnsley: Pen & Sword Books.

[24] Gilliom, John. 2001. *Overseers of the Poor: Surveillance, Resistance, and the Limits of Privacy*. London: The University of Chicago Press.

[25] Gregg, Pauline. 1965*. A Social and Economic History of Britain: 1760-1965*. London: George G. Harrap & Co.

[26] Guy, John. 1988. *Tudor England*. Oxford: Oxford University Press.

[27] Hobsbawm, Eric J. 2013. *The Age of Revolution 1789-1848*. London: Abacus.

[28] Kenyon, John P. 1969. *The Stuart Constitution: Documents and Commentary*. London: Syndics of the Cambridge University Press.

[29] Llewellyn, Alexander. 1972. *The Decade of Reform: the 1830s*. Newton Abbot: David & Charles.

[30] Marshall, Dorothy. 1956. *English People in the Eighteenth Century*. London: Longmans, Green and Co.

[31] Midwinter, Eric C. 1969. *Social Administration in Lancashire 1830-1860*. Manchester: Manchester University Press.

[32] Miller, Ian. 2013. "Feeding in the Workhouse: The Institutional and Ideological Functions of Food in Britain, Ca. 1834–70." *Journal of British Studies* 52 (4): 940–62. DOI: 10.1017/jbr.2013.176.

[33] Newman, Charlotte. 2013a. "An Archaeology of Poverty: Architectural Innovation and Pauper Experience at Madeley Union Workhouse, Shropshire." *Post-Medieval Archaeology* 47 (2): 359–77. DOI: 10.1179/0079423613z.00000000046.

[34] _____. 2013b. "To Punish or Protect: The New Poor Law and the English Workhouse." *International Journal of Historical Archaeology* 18 (1): 122–45. DOI: 10.1007/s10761-013-0249-7.

[35] Parry, Jonathan. 1993. *The Rise and Fall of Liberal Government in Victorian Britain*. New Haven: Yale University Press.

[36] Poor Law Commissioners. 1836a. *In Pursuance of an Act of parliament. To Assistant Poor Law Commissioners*. The National Archives, Kew, England. MH 10/2.

[37] _____. 1836b. *Union. To Assistant Poor Law Commissioners*. The National Archives, Kew, England. MH 10/2.

[38] Public General Act, 4&5 William IV, c.76 1834 (England and Wales).

[39] Rapport, Michael. 2005. *Nineteenth-Century Europe*. Houndsmill: Palgrave Macmillan.

[40] Roberts, David. 1963. "How Cruel Was the Victorian Poor Law?" *The Historical Journal* 6 (1): 97. DOI: 10.1017/s0018246x00000935.

[41] Royle, Edward, and James Walvin. 1982. *English Radicals and Reformers 1760-1848*. Brighton: The Harvester Press.

[42] Royle, Edward. 1987. *Modern Britain: A Social History 1750-1985*. London: Edward Arnold.

[43] Salvadori, Massimo. 1972. *European Liberalism*. New York: John Wiley and Sons.

[44] Stokes, Peter M. 2001. "Bentham, Dickens, and the Uses of the Workhouse." *Studies in English Literature 1500-1900* 41 (4): 711–27. DOI: 10.1353/sel.2001.0042.

[45] United Kingdom. *Poor Laws' Amendment-Committee*, 9 June 1834. Parliamentary Debates, Commons, vol.24, cc324-40.

[46] _____. *Poor Laws' England*, 17 April 1834. Parliamentary Debates, Commons, vol.22, cc874-98.

[47] _____. *Poor Laws' Amendment Committee*, 23 May 1834. Parliamentary Debates, Commons, vol.23, cc1276-304.

[48] _____. *Poor Laws' Amendment*, 11 August 1834. Parliamentary Debates, Commons, vol.25, cc1207-28.

[49] _____. *Poor Laws' Amendment-Report*, 27 June 1834. Parliamentary Debates, Commons, vol.24, cc913-35.

[50] _____. *Poor Laws' Amendment Committee*, 18 June 1834. Parliamentary Debates, Commons, vol.24, cc.520-49.

[51] _____. *Poor Laws' Amendment Committee*, 21 June 1834. Parliamentary Debates, Commons, vol.24, cc.715-9.

[52] _____. *Poor Laws' Amendment- Third Reading*, 1 July 1834. Parliamentary Debates, Commons, vol.24, cc.1027-61.

[53] Watson, J S. 1960. *The Reign of George III 1760-1815*. London: Oxford University Press.

Samuel Romilly and Jeremy Bentham's Decisions of Publication

Cheng Li, University of York, UK

Abstract

From 1788 to 1818, Samuel Romilly had been Jeremy Bentham's closest friend in his intellectual life. Due to his admiration of Romilly's reformism and rising reputation in the courts and parliament, every time Bentham wrote a new manuscript he would send it to Romilly for advice. Romilly's advice directly influenced Bentham's decision on whether or not to distribute or publish his increasingly radical writings. In Romilly's revisions, he would mark the dangerous passages of Bentham's manuscripts and replace them with safer expressions. For the most time, Bentham appreciated much of Romilly's revisions and accepted his suggestions. The following works of Bentham will be discussed: Truth versus Ashhurst; or, law as it is, contrasted with what it is said to be, the 'On the dispensing power exercised by the Duke of Portland and his confederates', the Elements of the Art of Packing and the Church-of-Englandism. The analysis reveals that Romilly's persuasion had two consequences for Bentham. Firstly, it stimulated Bentham's thinking about the freedom of speech and the nature of the existing libel laws. Secondly, it improved Bentham's judgement of publication.

I

The intellectual life at the salons of the Marquess of Lansdowne largely promoted the formation of Bentham's friendship with Romilly. By 1788 when Bentham had failed to attract the interest of the Russian empress and returned to England, Lansdowne's Parisian salons had attracted many British, French and American intellectuals and became a centre of innovative ideas (Andrew 2006, 170). On the eve of the French Revolution, both Romilly and Bentham were excited by an optimistic view that the Revolution would improve both French and British societies. Lord Lansdowne encouraged them to aid his French connections with their knowledge of British law. Through Lansdowne's arrangement, Romilly travelled to France and was asked by a French military officer for "some book which stated the rules and orders of proceeding in the English House of Commons" for the Estates-General (Romilly 1840 I, 101). This inquiry stimulated Romilly's interest in writing a manual by himself after a failed search for a suitable one. When Romilly returned to England, and shared the news with Bentham on Lansdowne's salons, Bentham began to write on a similar topic. Also, Romilly assisted Bentham in his French writing and passed Bentham's manuscripts on to Etienne Dumont, whose translation later built an international reputation for Bentham. Ro-

milly and Bentham enjoyed the intellectual life at Lansdowne's circle, particularly in his library which provided a plentiful source for them (de Champs 2015, 104). This association with Lansdowne is important also for the reason that through Lansdowne's transnational network, Romilly and Bentham for the first time were presented to a large Enlightened audience with the hope that their ideas could be heard and appreciated. When they worked closely together on writing on French issues and other reformist projects, they found more mutual appreciation.

Secondly, Romilly's rising reputation in the courts and parliament made Bentham seek assistance for marketing his projects. Though the nineteenth century saw that social entrepreneurs or "projectors" became more influential in some policy-making, the process of marketing or persuading the government still relied on personal friendships, and political patronage in many ways dominated the career of a social projector. In order to advertise the advantages of his projects, Bentham would seek friends within the system. In two of Bentham's projects, the Panopticon prison building and the Scottish court reform, Romilly provided continuous support and became one of Bentham's comrades. Romilly was named a king's counsel in 1800 and in 1802 he was considered as "the head of the profession both in point of legal accomplishments, general information, and respectability" (Horner 1843, 182). Due to Romilly's high reputation and his practice in the chancery court, when Bentham met difficulty in persuading the Lord Chancellor he often asked Romilly for support. Through Romilly's nudging, the Lord Chancellor responded to Bentham, saying why he disapproved of the Panopticon Bill. In 1806, Romilly was appointed as the Solicitor-General and this promotion encouraged Bentham's hope that his voice would be more appreciated. One ambition of the Whig ministry in 1806-7 was the reform of the Scottish civil courts. Some young Scottish lawyers came to London and visited Romilly's house, where they were supposed to learn Bentham's new writing on the same topic through Romilly's introduction (Dinwiddy 1988, 416). Through Romilly's connection, Bentham received a public invitation from the influential *Edinburgh Review* to give advice on the Scottish legal reform (Jeffrey 1807, 483). In 1808, at Romilly's encouragement, Bentham published the *Scotch Reform; Considered with reference to the plan, proposed in the late Parliament, for the regulation of the courts, and the administration of justice in Scotland* which guided the later reforms.

II

Due to Romilly's position and legal knowledge, it would be safer for Bentham to know Romilly's opinion before the distribution or publication of his ideas. Many of Bentham's radical expressions could cause the risk of prosecution. During the period of French Revolution and

the wars against France, many innovative and democratic ideas were interpreted as a threat to the national security. Without a precise code, judges were encouraged to use their discretionary power, but the opinion of a judge was uncertain. The political situations in that period also increased the difficulty in predicting a judge's decision. Though judges tended to use the law more as a threatening tool and were not keen on the actual enforcement (Harling 2001, 107-34), the harshness of punishments and costly court fees still produced much anxiety. In practice, corruption was rife before the trial. The jury trial had long been boasted of as the safeguard of English justice but in 1817, a national scandal happened when corrupt practice in the special jury selection in Crown prosecutions was exposed and led to unprecedented publicity. Newspapers revealed how the master of the Crown Office selected jurymen in favour of the government (Epstein 1994, 56-57).

Bentham's critiques of the practice and theory of the existing laws developed from the 1790s. December 1792 was a sensitive moment when the memory of the French September Massacres was so fresh, and the government had just conducted a trial of Thomas Paine for seditious libel. Bentham wrote a sweeping pamphlet, *Truth versus Ashurst; or, law as it is, contrasted with what it is said to be*. His words were most intense where the judge Ashurst boasted of the superiorities of the English law: Bentham refuted these as the abuses. Bentham also interpreted Ashurst's theory as the "dog-law" thinking in which judges designed the law as a tool to rule the people of inferior social status. The relationship between law makers and receivers was like the dog master and his dog. Judges deliberately made the laws vague in order to trap people for court fees: "when your dog does anything you want to break him of, you wait till he does it, and then beat him for it. This is the way you make laws for your dog" (Bowring 1843 V, 235).

Soon after finishing the *Truth versus Ashurst*, Bentham sent a copy to Romilly. Romilly accepted its arguments and planned to make extracts. However, his opinion was not to publish as "the praise given to the French would, I have no doubt, throw discredit on all the truth it contains" (Milne 2017, 415). Bentham then gave up its publication until 1823. Interestingly, this decision might have reflected their change of attitude towards the French Revolution. After the September Massacres, Romilly said, "how could we ever be so deceived in the character of the French nation as to think them capable of liberty" (Romilly 1840 II, 4). He also burned the copies of his new pamphlet as it contained his optimistic hopes. As their common friend Dumont said, "let us burn all our books, let us cease to think and dream of the best system of legislation, since men make so diabolical a use of every truth and every principle" (Romilly 1840 II, 6). Many French friends whom Romilly and Bentham knew through Lansdowne died in the Massacre. The death of the duc de La Rochefoucauld on 4 September was devastating to Bentham, who was to have dinner with Dumont and the duke's first cousin when the news arrived (de Champs 2015, 100). Since this event until the

mid-1795, Bentham hardly supported any political reformist projects and even associated any democratic polity with a series of negative features such as ignorance, violence, extravagance, discontent, frequent wars, and danger of violent revolution to any democratic polity (Schofield 2004, 392, 396).

While after the Massacres, Bentham found much to say about the merits of the British constitution he still highly criticised its legal system (Schofield 2004, 398; 2006, 114). At the end of March 1802, Bentham wrote a polemic against the Duke of Portland, the then Lord President of the Council and the former Home Secretary. The polemic was in the form of a letter which Bentham planned to send to Lord Pelham, the Home Secretary. Bentham declared the former Secretary and his assistants of exercising "a dispensing power, for the purpose of illegally obstructing, and if possible preventing the execution of an imperative Act of Parliament" (University College London Library, Bentham Papers [hereafter UC], box 120, fol. 470). The Act refers to the 1794 Penitentiary Act which gave approval to build a profit-making prison and to allow Bentham to make a contract with the government (34 Geo. 3 c. 84). For many disadvantageous conditions such as the aristocratic dislike of Bentham's chosen site and the Home Office's preference of the penal transportation policy, by 1802 Bentham had spent 8 years of money and energy to get the land but failed to secure the contract (Semple 1993, 224-5). In an angry mood, Bentham had written a few weeks earlier that "unfortunately as to the destruction of eight years … they have murdered my best days!" (UC, box 120, fol. 466). When Romilly read the polemic, he agreed with Bentham's argument but discouraged the publication for the reason that Bentham's violent expressions might be conceived as libel (Dinwiddy 1988, 154). Bentham used words such as "conspiracy" and "state crimes" (UC, box 120, fol. 470, 473). Also, Romilly reminded Bentham that since the 1794 Act concerned his personal interest, the publication might cause a public scandal which would injure his own reputation. Thirdly, as a legal expert, Romilly suggested that such a scandal would force the government to accuse Bentham the guilty of libel. Romilly further inquired Bentham "what has passed between you and the present ministers" and reminded Bentham that there was very little chance of persuading the new Home Secretary (Dinwiddy 1988, 155).

Bentham was alarmed by Romilly. Though still angry with Portland, he calmed down and began to revise the violent passages. He trusted Romilly's libel law knowledge and asked him to detect "any objectionable passages" and claimed that if Romilly would not do, he would risk publishing the polemic and took the chance "for seeing the inside of the King's Bench" (Dinwiddy 1988, 155). While the tone was tough, Bentham avoided danger.

The radicalization of Bentham's legal thought continued, although it remained unpublished until 1817. On 20 February 1809, *The Times* commented on the government's actions relating to the Duke of York Scandal. It claimed that 26 printers and publishers were under prosecution for libelling the Duke, and the public mind was under "no ordinary uneasiness". Some of these printers and publishers were close to Bentham's circle. For example, in April Bentham's friends James Mill, Francis Horner and Francis Burdett concerned and petitioned for the *Independent Whig* newspaper, whose publisher and printer were convicted of libel and imprisoned (Conway 1988, 26-7). Their parliamentary petition complained that the king's bench conducted an unfair trial and the special jury was selected in an irregular way which damaged the constitutional right of the accused (*Hansard* 24 Apr. 1809, 175-7). In the same month, Bentham was consulting sources for the *Elements of the Art of Packing, as applied to Special Juries, particularly in Cases of Libel Law* (UC, box 26, fol. 68; Conway 1988, 22). By October, the sheriff of London, who was responsible for the selection of the special jury, seemed supportive of Bentham's inquiry (Conway 1988, 47-9). At the end of 1809, Bentham completed a draft and judged that a suitable time for its publication was in the near future when he wrote "the current of public opinion has been turned against the Ministry, or rather against all Ministries, and in favour of Parliamentary Reform as the only remedy" (Conway 1988, 60).

In this context, Bentham sent a copy to Romilly. On 31 January 1810, Romilly replied that "I do most sincerely and anxiously entreat you not to publish it, --and I have not the least doubt that Gibbs [the then Attorney-General] would prosecute both the author and the printer" (Conway 1988, 60). Romilly stressed the point that the current Attorney-General was a very tough adversary and would prosecute and put Bentham in prison straightway. On 9 June 1809, Romilly had a conversation with Gibbs in the House of Commons. On that day, Gibbs had planned to move a Bill to strengthen the government's power to suppress seditious activities, but as other business occupied the House till past 12 o'clock at night, Romilly prevented him for the reason that he would oppose the Bill and it was too late to have another debate. Romilly was concerned that Gibbs' Bill was "a most insidious attack upon the liberties of the people" (Romilly 1840 II, 289-90). On 28 March 1811, Gibbs was criticised in the House of Lords for his harshness and partiality for his enthusiasm with which he filed ex officio information against the publishers. More relevant to Bentham's case, Gibbs was a strong loyalist supporter of the Duke of York and his overreaction in protecting the Duke's fame made these 26 publishers suffer (Melikan 2009, 3). Due to these considerations, especially Romilly's personal observation of Gibb's character, Bentham was persuaded as he repeated Romilly's warning to another friend later (Conway 1988, 94).

III

These experiences pushed Bentham into developing a cautious strategy of publication to avoid prosecution. In September 1817, Bentham had printed a book which criticised how the Church of England conducted a series of abuses by its system of education. Before the decision of publication, Bentham lent Romilly a proof copy for the possibility of prosecution. Romilly cautioned that Bentham's words might be viewed as disrespectful to Jesus (Conway 1989, 66). Bentham then asked Romilly "to mark the dangerous passages" and "set down in the margin what he regarded as safe substitutes" (Conway 1989, 66). However, by 7 January 1818, as Romilly still did not finish the revisions, Bentham lost patience and asked the return so that he could make use of whatever comments in it (Conway 1989, 143).

In fact, earlier in December 1817, in order to test the possibility of prosecution by the current Attorney-General Samuel Shepherd, Bentham had planned to publish a sample of the work which included extracts on the subject of blasphemy in a well-known newspaper. Shepheard was not like Gibbs and had defended Bentham's radical MP friend Burdett in 1810. The newspaper in Bentham's mind was either the *Morning Chronicle* or the *Examiner* (Conway 1989, 138). Bentham particularly admired the editor of the *Morning Chronicle*, John Black who worked closely with James Mill and Francis Place. Meanwhile, William Hone's blasphemy case encouraged enormous publicity and mobilized the public opinion against the government (Marsh 1998, 28). Bentham had hoped to publish the sample before the trial so that his arguments could produce its best effect. Eventually, he managed to publish the sample on 18 January 1818 in the *Examiner* with the promise that it could be revised freely with the editor's discretion (Conway 1989, 139).

Bentham also developed another method to secure safe publication. On 24 January Bentham contacted William Smith, a unitarian MP and friend of the Archbishop of Canterbury, to enquire about the Archbishop's attitude towards the blasphemy laws as Smith worked together with the Archbishop in 1813 in a reform campaign. One of Bentham's questions was "whether it be not true, that a Bill, either drawn or approved by the Archbishop of Canterbury, gave to the liberty of printing and publishing" (Conway 1989, 151-3). In this way, Bentham aimed to get the endorsement of the Archbishop for by adding the MP's account into the preface of his new work. On 16 February, the MP provided the account. Later, Bentham published the *Church-of-Englandism* safely.

In short, Romilly deserves the credit for improving Bentham's judgement about publication. The bitter Panopticon experience not only pushed Bentham into a novel form of radical

politics but also made his rhetoric inappropriately violent sometimes. Bentham was too politically optimistic and misjudged the tide of popular opinion and the stability of the Tory government many times. In this sense, it was fortunate for Bentham to have a critic lawyer friend who could tell him other sides of the real world.

References

[1] Andrew, Edward. 2006. *Patrons of Enlightenment*. Toronto: University of Toronto Press.

[2] Bowring, John, ed. 1843. *The Works of Jeremy Bentham*. *Vol. V.* Edinburgh: Tait.

[3] Conway, Stephen, ed. 1988. *The Correspondence of Jeremy Bentham*. *Vol. 8: January 1809 to December 1816*. Oxford: Clarendon Press.

[4] _____, ed. 1989. *The Correspondence of Jeremy Bentham. Vol. 9: January 1817 to June 1820*. Oxford: Clarendon Press.

[5] De Champs, Emmanuelle. 2015. E*nlightenment and Utility: Bentham in French, Bentham in France.* Cambridge: Cambridge University Press.

[6] Dinwiddy, J. R., ed. 1988. *The Correspondence of Jeremy Bentham. Vol. 7: January 1802 to December 1808*. Oxford: Clarendon Press.

[7] Epstein, James. 1994. *Radical expression: political language, ritual, and symbol in England, 1790-1850*. Oxford: Oxford University Press.

[8] *Hansard* (U.K.) House of Commons Debates, 24 April 1809, vol. 14, 175-7, accessed via Hansard. July 1, 2020. https://api.parliament.uk/historic-hansard/commons/1809/apr/24/ petition-of-mr-henry-white

[9] Harling, Philip. 2001. "The Law of Libel and the Limits of Repression, 1790-1832." *The Historical Journal* 44 (2): 107-34.

[10] Horner, Francis. 1843. *Memoirs and Correspondence of Francis Horner. 2 vols*. London: John Murray.

[11] Jeffrey, Francis. 1807. "Proposed reform of the Court of Session [in Scotland]" *The Edinburgh Review* 9 (January): 462-92.

[12] Marsh, Joss. 1998. *Word crimes: blasphemy, culture, and literature in nineteenth-century England*. Chicago: The University of Chicago Press.

[13] Melikan, R. A. 2009. "Gibbs, Sir Vicary." *Oxford Dictionary of National Biography*: 1-7.

[14] Milne, Alexander Taylor, ed. 2017. *The Correspondence of Jeremy Bentham. Vol. 4: October 1788 to December 1793*. London: UCL Press.

[15] Romilly, Samuel. 1840. *Memoirs of the life of Sir Samuel Romilly. 3 vols*. London: John Murray.

[16] Schofield, Philip. 2004. "Jeremy Bentham, the French Revolution and political radicalism." *History of European Ideas* 30: 381-401.

[17] _____. 2006. *Utility and Democracy: The Political Thought of Jeremy Bentham*. Oxford: Oxford University Press.

[18] Semple, Janet. 1993. *Bentham's Prison: A Study of the Panopticon Penitentiary*. Oxford: Oxford University Press.

[19] University College London Library, Bentham Papers. Transcript of Box 26, fol. 68, "Bentham Papers, UCL Special Collections.", accessed via *Transcribe Bentham*. November 18, 2018. http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham

[20] _____. Transcript of Box 120, fol. 466, 470, 473, "Bentham Papers, UCL Special Collections.", accessed via *Transcribe Bentham*. November 18, 2018. http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham

Killing Animals: The Badness of Death, Value and Replaceability

Fayna Fuentes López, Macquarie University Sydney, Australia

Abstract

Whilst animal suffering is usually considered morally relevant in utilitarian ethics, it is commonly argued that killing an animal can be a morally neutral act, so long as some conditions are met. In particular, Singer has argued that non-persons (both human and non-human) are replaceable, that is, we can painlessly kill them as long as we bring a similar being into existence to compensate for the net loss of utility that the killing would have resulted in the universe. This, known as the replaceability argument, is arguably Singer's most controversial argument. In my paper, I will argue that death can be a misfortune for the victim due to the deprivation it causes. Death leaves the victim comparatively worse off (if the balance of well-being lost is positive), and therefore, death can be a misfortune for the victim. Furthermore, as (sentient) animals can experience different levels of well-being, they are too affected by their deaths. Thus, in cases where the animal is harmed by their death, utilitarians have a direct *pro tanto* reason to oppose their killing. However, replaceability means that this loss can be compensated by bringing a being with similar level of well-being into existence, leaving utilitarians with no direct reasons to condemn the killing.

Importantly, I argue that the scope of the replaceability argument is wider than commonly recognised, and that once replaceability has been introduced, it also applies to persons, that is, to self-aware beings. Thus, those accepting replaceability may also need to accept that adult human beings are replaceable. This is a highly controversial position and, I contend, should make us pause and consider the plausibility of those versions of utilitarianism that endorse replaceability.

Introduction

While utilitarian thinkers commonly regard animals as moral patients, and as such take a strong stance against animal suffering, the same cannot be said about the killing of animals. For some utilitarians, killing an animal can be considered a morally neutral act, as long as the killing is done without causing any suffering, and the animal is not self-aware. This evaluation of killing animals depends on the idea that death does not harm (merely sentient) animals, or if it does, this negative utility can be compensated, a position known as the *replaceability argument*. In particular, the replaceability argument holds that the killing of an innocent being can be a morally neutral act, as long as the victim is replaced with a similar

creature, which will restore the previous level of overall utility in the universe. This argument was first presented in its most prominent form by Peter Singer in *Practical Ethics* (2011, 106-7), and has been intensely debated since.[1]

I will argue that utilitarians can account for the wrongness of killing, in both hedonistic and desire-based approaches, however, some versions of this ethical theory need to endorse replaceability. And those accepting replaceability may be unable to deem the killing of (at least some) innocent beings as wrong, leading them to highly controversial implications. To show this, I will analyse the utilitarian reasons against killing in both hedonistic and desire-based accounts, and how these reasons can be applied to the case of non-human animals. Then, I will elaborate on how the replaceability argument, as devised by Singer, undermines these reasons. Finally, I will discuss the main issues with the replaceability argument, in particular its supposed scope, to highlight the unsettling implications of this argument when taken seriously.

I The Wrongness of Killing

Before we examine the utilitarian assessment of the wrongness of killing, a couple of clarifications are needed. First, for simplicity's sake, I will assume that the killings are done painlessly and without creating any distress for the creature. Secondly, in my discussion, I will assume (unrealistically) that the killing will not affect other sentient beings. That is, I will ignore indirect reasons against the killing, such as the grief or fear that the killing will produce in others, in order to elucidate whether killing sentient creatures harms them.

First, let us examine the utilitarian reasons against killing. In the case of killing, and leaving aside indirect reasons, a utilitarian could argue that painlessly killing someone negatively impacts on their well-being. As their well-being has diminished, they have been harmed, and therefore, and other things being equal, we have committed a morally wrong act. However, to accept this as a direct reason against the killing of innocent beings, first it is necessary to determine whether it is true that death has a negative effect on the victim's well-being.

[1] It has been argued (Dombrowsky 1997, 43-44; Uniacke 1997, Kemmerer 2007) that replaceability implies that utilitarian theory cannot account for the wrongness of killing merely-conscious life. Consequently, accepting replaceability may imply that utilitarianism fails to account for the value of (at least some) life. And even more, some have argued that this includes, in some instances, the wrongness of killing self-conscious beings (Lockwood 1979). This is frequently regarded as a fatal criticism of utilitarian theory (Regan 2004, 206-11).

Traditionally, philosophers have considered that death harms the victim due to the deprivation it causes. Death takes our future from us, and with it, all future well-being. This position is known as the deprivation account of the badness of death. However, there is a serious problem for the deprivation account: even if death may curtail their possibility of future positive well-being, the victim is not around any more to experience that deprivation. So, although it is true that they are deprived of all future positive goods, they do not experience it. This, sometimes known as the experience condition, is the argument Epicurus (1994, 78) used to defeat the idea of the badness of death.

The Epicurean challenge can be responded in several ways. A popular response is to claim that we can elucidate the effect death has for an individual in a comparative way. We could compare, for instance, two possible worlds, one where the victim dies at time X and another one where the victim continues to live until it dies at a later time Z. If all other variables are held constant, the difference in value between these two worlds is the value of the victim's death. Whilst it is true that the victim cannot experience this frustration, they have been affected by their death: their overall well-being level is lower than it could have been. Thus, it is possible to claim that the victim has been harmed.

When discussing the morality of killing, Singer appeals to the loss of future pleasure that the victim will suffer. However, he adds that the victim will not be present to suffer this deprivation and thus, is not affected by it. Singer says

> This means that we cannot move automatically from valuing a pleasant life rather than an unpleasant one, to valuing a pleasant life rather than no life at all. For, it may be objected, being killed does not make us worse off; it makes us to cease to exist. Once we have ceased to exist, we shall not miss the pleasure we would have experienced. (2011, 87)

Hence, it would appear that Singer accepts some version of Epicurus' existence condition: he acknowledges that pleasure may be lost with the killing, but this loss is not suffered by the victim herself.

Let us clarify Singer's position and his acceptance of the existence condition. Imagine I need to decide whether to kill Sally, whose future life is destined to be a happy one, with a positive balance of overall well-being. If we are measuring utility in a hedonistic way, we need to account for the loss of value that her death will cause: pleasure will be eliminated. However, note that whilst value has been destroyed, no disvalue has been created: assuming that her death was painless and caused no fear, the killing has not created a surplus of suffering. Furthermore, this loss is merely accounted in impersonal terms, as the victim does not experience her loss. Thus, according to Singer, although pleasurable lives are valuable, the killing may not harm the victim herself.

However, I disagree with Singer. First, even in Singer's own terms, we can still deem her killing as wrong, given that the amount of pleasure in the universe has been reduced, making the universe worse (although not worse for anyone in particular). Hence, as reducing the amount of utility in the universe (in both personal and impersonal terms) is wrong, we can conclude that killing the creature is wrong (other things being equal). Secondly, I claim that Singer is mistaken when he declares that the victim has not been harmed. Even if the death has created no additional disvalue, the well-being level of the victim has been negatively affected by her death, thus making her worse off. Moreover, as she has lost all possibilities for positive value in her life, I contend that this harm is a great one, giving us a strong personal reason to condemn her killing.

Let us now turn to preference utilitarianism. The response of preference utilitarians is similar in the case of merely-conscious beings, that is, their killing will prevent the satisfaction of future preferences. Note that, similarly, the victim will not experience this frustration, so the loss of utility may also be impersonal. However, their response goes further in the case of self-conscious beings: a creature that is self-aware has an idea of itself through time, and can project themselves into the future, creating plans and holding preferences for their lives as a whole. Killing a self-conscious being will mean that their present preferences about the future will be frustrated. This frustration means that killing a self-conscious creature is worse, as along with the impersonal loss of value, there will be a personal one. Hence, killing them is a seriously wrong act.

Thus, utilitarians can, in principle, deem the killing of innocent beings as morally wrong. This may not be enough to award these creatures the right to live, as utilitarians may agree to harm a creature if it maximizes overall utility; but it gives them a *pro tanto* reason to oppose the killing of innocent beings. How is it possible that some sentient beings are considered replaceable then? Here, it is necessary to explore the question of the creation of disvalue. Although the mere loss of value is enough to resolve the initial question of the wrongness of killing, the fact that disvalue is not being created is relevant to the discussion of replaceability.

II Killing and Replaceability

First, let us consider hedonism. When discussing replaceability, note that, although the loss of pleasure created by the victim's death could be used to condemn her killing, this loss can be compensated. Namely, it is possible to restore the previous level of pleasure in the universe by bringing a similar being into existence, as long as its life will be as pleasurable as

the victim's. Importantly, as according to Singer the loss is only measurable in impersonal terms, it makes no difference who experiences the pleasure. This implies that sentient beings are replaceable: it is morally acceptable to kill them, provided that they are replaced by a future being with a similar level of well-being.

At this point, Singer needs to determine whether all sentient beings are replaceable, and if not, which of them are. In accord with a lengthy philosophical tradition (Locke 2008; Rachels 1975; Tooley 1983), Singer alludes to the division between persons (self-conscious beings) and non-persons (merely conscious beings). This, together with a desire-based approach to utilitarianism, will allow him to claim that, while merely conscious beings are replaceable, self-conscious beings are not. As we have seen, persons not only have pleasurable states, but also preferences for the distant future and for their life as a whole. Importantly, these future-oriented preferences will be frustrated if they are killed, creating additional disvalue. Thus, these beings have a personal interest in continued existence, and for this reason, Singer considers them not to be replaceable.

At this point we could ask why cannot this wrongness be compensated by bringing a similar being into existence, as done with merely-conscious beings. To reply to this question, we need to elucidate the values attached to the killing.[2] Imagine two creatures. The first one is Sally the cow, a merely-conscious being (for the sake of the argument let us assume that cows are not self-conscious beings) The other being is Kelly, a self-conscious human being. Both their lives have a positive balance of well-being, faring at a five of positive utility. In the case of Kelly, though, of those five points, three are related to present concerns, but the remaining two are linked to plans that are projected into the future, such as the desire to see her children grow, or the preference for a continued existence. So, how will each version of utilitarianism address the loss suffered by their death and their possible replacements?

In a hedonistic approach, when Sally the cow dies she loses five utility points. Similarly, when Kelly the human dies, she also loses her five utility points. Furthermore, it can be argued that both these losses are only accounted in impersonal terms, as they do not experience their loss. Moreover, and importantly, as their deaths have been void of suffering, the amount of negative utility in the universe has not increased. This means that all we need to do to compensate for their deaths is to bring into existence an equally happy being, that will restore the previous level of utility in the universe. Thus, in a hedonistic approach, both Sally and Kelly are replaceable in a similar manner. Some will try to counter this evaluation arguing that a human being will score higher in happiness than a non-human animal. How-

[2] My example is based on Jamieson's (1983) take on replaceability.

ever, note that even if we accept this, it does not mean that human beings are not replaceable, it only means that they are not replaceable by a cow, as their score would be lower. Nevertheless, it is still possible to replace a human with another human (or, perhaps, with multiple cows).

Here is where preference utilitarianism helps Singer avoid the undesired implication that persons are replaceable. When Sally the cow dies, she loses five utility points. However, when Kelly the human dies, she loses all her five points of utility, but given that her long-term preferences have been frustrated, she now scores a negative two. The key difference here is that, in a desire-based approach, the death of a self-conscious being creates disvalue, as the long-term preferences have been frustrated. This negative score implies that Kelly cannot be replaced by the creation of a similar creature.

One could question, however, whether the frustration of future-oriented preferences is enough to deem self-conscious beings irreplaceable. On the contrary, I argue, that this disvalue could still be compensated by the creation of other beings. It is possible, for instance, to create a being that will exceed the happiness level of the previous one. For instance, if we bring Mark into existence, he will enjoy an optimistic personality that will allow him to fare a positive well-being of seven. This is enough to compensate for the negative two of Kelly, and still restore those five impersonal points lost. Alternatively, it would also be possible to bring into existence, two cows or four rabbits, and their aggregated positive welfare will in fact exceed that of Kelly. Therefore, it would seem that, even in a desire-based approach, persons are still replaceable.

I have argued that both versions of utilitarian theory can account for the wrongness of killing, as long as the victim was destined to enjoy a positive balance of overall well-being. Killing the victim eliminates what utilitarianism considers to be the unique value, positive welfare, thus robbing the victim of all the good things in their life. This loss negatively affects the overall well-being of the victim, therefore harming her. Consequently, and other things being equal, killing the victim is morally wrong. In this way, utilitarians have a *pro tanto* reason reason to condemn the killing. The problems arise, however, when replaceability is introduced. As the harm inflicted on the victim can be compensated by the creation of another being, this harm becomes irrelevant, leaving utilitarians with no means to justify the wrongness of killing. Furthermore, we have seen that efforts to restrict replaceability to merely-conscious beings by appealing to long-term preferences may be unsuccessful too, leaving persons equally unprotected.

Importantly, although utilitarians are commonly criticised as having too weak a stance against killing, the problem created by replaceability is far more problematic. Critics frequently contend that the utilitarian *pro tanto* reasons against killing fail to adequately protect potential victims, as such reasons can be trumped if the act maximises utility. This is a serious criticism, as rules against the killing of innocents are frequently considered among our most basic moral intuitions. However, as we have seen, utilitarians who accept replaceability lack any direct reasons to condemn killing, including the killing of self-aware beings, such as adult human beings. This leads us to the perplexing conclusion that, as long as we bring new beings into existence, the harm done to the victim is irrelevant to the evaluation of the killing, even in the case of persons.

Does this mean that utilitarianism is a failed moral theory? Here it is relevant to note that not all versions of utilitarianism accept replaceability, as this position mainly depends on the assumption that value is to be accounted in an impersonal way. This is the type of approach Singer takes, a position known as the Total View. Nevertheless, this is a controversial view in utilitarian ethics, as are, in fact, all the other potential positions on how to account for utility in the universe. However, I believe that the implications that the replaceability argument has for the ethics of killing are serious enough to count against the plausibility of those views supporting it.

References

[1] Dombrowski, Daniel. 1997. *Babies and Beasts: The Argument from Marginal Cases*. Chicago: University of Illinois Press.

[2] Epicurus. 1994. In *The Epicurus Reader: Selected Writings and Testimonia,* edited by Brad Inwood and Lloyd Gerson. Indianapolis: Hackett.

[3] Jamieson, Dale. 1983. “Killing Persons and Other Beings.” In *Ethics and Animals,* edited by Harlan Miller and William Williams. Clifton: Humana Press.

[4] Kemmerer, Lisa. 2007. “Peter Singer on Expendability.” *Between the Species* 13 (7): 1-10.

[5] Locke, John. 2008. *An Essay Concerning Human Understanding.* Oxford: Oxford University Press.

[6] Lockwood, Michael. 1979. “Singer on Killing and the Preference for Life.” *Inquiry* 22 (1-4): 157-70.

[7] Rachels, James. 1975. “Active and Passive Euthanasia.” In *Bioethics: An Introduction to the History, Methods, and Practice*, edited by Nancy S. Jecker, Albert R. Jonsen, and Robert A. Pearlman, 77-82. Washington: Jones & Bartlett Learning.

[8] Regan, Tom. 2004. *Empty Cages: Facing the Challenge of Animal Rights*. Los Angeles: Rowman & Littlefield.

[9] Singer, Peter. 2011. *Practical Ethics*. Cambridge: Cambridge University Press.

[10] Tooley, Michael. 1983. *Abortion and Infanticide*. New York: Oxford University Press.

[11] Uniacke, Suzanne. 1997. “Replaceability and Infanticide.” *The Journal of Value Inquiry*, 31 (2): 153-66.

From Utilitarianism to Prioritarianism

An Empathy-Based Internalist Foundation of Welfare Ethics

Christoph Lumer, University of Siena, Italy

Abstract

The article develops an internalist justification of welfare ethics based on empathy. It takes up Hume's and Schopenhauer's internalistic (but not consistently developed) justification approach via empathy, but tries to solve three of their problems: 1. the varying strength of empathy depending on the proximity to the object of empathy, 2. the unclear metaethical foundation, 3. the absence of a quantitative model of empathy strength.
1. As a solution to the first problem, the article proposes to limit the foundation of welfare ethics to certain types of empathy. 2. In response to the second problem, an internalistic metaethical conception of the justification of moral principles is outlined, the result of which is: The moral value of the well-being of persons is identical to the expected extent of (positive and negative) empathy arising from this well-being. 3. The contribution to the solution of the third problem and focus of the article is an empirical model of the (subject's) expected extent of empathy depending on (an object's) well-being. According to this model, the extent of empathy is not proportional to the expected empathy, but follows a concave function and is therefore prioritarian. Accordingly, the article provides a sketch of an internalist justification of prioritarianism.

I The Search for a Justification of Utilitarianism and a New Proposal - With a Prioritarian Outcome

The justification of utilitarianism is not exactly a success story. Mill's justifications (1998, ch. 4, par. 3-9), for example, are paradigmatic fallacies. Several justifications, in an intuitionistic, question-begging way, already presuppose certain moral principles – Hare (disguised by semanticism) (1981, sects. 1.3; 1.6) and Singer (1993, 11-12; 2011, 91-93; 100-102; 113-14) presuppose a certain form of universalization, Harsanyi (1953) presupposes ignorance of one's own identity (thereby operationalizing impartiality like Rawls) or the Pareto Principle plus the application of Bayesian Rationality to moral decisions (Harsanyi 1955). Still others build on – questionable – rationality-theoretical premises – in particular the equalization of one's own future time slices and the time slices of other persons (Sidgwick 1982, 381-82; 418-19; Parfit 1992, 281-82; 342; 346; Broome 1991, 231-37; 239-40). Most utilitarians do not even give a reason and only rely on their intuitive acceptance of utilitarianism (e.g.

Smart 1973, 3-8). But the research on the rational foundations of utilitarianism also contains unexploited potential, e.g. Hume's reflections.

This article develops a justification of a welfarist moral value function based on empathy, or, in Hume's (1978, 317-19) terminology, on sympathy. Here I will use the terms "*empathy*", "*sympathy*" and "*compassion*" interchangeably and with them mean: an emotion evoked by considering some person's or sentient being's well-being, that leads to the compassionate emotion, which may be negative or positive, according to the object's assumed negative or positive well-being. My justification takes up Hume's (1978, sects. III.2.2; 3.1-3) and Schopenhauer's (1977, §§15-6) internalistic (but not consistently developed and empirically flawed) approaches, but tries to solve three of their problems. The first problem, seen by Hume himself (but not satisfactorily solved), is: Morality formally requires universality and impartiality, while empathy varies with the temporal, spatial, social and personal distance from the object of empathy (1978, 580-82; 603). The second problem is the unclear metaethical basis of Hume's and Schopenhauer's approaches. The third problem, seen by neither of them, is that empathy is not proportional to the well-being of the empathy object:[1] An empirical study I conducted shows that compassion with negative well-being is more intensive than happiness about others' positive well-being.

My proposal for solving the first problem is that, in order to achieve universality and impartiality, which are necessary for the purpose of morality, the moral justification should be based only on certain universalistic forms of empathy: empathy that arises when considering the effects of one's own actions on the well-being of others (and not, for example, the empathy that arises from direct contact with others) (Lumer 1999). Unfortunately, this is only a very weak component of our total empathy but the only one which is subject-universalistic, i.e. leads to interpersonally identical valuations of the same objects (though there will rarely be valuations of the same object by different persons). The problem of the emotion's and therefore also the appertaining motivation's weaknesses may be resolved by taking the empathic emotion only as the signal which informs us about its object's moral value. This signal then has to be amplified by other motives which follow its lead. The most important such amplifiers are socially valid norms (Schopenhauer also suggested this (1977, 257-58)) and our feeling of moral self-worth. In the following I will not deal any further with this problem but will focus on the first and third problems.

[1] Hume, instead, seems to presuppose some proportionality between the pleasure of the persons affected and the spectators' sentiments: sympathy for the affected, love and hate for those changing their fate (1978, 591).

The proposal for the solution of the third problem is to study empirically how the degree of other persons' well-being influences our empathy.[2] More precisely: In the following an empirical model is developed, that calculates which extent of empathy (i.e. the integral of positive and negative empathy over time) occurs depending on the average well-being of an object of empathy. The expected extent of empathy is then the hedonistic and internalist moral reason for empathy-optimizing actions; and this empathy is also the basis and source of the internalist morality: The proposal equates the expected extent of empathy – which is identical to its expected hedonic desirability for the empathetic subject – with the moral value of the object's underlying well-being. The most important outcome of the model below is: Because of the greater intensity of negative empathy, the resulting moral value function is not utilitarian (linear function from well-being to moral desirability) – as a Humean may have guessed –, but prioritarian (concave function from well-being to moral desirability). This means the model provides a justification[3] and quantitative specification of prioritarianism.

In the following I will first (II) briefly explain the metaethical basis of the justification developed here and thereby outline my solution of the second, metaethical problem; this is only for understanding the approach, a further justification of this basis is not possible here. (III) Subsequently, I will present the empirical model of expected empathy in order to (IV) draw normative-ethical consequences.

[2] I have developed the model set out below in my professorial dissertation from 1992, which, however, was published only in 2000, 2nd edition 2009 (Lumer 2009). This paper is the first English presentation of the model.

[3] If prioritarianism is justified at all, exceptions aside, it is justified only intuitionistically, in particular as a middle way between utilitarianism, which is economic but does not intrinsically care about distributive justice, and maximin or leximin, which cares about distributive justice by giving priority to those who are worst off but in an extremist way. An exception is Hurley's (1989, 360-82) idea to introduce a risk-averse, concave weighting of prospects into a Rawlsian/Harsanyian framework of rational decision under uncertainty about one's identity. The result would be a concave, today we would say: prioritarian, moral value function. But Hurley did not elaborate this idea nor bring it together with the critique of utilitarianism and Rawls' difference principle; she envisioned her idea as something egalitarian – prioritarianism at that time was not yet a theoretical movement.

II Metaethical Foundations of the Justification of Morals [4]

What is a valid justification of morals at all? Justifications of morals, firstly, contain an *epistemically rational* component: By justifying these morals, one gains insights which distinguish them as something special. Secondly, valid justifications of morality contain a *practical* component: they are to have the consequence that the addressee of the justification adopts the justified morality as his own and, if possible, also acts on this basis.

The simplest and clearest way to bring the epistemic and the practical requirements together is to design moral justifications as arguments for a thesis about the object of justification, i.e. about the moral principle, etc. However, this cannot be *any* thesis; but the justification for this thesis must meet certain conditions. A thesis which fulfils these conditions is the *justification thesis* for moral principles. In this way, the epistemic requirement can be met by the fact that the justification still consists in an argumentatively valid and adequate argument which leads to a justified belief; and the practical and moral requirements can be met by selecting a particular thesis about the object to be justified, i.e. the justification thesis that this object has a certain justificatory quality *F*. I have developed several adequacy conditions for selecting this property *F*:

Adequacy Condition 1 (AC1): Motivation or practical requirement: Moral justification theses about moral principles are motivating in the sense that if a prudent addressee (i.e.: an epistemically and practically rational addressee with certain relevant information) is justifiedly convinced of the justification thesis (i.e. that the moral principle in question is *F*), he is motivated, at least to some extent, to adopt and observe the moral principle.

The motivation requirement is the specifically *practical* component of the conception for justifying moral principles. It makes the justification internalistic.

Adequacy Condition 2 (AC2): The motivating effect's stability with respect to new information: The motivating effect of a justified conviction of a justification thesis is stable with respect to new information, i.e. it is not lost as a consequence of acquiring additional true information.

Stability with respect to new information is the *rational* component of the concept of justifying moral principles. The only thing we can rationalize (in the sense of making it rational)

[4] Unfortunately, for reasons of space, this section is rather apodictic. A detailed explanation and justification of the presented metaethical approach can be found in: Lumer 2009, 30-127; 577-632; 2015.

directly are beliefs, indirectly also actions and other things. And the two main directions of that rationalization are: first, to make our beliefs true, i.e. to acquire possibly only true beliefs (or to correct false beliefs) by observing epistemological rules and, second, to increase the number of true beliefs. The requirement of the motivation's stability with respect to new information introduces the practically relevant maximum of epistemic rationality into the conception of practical justification.

Adequacy condition 3 (AC3): Moral instrumentality: Principles for which the justification thesis is true fulfill the function of moral principles, they meet the instrumental requirements for such principles and for morals in general.

Moral instrumentality is the specifically *moral* component of the conception of justification. If the "justified" moral principles do not fulfill the function of morality we are no longer dealing with a justification of a morality.

What is the function of morality? One can facilitate answering this question by distinguishing the structural components of morality. Normative morality consists mainly of a moral desirability function and moral norms, institutions and virtues. Once the moral desirability function has been established, it can be used to justify the other components of morality as more or less good means for realizing moral values. So, proceeding in this way, only the practical function of the moral value order has to be determined.

One can distinguish an *individual* and a *socially binding morality*, where the latter is designed to regulate social relations in an intersubjectively binding way. Here I will mainly deal with the second type. The sense of a socially binding moral desirability function could be *prudential-consensualistic*:

1. *Consensualistic requirement:* Socially binding moral evaluation criteria constitute a common moral value system that provides the intersubjectively shared standard (i) for assessing socially relevant measures, (ii) for planning social projects and (iii) for consensual arbitration of interpersonal conflicts of interest. In addition, for individuals the purpose or sense of such an intersubjectively shared value system could be to procure a benchmark for self-transcendent ego ideals and actions. I call this quality of the desired moral value functions "*subject universalism*", i.e. the value of all value objects (or more precisely the value relation of every two value objects p and q (= $U(p)/U(q)$) of this value function is roughly identical for all (or nearly all – except e.g. for psychopaths) moral subjects of the moral community.

2. *Prudential requirement:* The prudential requirement is that the subjective value functions to be compared according to subject universalism be parts or components of the subjects' prudential desirability functions. *Prudential desirability functions* express what is good for

the respective subject and hence, rationally or from a prudential point of view, should be the guideline of the subject's decision. Prudential desirability functions are constructed similarly to the utility functions of rational decision theory but with much stricter, philosophically developed standards, which also permit the criticism and correction of the subject's present instrumental or even intrinsic preferences (cf. e.g. Brandt 1979, part I; Lumer 2009, 241-428; 521-48). Prudential desirability functions are intersubjectively different – that I have a headache is mainly bad for me and neutral for you, and the reverse holds for your headache –; otherwise they could not express the *personal* good. Therefore, the subject-universalistic requirement is not intended to refer to complete prudential desirability functions but only to components thereof, i.e. parts of the total value which derives from particular types of consequences of the value object.

This concludes the metaethical considerations regarding the justification of morals; now the exposed conception has to be applied. The next step is empirical, viz. to enquire empirically, with the help of empirical decision theory and moral psychology, which component *V* of the prudential desirability function *U* is subject-universal and hence can be adopted as the moral desirability function. The result of a respective scrutiny is that interpersonally (nearly) identical components of our prudential desirability functions arise from our expected *compassion* and our expected *feelings of respect*. Of these two subject-universal feelings and motives, however, compassion is much better suited as the basis of the moral desirability function. For unlike compassion, one can hardly specifically optimize one's feelings of respect; respect is rather passive, it evokes motives for defending the respected object, but not motives for creating or improving respected objects. Therefore, in the following I will develop a model of a prudential desirability function based on empathy, or more precisely: a model of expected empathy depending on the well-being of other people. This expected empathy, in turn, corresponds to its hedonic desirability for the empathic subject. Ultimately, desirability procured through empathy is the sought-after subject-universal component of the prudential desirability function, which defines moral desirability. In short: The extent of expected empathy (according to the empirical model) is equated with moral desirability.

III An Empirical Model of the Expected Extent of Empathy

So the present task is to develop a – simplified – quantitative model of how the well-being of other persons whom we neither particularly like nor dislike is reflected in our expected

sympathy, i.e. the expected amount of our feelings of positive and negative sympathy. In short the model informs about the (expected) extent of our sympathy depending on other persons' well-being. The model's most important simplifying assumptions are these: 1. The object of our sympathy is the assumed well-being of the person(s) for whom we feel sympathy. 2. Errors in our assumptions about other persons' well-being statistically offset each other. 3. The model deals with *universal* sympathy only, i.e. a kind of sympathy we feel for strangers whom we neither like nor dislike in a particular way and whose behavior we do not judge in a moral way. 4. In a very flexible society like ours, the chances to be confronted with the lot of other people are equal for all objects of sympathy. And the salience of the fate of other people is equally distributed statistically. 5. The intensity of our compassion depends on the intensity and duration of considering it. But again, the expected values of these two quantities are intersubjectively equal for all objects of sympathy. 6. Prudent subjects have feelings of sympathy and do not try to avoid them.

The first step in developing this model is to determine the intensity of our sympathy depending on the assumed condition of the object. Consider figure 1.

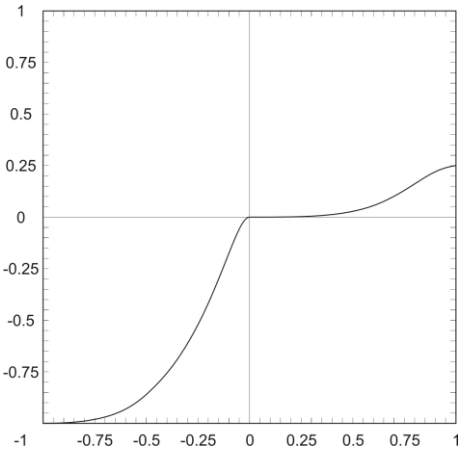


Fig. 1: Sympathy $S(x)$ depending on assumed well-being x

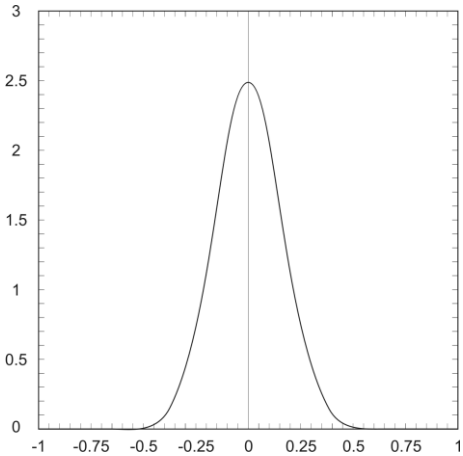


Fig. 2: Distribution *PD(x)* of well-being x for $x_{\mu}=0$

The x-axis represents the object's well-being; positive values represent pleasant feelings, negative values represent unpleasant feelings. The y-axis represents the appertaining sympathy, negative values representing pity and positive values representing pleasant feelings of sharing joy or the other person's flourishing. The other person's well-being as well as the sympathy are normalized into the interval [-1;1] with 0 being the point of indifference. Plausible assumptions about the function from well-being to sympathy are: The sympathy function ascends monotonously. To neutral well-being we are sympathetically indifferent; i.e. the function includes the point (0;0). Negative sympathy, i.e. pity, is much more intense

than positive sympathy. At the time when I developed this model I conducted interviews for testing the willingness to exchange packages of such feelings with different durations. This kind of willingness was then hedonistically reinterpreted as the subject's comparative judgement of the respective extents of sympathy. According to these calculations, pity for the most extreme sort of suffering was 4 to 10 times more intensive than positive sympathy with the most extreme form of joy. Conservatively I have taken 4 to be the right relation. The most extreme points of the function of figure 1 then are (-1; -1) and (1;0.25). Empirically our normal well-being ranges between 0 and 0.4; our sympathetic reaction to this kind of normal well-being is minimal. Outside of this region of normalcy sympathy's intensity increases rapidly, though much more rapidly versus negative than versus positive. When approaching extreme states of well-being sympathy will be satiated. – From these assumptions one gets the sympathy function designed in figure 1.

The most important feature of this function is that it is not linear: Pity is much more intense than positive sympathy; and normal states of well-being (between 0 and 0.4) are nearly neglected by our sympathy.

The second step of the model is to find out the intrasubjective distribution of well-being for different objects of sympathy over their life-time. For establishing the extent of sympathy, we need not know the exact course of the object's well-being but only the proportional duration of the single levels of well-being during the whole life. Again simplifying, I assume that these well-being levels are distributed normally. The open parameters of such a normal distribution are, first, the mean μ and, second, the spread σ. Empirical research on well-being has revealed that the intersubjectively most extreme long-term means of well-being of the unhappiest and the happiest people, positively-linearly transformed in our scale (-1;1), lie between 0 and 0.4 ($0 \leq \mu \leq 0.4$), so that the happiest people in the long run arrive at a mean of 0.4. Continuing the simplification, I assume that the mean levels of well-being of happy and unhappy people are intersubjectively different, but that the spread remains the same. Relying on some plausible assumptions about the absolute duration of very extreme feelings, the spread can be calculated as being equal to $\sigma=0.16$. The resulting curve for $\mu=0$ is shown in figure 2. In this way one gets a bundle of curves of normal distributions each representing the distribution of different well-being levels for typical more or less happy individuals; all these curves are equally shaped but their means range from 0 to 0.4 – according to the individual happiness –; i.e. the curves are shifted to the left or to the right with the top of the curves ranging between 0 and 0.4.

The third step is to multiply the probabilities given by the normal distribution of well-being with the sympathy function and to calculate the integral from -1 to 1 over this product function. The result of this operation is the expected extent of sympathy, i.e. the sum of all

feelings of sympathy which one expects to feel for a given person depending on the mean well-being μ of this person. This operation can be repeated for all the long-term means μ of well-being from the empirically expected range of such means, i.e. the interval from 0 to 0.4. The result is the function of the extent of sympathy depending on the long-term mean level μ of well-being. Normalizing the mean levels of well-being as well as the resulting extents of sympathy by a positive-linear transformation into the interval [0;1] one gets the normalized function of the extent of sympathy: ESN(m). This function is represented in figure 3.

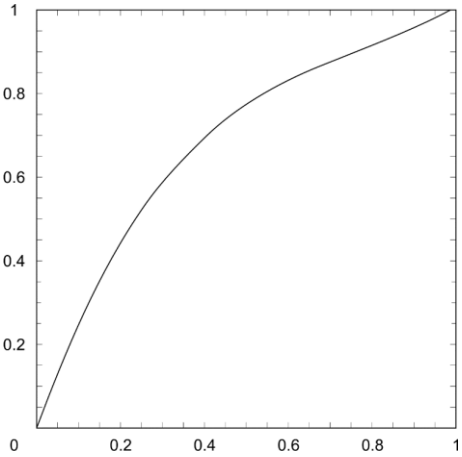


Fig. 3: Normalized extent of sympathy *ESN*(*u*) depending on the long-term mean level *u* of well-being

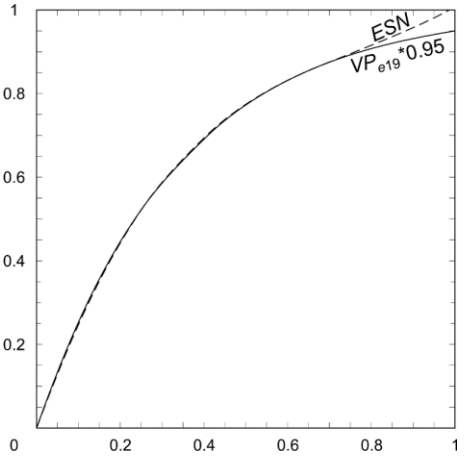


Fig. 4: Comparison of the normalized extent of sympathy *ESN*(*u*) with utilex VP_{e19}·0.95 (exponential value function)

In this function the *x*-axis represents the normalized lifetime mean-levels of well-being; and the *y*-axis represents the normalized expected extent of sympathy resulting from facing persons having the respective mean-level of well-being.

If somebody wants to value some social order from a purely sympathetic perspective he can assess the various mean levels of well-being of the people living in this society, find out the appertaining extent of sympathy and, finally, sum up these extents of sympathy. This, of course, is the same procedure which a hedonist prioritarian has to use to assess the prioritarian value of this social order. The only difference is that the prioritarian uses the prioritarian welfare function instead of the function of the extent of sympathy.

For formal mathematical reasons, but above all for metaethical reasons, one would like to have functions with certain properties as prioritarian weighting functions: They should be concave throughout, i.e. have a constantly decreasing gradient, rise monotonously, etc. For

this purpose I have discussed several mathematical curve families (Lumer 2005, sect. 3.1). The most suitable of these curve families are exponential curves:

for $e > 1$: $VP_{ee}(u) = e/(e\text{-}1)) \cdot (1\text{-}e^{-u})$; and

for $e = 1$: $VP_{e1}(u)=u$; this is identical to the right-hand limes of $VP_{ee}(u)$ for $e \rightarrow 1$ (see figure 5).

$VP_{ee}(u)$ is the family of **e**xponential **P**rioritarian **V**alue functions with the parameter *e*, where "*e*" within the function is a parameter equal to or larger than 1 (and does not mean Euler's number), which expresses the degree of priorioritarianism: the higher the number *e*, the stronger the prioritarian inclination. With *e*=1 the prioritarian inclination does not exist; the curve coincides with utilitarianism. With extremely high values for *e* the function creates leximin preferences. *e*-values between these extremes represent more or less radical forms of prioritarianism.

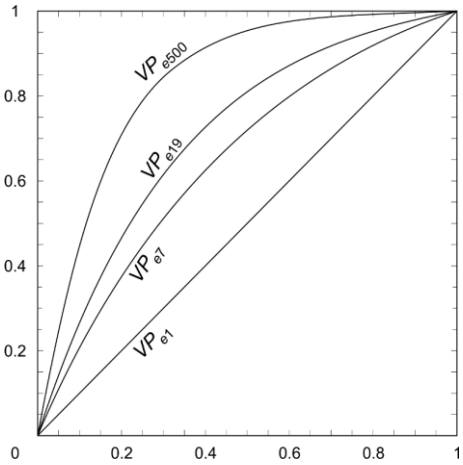


Fig. 5a: Exponential functions: VP_{e1}, VP_{e7}, VP_{e19}, VP_{e500}

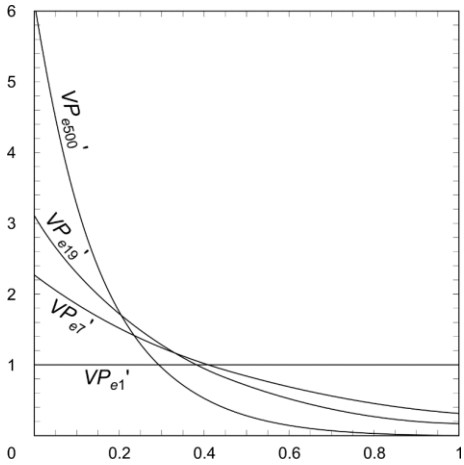


Fig. 5b: First derivations VP_{ee}' of exponential functions

One can now compare the empirically established function of the extent of empathy with these ideal prioritarian curves. The one that fits best is the curve for *e*=19. The two curves are compared in Figure 4. (The prioritarian function has been compressed by the factor 0.95 in order to facilitate the comparison.) One can easily see that, for a big stretch the two functions are more or less identical. That is why I have proposed the exponential prioritarian curve with *e*=19 ($VP_{e19}(u)$) as the internalistically justified prioritarian weighting function.

The function of the extent of sympathy just presented is based on some rather provisional measurements. But its general prioritarian shape is rather stable with respect to changes of

these assumptions and measurement results. So the exact function may be changed by re-measuring but the prioritarian shape will remain, because it depends only on the stronger intensity of pity as compared to positive sympathy.

IV Conclusion

On the basis of all these considerations we can now draw the conclusion: The internalist justification strategy for value ethics based on the adequacy conditions presented in section 2 and the prudential-consequentialistic determination of the function of socially binding morals, via an empirical scrutiny of possible subject-universal components of the prudential desirability functions has led to identifying empathy with others whom we neither like nor dislike in a particular manner as the sought source of the moral desirability function. On the basis of prudential hedonism, the empirical model of the expected extent of sympathy depending on other persons' (mean life-time) well-being provides the quantitative specification of this prudential desirability function. This function is mathematically simplified as $VP_{e19}(u)$, so that this function is therefore proposed here as the internalistically justified moral value function. This value function is universalistic, welfaristic and prioritarian. In the next parts of the theory, on the basis of this value function, certain moral norms, institutions, virtues, etc. can be justified as good means of realizing moral values.

What has been achieved with the study presented here? 1) If one tries to justify welfare ethics internalistically in the manner outlined above through compassion, the result is a version of *prioritarianism*, not utilitarianism (i.e. a concave not a linear moral value function). 2) In this way, prioritarianism has been justified internalistically, i.e. with recourse to (pre-moral) motives. This goes far beyond a merely intuitive acceptance of prioritarianism. 3) Prioritarianism has been quantitatively specified, beyond a vague comparative intuition, in a way that is needed for complex moral assessments with the comparison of many different consequences for different persons. From an infinite spectrum of more or less radical forms of prioritarianism, a specific one is distinguished as internalistically justified.

References

[1] Brandt, Richard B. 1979. *A Theory of the Good and the Right*. Oxford: Clarendon.

[2] Broome, John. 1991. *Weighing Goods: Equality, Uncertainty and Time*. Oxford/Cambridge, MA.: Blackwell.

[3] Hare, Richard M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon.

[4] Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61: 434-35.

[5] _____. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63: 309-21.

[6] Hume, David. 1739-40/1978. *A Treatise of Human Nature*. Edited, with an Analytical Index by L. A. Selby-Bigge. 2nd edition with text revised and variant readings by P. H. Nidditch. Oxford: Clarendon

[7] Hurley, Susan L. 1989. *Natural Reasons: Personality and Polity*. New York/Oxford: Oxford University Press.

[8] Lumer, Christoph. 1999. "Intergenerationelle Gerechtigkeit: Eine Herausforderung für den ethischen Universalismus und die moralische Motivation." In *Was heißt Gerechtigkeit?: Ethische Perspektiven zu Erziehung, Politik und Religion*, edited by Reinhold Mokrosch and Arnim Regenbogen, 82-95. Donauwörth: Auer.

[9] _____. 2000/2009. *Rationaler Altruismus: Eine prudentielle Theorie der Rationalität und des Altruismus*. 2nd supplemented edition. Paderborn: mentis.

[10] _____. 2005. "Prioritarian Welfare Functions – An Elaboration and Justification." *Working paper*, Department of Philosophy and Social Sciences, University of Siena. http://www.lumer.info/wp-content /uploads/2020/07/A066_Lumer_PrioritarianWelfareFunctions.pdf.

[11] _____. 2015. "Ethical arguments for moral principles." In *Proceedings of the 8th International Conference of the International Society for the Study of Argumentation, Amsterdam, July 1-4, 2014*, edited by Bart Garssen, David Godden, Gordon Mitchell and A. Francisca Snoeck Henkemans. Amsterdam: Sic Sat. http://rozenbergquarterly.com/issa-proceedings-2014-ethical-arguments-for-moral-principles/?print=pdf.

[12] Mill, J. S. 1861/1998. *Utilitarianism*. Edited by Roger Crisp. Oxford/ New York: Oxford University Press.

[13] Parfit, Derek. 1984/1992. *Reasons and Persons*. Oxford: Clarendon.

[14] Schopenhauer, Arthur. 1840/1977. “Preisschrift über die Grundlage der Moral.“ In *Werke in zehn Bänden. Zürcher Ausgabe, Vol. VI,* 143-315. Zürich: Diogenes [Engl. Transl.: *On the Basis of Morality*. 2000. Translated by E. F. J. Payne. 2nd edition. Indianapolis: Hackett.]

[15] Sidgwick, Henry. 1874/1982. *The Methods of Ethics*. Indianapolis/ Cambridge: Hackett.

[16] Singer, Peter. 1979/1993. *Practical Ethics*. 2nd [enlarged] edition. Cambridge: Cambridge University Press.

[17] _____. 1981/2011. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. 2nd edition. Princeton/Oxford: Princeton University Press.

[18] Smart, J. J. C. 1961/1973. “An outline of a system of utilitarian ethics.” In *Utilitarianism for and against*, edited by Idem and Bernard Williams. 3-74. Cambridge: Cambridge University Press.

How to Define ‘Prioritarianism’ and Distinguish It from (Moderate) Egalitarianism

Christoph Lumer, University of Siena, Italy

Abstract

In this paper, first the term ‘prioritarianism’ is defined, with some mathematical precision, on the basis of intuitive conceptions of prioritarianism, especially the idea that "benefiting people matters more the worse off these people are". (The prioritarian weighting function is monotonously ascending and concave, while its first derivation is smoothly descending and convex but positive throughout.) Furthermore, (moderate welfare) egalitarianism is characterized. In particular a new symmetry condition is defended, i.e. that egalitarianism evaluates upper and lower deviations from the social middle symmetrically and equally negatively (as do e.g. variance and Gini). Finally, it is shown that this feature distinguishes egalitarianism also extensionally from prioritarianism.

Introduction: Open Problems of Prioritarianism and the Aims of This Paper

Egalitarianism and prioritarianism are important ways of correcting utilitarianism for considerations of justice (others are sufficientarianism and leximin). (Telic) egalitarianism aims at diminishing (or eliminating) intersubjective differences in personal goods, in particular individual utilities (Parfit 1997, 204). Prioritarianism on the other hand, wants each person to fare as well as possible, but is especially concerned with those who are worse off. From this idea we get Parfit’s *prioritarian slogan*: "Benefiting people matters more the worse off these people are" (Parfit 1997, 213). While egalitarians are concerned with relativities, i.e. how each person’s level compares with the level of other people, prioritarians are concerned with absolute levels, giving the higher priority to improving the situation the lower the beneficiaries fare in absolute terms (Parfit 1997, 214).

Prioritarianism has many advantages with respect to other criteria of distributive justice, which, however, I can not discuss here. Despite these advantages, up to the present prioritarianism has not been elaborated that much and – among others – the following problems still have to be resolved: 1. (Moderate welfare) egalitarianism as well as prioritarianism, both fulfil the Pigou-Dalton condition and can be represented by concave welfare-functions. Does there then remain any difference between these two approaches and, if yes, what does it consist in? 2. More generally, how can ‘prioritarianism’ and ‘(moderate welfare)

egalitarianism' be precisely defined? 3. If prioritarianism is to be applied in practice the degree of priority has to be established. What exactly is the prioritarian welfare function? 4. Prioritarians have described their *intuitions* about priority. Is there any deeper, in particular internalist, justification of prioritarianism? – In this article I sketch an answer to the first two questions. In a parallel paper (Lumer 2020) I propose an answer to the fourth question; and in Lumer 2005 (22-32) I have provided an answer to the third question.[1]

I Defining 'Prioritarianism'

Parfit has summarized and systematized the ideas of a number of other philosophers, given this system the name "*priority view*" and coined the prioritarian slogan: "Benefiting people matters more the worse off these people are" (Parfit 1997, 213). A somewhat different way of explicating prioritarian intuitions is to take prioritarianism as a synthesis of utilitarianism and leximin somewhere between these two systems, which preserves the advantages of both, utilitarianism's efficiency and leximin's concern for those badly off, and removes their respective one-sidedness's, utilitarianism's neglect of distributive justice and leximin's inefficient and hard-hearted intrinsic disregard of improvements for those better off (even the second worst off) (Lumer 1997, 102; 2009, 628-32; Temkin 2003).

'Prioritarianism' may informally be defined like this:

Prioritarianism is a way of intrinsically morally valuing individual situations, according to which (small) changes in personal well-being, or more generally: personal desirability of the situation, are morally valued in strict positive correlation to these changes but giving more – though not infinitely more – weight to changes for people being badly off; this weight declines continuously and smoothly with increasing personal desirability, however without ever reducing to zero – not even for the highest levels of personal desirability.

The different weights express the degree of our moral concern, i.e. how close improving the lot of the person in question is to the moral subject's heart. Because this desirability function is applied to life situations of *individuals*, the moral value of a *group's* state can be established additively.

[1] The present article to a great extent is an abridged version of a part of an unpublished working paper of mine (Lumer 2005, sect. 2). The parallel article (Lumer 2021) mainly relies on material – so far published only in German – of my habilitation thesis from 1992: Lumer 2009, 589-632.

The straightforward way of formally modeling prioritarian valuing is to define a one-adic moral value function over normalized personal desirabilities – which may range from 0 to 1. This is represented in figure 1a (source: Lumer 2005, sect. 2.1).

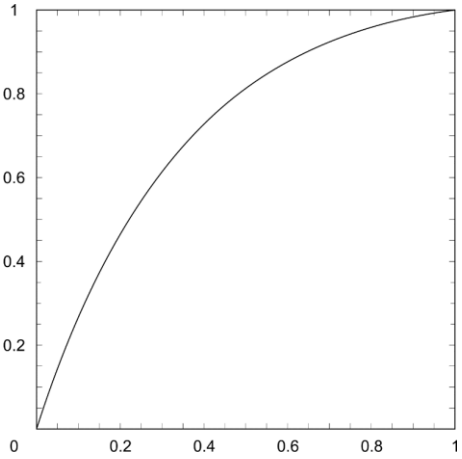


Fig. 1a: Prioritarian value function *VP* (VP_{e19})

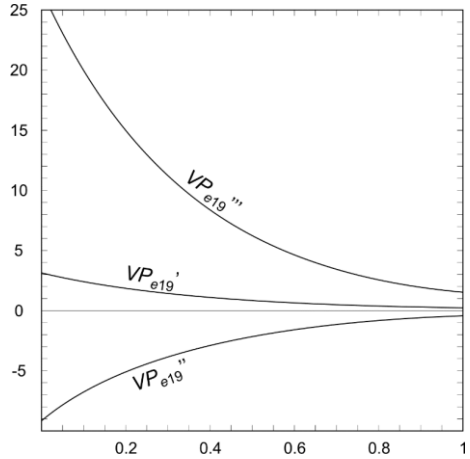


Fig. 1b: Derivations of prioritarian value function

The moral value function increases monotonously because of the strict positive correlation between personal desirability changes and their moral assessment. Therefore, and because of the normalization it has to cross the points (0;0) and (1;1). But the moral value function is concave; it increases less and less steeply, without ever arriving at a slope of zero. The first derivation of this moral value function is represented by the middle curve of figure 1b (VP_{e19}'). It expresses more intuitively the idea of prioritarianism as it is coined in the slogan than the value function itself, namely the degree of our concern for, the weight we attribute to *changes* of other people's well-being. This weight is positive allover but it decreases monotonously and smoothly; and because it never reaches zero, not even for the highest well-being, the curve of the first derivation has to be strictly convex (otherwise it would intersect the *x*-axis at some point). Mathematically this means that the second derivation has to be negative allover and must be monotonously increasing (see fig. 1b, VP_{e19}''). The welfare of a group or a society, finally, is defined as the sum of the moral desirability of the situations of its members.

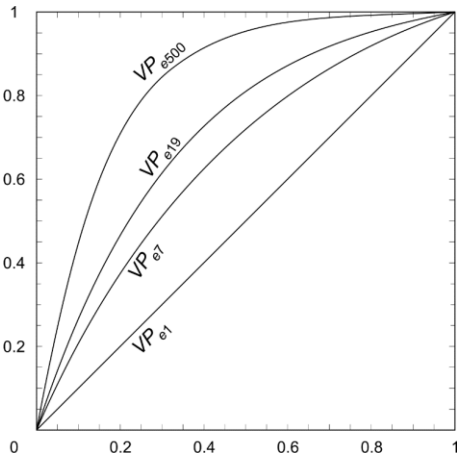


Fig. 2a: Prioritarian functions: VP_{e1}, VP_{e7}, VP_{e19}, VP_{e500}

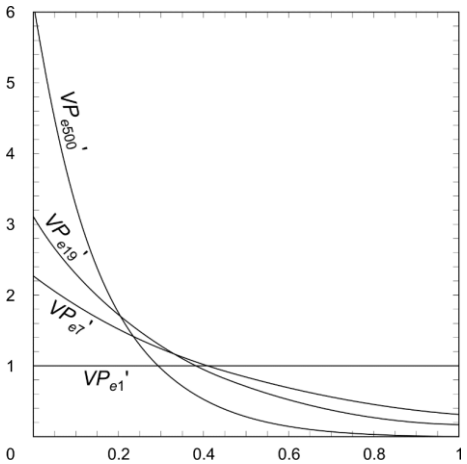


Fig. 2b: First derivations VP_{ee}' of prioritarian functions

As suggested by figure 2a, there are infinitely many functions having the features just described. The two limiting cases are, first, the diagonal itself, i.e. the identity function, according to which the moral value is identical to the personal desirability; this is the utilitarian way of valuing with a zero-degree of priority so to speak. And, second, there is the right angle connecting the points (0;0), (0;1) and (1;1), or more precisely a function which approaches this angle and cannot be visually distinguished from it; this function represents leximin. Prioritarian desirability functions have to be between these two limiting cases, which expresses that prioritarianism is a synthesis of utilitarianism and leximin.

The features explained so far are sufficient for formally defining 'prioritarianism':

Abbreviations:

VPT(*a*) = prioritarian value function (under certainty) over objects *a* (e.g. actions).

[*VPP*(*a*) = prioritarian value function (under risk) over objects (prospects) *a* (e.g. actions).]

VP(*x*) = prioritarian weighting function over personal desirabilities *x*.

$U_i(a)$ = personal utility / desirability of object *a* for person *i*.

Definition:

Prioritarianism is a way of moral valuation that can be represented by an

(P1) additively separable moral value function $VPT(a)$ of the form:

$VPT(a) := \sum_i VP(U_i(a)) = \sum_i VP(u_i)$ for certain prospects a

(readers not interested in the valuation of risky prospects can skip conditions P2 and P3)

[(P2) and $VPP(a) := R[\langle VPT(a_1), P(a_1)\rangle, ... \langle VPT(a_m), P(a_m)\rangle]$ for risky and uncertain prospects $a = (\langle a_1, P(a_1)\rangle, ..., \langle a_m, P(a_m)\rangle)$ – a_i is a possible outcome of a, and $P(a_i)$ is its probability –, where]

[(P3) $R(x_1, ..., x_m)$ is a suitable monotonously increasing weighting function for not certain prospects with $R(0) = 0$ and $R(\langle VPT(a), 1\rangle) = VPT(a)$,]

(P4) and where $VP(u)$ is a three times differentiable value function with

(P4.1) $VP'(u) > 0$ for all u,

(P4.2) $VP''(u) < 0$ for all u,

(P4.3) $VP'''(u) > 0$ for all u, and

(P5) for which a set of real (at some point in history) options $\{a, b\}$ exists with $VPT(a) > VPT(b)$ which is in contrast to the leximin valuation (because a entails some greater utility for people better off than b for some people worse off).

For the subsequent comparison of prioritarianism to egalitarianism it is helpful to consider the following feature. Prioritarian value functions remain above the diagonal (see figure 2a) so that one can examine the mathematical qualities of the piece over the diagonal, too, i.e. the curve which results from subtracting the diagonal from the desirability function. This difference function may be called the "*surplus function*"; it is shown in figure 3 (SP_{e19}, i.e. the more horizontal graph; the other graph in figure 3 represents the first derivation of the surplus function, SP_{e19}').

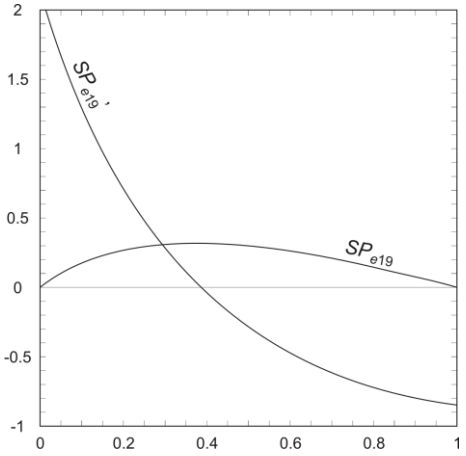


Fig. 3: Prioritarian surplus function: *SP*(*u*) := *VP*(*u*)-*u*

In prioritarianism this surplus function intuitively makes no sense; it is just the result of a mathematical transformation; but it can be compared to egalitarian *surplus* functions. The prioritarian surplus function goes from (0;0) to (1;0), is concave and constitutes a hill between these points. However, the characteristic property, which distinguishes prioritarian surplus functions from some egalitarian surplus functions, is that the prioritarian surplus functions are right-skewed: they ascend steeper on the left side than they descend on the right side. (Its first derivation is identical to the first derivation of the prioritarian value function shifted downwards by one unit. The second and third derivations of the surplus function are identical to those of the prioritarian value function itself.) I will come back to this feature below.

II Trying to Define 'Egalitarianism' in Opposition to Prioritarianism

What is egalitarianism? Parfit has distinguished telic egalitarianism from deontic egalitarianism, where the former is interested in the final distribution, intrinsic from instrumental egalitarianism, and moderate from radical (or pure) egalitarianism, where the former satisfies the Pareto-principle. In addition, egalitarianisms have to be distinguished according to the good they hold to be distributed equally (Parfit 1997, 203-9). In the following I will speak only of moderate, telic and intrinsic egalitarianism of utilities because this version is the most difficult to distinguish from prioritarianism. *Moderate* egalitarianism is not only interested in equality but also in a high sum of personal utilities.

Some theorists have a rather loose way of using the label "egalitarianism" so that egalitarianism includes prioritarianism or leximin. Here I will use the expression "egalitarianism" in a narrower, more specific sense, namely – provisionally –: egalitarianism cares about *equality* in the sense that it tries to *diminish inequalities*; it values lower and upper deviations from the middle (mean, median …) negatively, the greater they are the more negative.

How can egalitarianism and prioritarianism be distinguished? In the literature several differences are recognized: 1. different "justifications", or better: different aims (equality vs. priority); 2. interest in relativities vs. absolute levels; 3. lacking vs. present additive separability; 4. lacking vs. present strong separability; 5. interest in distribution patterns vs. interest only in absolute levels. However, so far there is still no *proof* of a compelling and extensionally relevant difference between egalitarianism and prioritarianism (for decisions under certainty). In the following I try to prove that there is such difference, which goes beyond the just mentioned: 6. symmetrical and increasing depreciation of deviations from the mean vs. smoothly decreasing care for those better off, which implies: 6.1. symmetrical vs. right-skewed surplus functions and 6.2. lack vs. presence of strong separability. In the following only feature 6.1 can be dealt with.

III The Essence of Egalitarianism: Symmetrical and Increasing Devaluation of Deviations from the Middle

So what is the essential core of moderate egalitarianism that leads to the demarcating formal, mathematical differences and then also extensional differences to prioritarianism? Despite egalitarianism's lack of additive separability (in contrast to prioritarianism, see P1), one may isolate individualized components of the egalitarian welfare function, i.e. vary the personal desirability for one person only (and keeping the desirability levels of all other persons constant, so that the social mean remains virtually unchanged) and see how these changes affect the egalitarian total welfare. If we consider such individualized functions, the purely egalitarian component of egalitarian value functions can be formulated in a negative way: Egalitarianism as such values deviations from a (hypothetical) state of equality as negative, the bigger these deviations are, the more negatively (more than proportionally) they are valued. This holds for downward deviations as well as, *ceteris paribus*, for upward deviations, which in this respect are valued symmetrically, i.e. equally negative, depending on the absolute value of the deviation alone. This symmetry is essential for egalitarianism be-

cause if somebody is exclusively interested in *equality*, the *direction* of deviation from equality should not matter; and if he is interested in equality only among other aspects – like, additionally, high sum or mean of individual desirabilities – the direction of deviation should not matter for the egalitarian aspects of his valuations. To summarize, pure egalitarianism and the egalitarian component in moderate egalitarianism here are characterized by two conditions, the *symmetry condition*, which says that upward and downward deviations from some middle must be valued equally negatively, and the *increasing weight condition*, which says that greater deviations should be valued increasingly, over-proportionally stronger. Moderate egalitarianism then may add the sum of utilities to this pure egalitarianism.

In contrast to this interpretation of 'egalitarianism', however, various contemporary theorists characterize (moderate desirability) egalitarianism by very broad conditions that do not imply the symmetry condition – e.g.: intrinsic badness of inequality, intrinsic badness of some being worse off than others, optimality of equality, Pigou-Dalton condition (Parfit 1995, 4; 1997, 204; Temkin 2003, 62-63; Tungodden 2003, 2; Fleurbaey 2015, 207; Voorhoeve 2015, 201). According to the argument just put forward, this would be too broad (so also: Broome 2015, 219). And this missing confinement of the concept 'egalitarianism' is confusing for the ethical systematics. For not only egalitarianism, but also prioritarianism fulfils these conditions at least extensionally. Thus, these conditions are not suitable for the demarcation of egalitarianism and prioritarianism. Prioritarianism, on the other hand, does *not* fulfil the *symmetry* condition (details below). Therefore, symmetry and increasing weight are not only characteristic of egalitarianism, but also suitable for the differentiation from prioritarianism.[2]

[2] Because Fleurbaey defines 'egalitarianism' very broadly – namely via the principles: 'equality is the best distribution' and 'inequality is intrinsically bad', to which he often adds the Pigou-Dalton condition and, in the case of moderate egalitarianism, also the Pareto Principle (Fleurbaey 2015, 207-8) – he considers prioritarianism extensionally only to be a special form of egalitarianism (ibid. 203; 207): prioritarianism "can be represented as a combined function of the average (or total) amount of benefit and of an inequality index" (ibid. 208). According to the argument just presented, Fleurbaey overlooks an essential characteristic of egalitarianism, viz. the symmetry condition, which leads to a narrower meaning of "egalitarianism". And he has a much too broad concept of 'inequality index', which also includes the prioritarian surplus function as the core of an inequality index, though it is right-skewed and completely detached from the social mean.

IV Specification of the Essential Conditions of Egalitarianism and Formal Demarcation from Prioritarianism

The symmetry and the increasing weight conditions, which have just been characterized informally, are now to be more precisely defined mathematically in order to specify the difference between prioritarianism and the forms of egalitarianism which are closest to prioritarianism in mathematical terms as well. There is a wide variety of egalitarian welfare functions with very different constructive features. Therefore, egalitarian welfare functions altogether are hard to compare to prioritarian ones. But at least some of them are constructed in a way that they subtract some measure of inequality from the sum of individual desirabilities. And again, some of these inequality measures are symmetrical in the sense that they count lower and upper deviations from some mean in the same way, furthermore they fulfil the increasing weight condition: e.g. variance, Gini-coefficient, Rescher's (1966, 33; 35-36) effective-average principle, Trapp's Utilitarianism incorporating justice (cf. Trapp 1988, 356; 1990, 365). The appertaining welfare functions are ideal types of egalitarian welfare functions. One can construct such ideal egalitarian welfare function as follows. First, one models the pure egalitarian part, as it is exemplified in figure 4a. (Figure 4a shows egalitarian surplus functions which lead to variations of the variance as inequality measure: $IC_{VARSp}(u) = -0.5 \cdot |0.5-u|^p + 0.5 \cdot 0.5^p$, with $p>1$; figure 4a represents the graphs for $p=1.5$, $p=2$, $p=3$.)

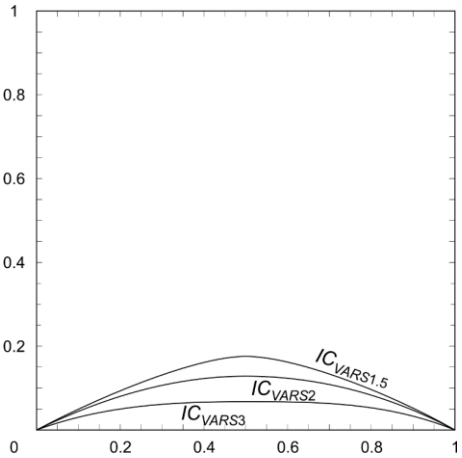


Fig. 4a: Inequality contribution (= equality surplus) of individual utility (u_μ fixed (u_μ=0.5)): $IC_{VARS1.5}(u)$, $IC_{VARS2}(u)$, $IC_{VARS3}(u)$

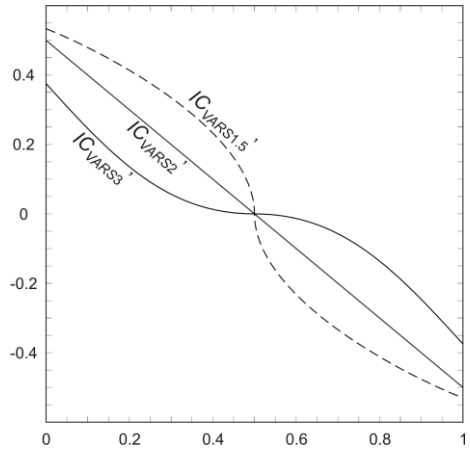


Fig. 4b: Derivations of inequality contributions (= equality surplus) of individual utilities $IC_{VARS1.5}'(u)$, $IC_{VARS2}'(u)$, $IC_{VARS3}'(u)$

The two conditions of symmetry and increasing weight are implemented by a surplus function, which gives some surplus value to the fact that a person's well-being is close to the social mean – which for simplicity we assume to be 0.5. The farther away a person's well-being is from this mean, the more the surplus value decreases. In addition, this deviation from the mean is valued over-proportionally so that we have a concave and not a linear decrease of the surplus function at both sides of the mean. In the usual models of moderate egalitarian value, a measure of inequality is *subtracted* from the utility sum (or average utility). The surplus function corresponds to the function of this inequality measure except that the surplus function is shifted upwards, so that the surplus values are positive for the normal utility interval [0;1]. This allows the comparison with prioritarianism without changing the order of preference.) In function IC_{VARS2}, shown in figure 4.a, for example, the inequality measure is shifted upwards by 0.125. The deductions for the deviation from the center thus become a surplus for the proximity to the center. This surplus function can now nicely be compared to the prioritarian surplus function – shown in figure 3: 1. *Egalitarian* surplus functions (fig. 4a) are axially symmetrical with respect to the social mean, whereas the prioritarian surplus functions (fig. 3) are right-skewed. For reasons of space I will not go into the mathematical details, but the axial symmetry of the egalitarian surplus functions, of course, has many further mathematical consequences for the derivations: point symmetry of the first derivation (fig. 4b), axial symmetry of the second derivation, point symmetry of the third derivation with respect to the point (u_μ;0). Right-skewness of the *prioritarian* surplus function, instead, means that upper deviations from the peak are valued less negatively than lower deviations. This feature of prioritarianism makes sense in the context of assessing welfare, i.e. *desirability* distributions: We can neither redistribute desirabilities from above to below, as is presupposed in resource egalitarianism; nor does above-average well-being directly cause harm to those badly off, as is presupposed in egalitarianism of power, rights and status for these distribuenda. Rather, right-skewness is only the mathematical consequence of a heavier weighting of changes for people who are badly off – completely independent of social distributions of individual desirabilities. 2. By definition, the peak of the egalitarian surplus function is attributed to the social mean (i.e. $IC(u_\mu)$). The position of the peak of the prioritarian surplus function, on the other hand, has no defined meaning, it can only be calculated; and it changes with the degree of prioritarianism: the stronger the degree of prioritarianism, the further to the left is the peak (i.e. the smaller is the u_x above which the peak is collocated).

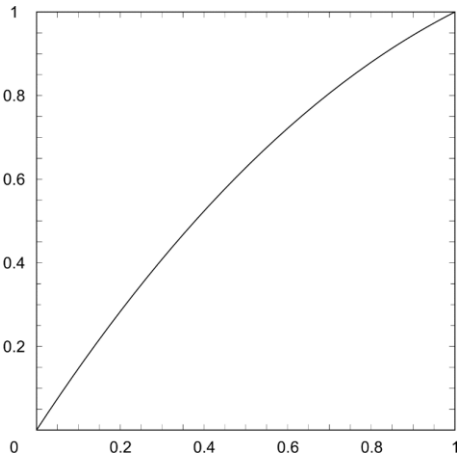


Fig. 5a: Egalitarian value function VE_{VARS2} based on variation, individual contribution u_μ fixed (u_μ=0.5)

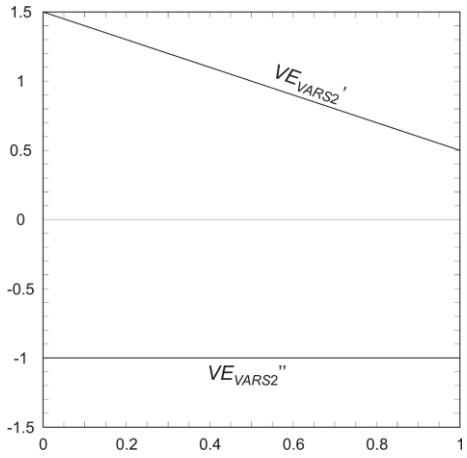


Fig. 5b: Derivations of egalitarian value function based on variation, individual contribution u_μ fixed (u_μ=0.5)

As a second step in modelling welfare egalitarianism one might want to include moderate egalitarianism, which apart from caring about equal distributions is also interested in increasing overall social well-being. This concern can be modelled by simply adding a utilitarian component to the surplus function, i.e. the diagonal – mathematically speaking. As a result, as can be seen in figure 5a, we get the concave value functions, which at first sight are similar to prioritarian value functions. But now we know that there are clear mathematical differences – at least between ideal egalitarian and prioritarian welfare functions –, which can easily be read from the surplus functions. The egalitarian surplus function is axisymmetric (with respect to x=u_μ) (fig. 4a), whereas the prioritarian surplus function is right-skewed (fig. 3). Therefore, the first derivation of the examined egalitarian surplus function is point-symmetric with respect to the point $\langle u_\mu, 0\rangle$ (fig. 4b) – as opposed to the convex first derivation of the prioritarian surplus function (function of fig. 2b shifted downwards by 1; the slant graph in fig. 3).

Do the mathematical differences between prioritarianism and moderate welfare-egalitarianism have any practical significance – in terms of different preference orders? Consider the following desirability distributions:

Symmetry litmus:

$a = \langle 0.75, 0.75, 0\rangle$,

$b = \langle 1, 0.25, 0.25\rangle$ (cf. Lumer <2000> 2009, 631).

a and *b* have the same sum of utilities and the same mean, namely u_{μ}=0.5. They are constructed in such a way that in *a* there are two upper deviations of 0.25 and one lower deviation of 0.5 from this middle, whereas in *b* these deviations are exactly reversed: two lower deviations of 0.25 and one upper deviation of 0.5 from the middle. So the structure of the example is: a = ⟨*m*+*d*, *m*+*d*, *m*-2*d*⟩, *b* = ⟨*m*+2*d*, *m*-*d*, *m*-*d*⟩, where *m* is the mean. Hence the utilitarian aspects of *a* and *b* are identical, and their egalitarian aspects are symmetrical. Therefore, all (real) egalitarian value functions, which fulfil, the symmetry condition, have to value *a* and *b* as equivalent. Prioritarian valuations, on the other hand, prefer *b* to *a*, and they have to do so because of the definitional properties of prioritarian evaluation functions, namely the continuously decreasing moral weight of desirability changes with increasing desirability level (represented by the first derivation of the prioritarian value function and which leads to the right-skewness of the respective surplus function). The example can therefore be used as a litmus test for fulfilling the symmetry condition and thus for distinguishing between truly egalitarian and other, especially prioritarian evaluations. This preference for *b* is generated even with minimal degrees of priority.

The prioritarian value difference between *a* and *b* is also reflected in many people's intuitions. In studies conducted in 2002-2004 with 79 participants who had to choose according to their moral intuitions between alternatives constructed in the fashion of *a* and *b* 81.0% (*n*=64) of the subjects preferred the analogue of *b*, i.e. decided in a prioritarian way; 13.9% (*n*=11) preferred the analogue of *a*; and only 1.3% (*n*=1) found the analogues of *a* and *b* equivalent, i.e. decided in a welfare egalitarian way (3.8% (*n*=3) gave no clear answer). This means, first, that the difference between *a* and *b* is not only technical gimcrackery but is intuitively seen as making a practical difference and, second, there are more prioritarians around than is usually assumed.

In summary, we have thus found an important, also extensionally relevant difference between prioritarianism and egalitarianism – if egalitarianism is understood only in a sufficiently specific way –: namely symmetrical and increasing depreciation of deviations from the mean (egalitarianism) vs. smoothly decreasing concern for the better-off with growing well-being (prioritarianism). This difference then implies the symmetry vs. right-skewness of the surplus function.

Acknowledgements

I thank Vuko Andrić, Annette Dufner and Rudolf Schüßler for inviting me to the workshop "Prioritarianism" (July 2018 in Karlsruhe), the participants of this workshop for their critical comments on an earlier version of this paper and Nils Holtug as well as Kacper Kowalczyk for their bibliographical indications.

References

[1] Blackorby, Charles, and David Donaldson. 1978. “Measures of Relative Equality and their Meanings in Terms of Social Welfare.” *Journal of Economic Theory* 18: 59-80.

[2] Broome, John. 2015. “Equality versus Priority: A Useful Distinction.” *Economics and Philosophy* 31: 219-28.

[3] Fleurbaey, Marc. 2015. “Equality versus Priority: How Relevant Is the Distinction?” *Economics and Philosophy* 31 (2): 203-17.

[4] Jensen, Karsten Klint. 2003. “What is the Difference between (Moderate) Egalitarianism and Prioritarianism?” *Economics and Philosophy* 19: 89-109.

[5] Lumer, Christoph. 1997. “Utilex – Verteilungsgerechtigkeit auf Empathiebasis.” In *Current Issues in Political Philosophy: Justice in Society and World Order*, edited by Peter Koller, and Klaus Puhl, 99-110. Wien: Hölder-Pichler-Tempsky.

[6] _____. 2000/2009. *Rationaler Altruismus: Eine prudentielle Theorie der Rationalität und des Altruismus*. 2nd, supplemented edition. Paderborn: mentis.

[7] _____. 2005. “Prioritarian Welfare Functions – An Elaboration and Justification.” (Unpublished) *Working paper*, Department of Philosophy and Social Sciences, University of Siena. http://www.lumer.info/wp-content/uploads/2020/07/A066_Lumer _PrioritarianWelfareFunctions.pdf.

[8] _____. 2021. “From Utilitarianism to Prioritarianism – an Empathy-Based Internalist Foundation of Welfare Ethics.” In *Proceedings of the ISUS 2018 Conference, Karlsruhe, July 23-26, 2018*, edited by Michael Schefczyk, and Christoph Schmidt-Petri. Karlsruhe: KIT Scientific Publishing.

[9] Parfit, Derek. 1995. *Equality or Priority? The Lindley Lecture, University of Kansas, November 21, 1991*. Lawrence, KS: University of Kansas. [Reprinted in *The Ideal of Equality*. 2000. Edited by Matthew Clayton, and Andrew Williams. New York: St. Martin’s Press.]

[10] _____. 1997. “Equality and Priority.” *Ratio, New Series* 10: 202-21.

[11] Rescher, Nicholas. 1966. *Distributive Justice: A Constructive Critique of the Utilitarian Theory of Distribution*. Indianapolis/New York/Kansas City: Bobbs-Merrill.

[12] Temkin, Larry S. 2003. “Equality, Priority, or What?” *Economics and Philosophy* 19: 61-87.

[13] Trapp, Rainer W. 1988. "*Nicht-klassischer" Utilitarismus. Eine Theorie der Gerechtigkeit*. Frankfurt: Klostermann.

[14] _____. 1990. "'Utilitarianism Incorporating Justice'. A Decentralised Model of Ethical Decision Making." *Erkenntnis* 32: 341-81.

[15] Tungodden, Bertil. 2003. "The Value of Equality." *Economics and Philosophy* 19: 1-44.

[16] Voorhoeve, Alex. 2015. "Introduction to the Symposium on Equality Versus Priority." *Economics and Philosophy* 31: 201-2.

The Reasons of Objective Consequentialism and Collective Action Problems

Susanne Mantel, Saarland University, Germany

Abstract

Objective consequentialism faces a challenge from action guidance. Consequentialists typically respond by drawing the distinction between a criterion of rightness and a decision criterion. Many objective consequentialists think that this challenge is thereby solved. However, if the challenge is reformulated as a challenge concerning the *normative reasons* which are implied by objective consequentialism, it might initially seem to be more troubling.
Normative reasons are typically thought to do both, determine rightness *and* serve as a decision criterion. A criterion of rightness could not amount to a normative reason, it is sometimes said, unless it can guide the agent in deliberation towards doing the right thing. However, there is a response available to this version of the challenge which is similar to the one given to the original version of the challenge from action guidance: We may use a distinction between right-making reasons and good deliberative reasons. Plausibly, deliberative constraints may hold for good deliberative reasons but need not hold for right-making reasons. Right-making reasons may thus consist in even those consequences of actions which cannot guide the agent in deliberation.
There remain various other problems for objective consequentialism, for instance the challenge from uncoordinated collective action. I discussed such a case which helpfully illuminates how right-making reasons might come apart from good deliberative reasons. There are different ways in which both right-making reasons and good deliberative reasons could be understood in this case. Thinking about similar problem cases with the distinction between different notions of normative reasons in mind might shed new light on the debate about objective consequentialism.

Introduction

It is a well-known objection to objective consequentialism that it does not provide sufficient help in guiding the agent in deliberation about what to do.

A version of this challenge might be expressed in terms of reasons: objective consequentialism is unable to come up with a satisfactory account of consequentialist practical reasons, since practical reasons must be capable of guiding us in reasoning about what to do. It might seem that this reasons-focused version of the challenge is more worrisome. Maybe the consequentialist can accept that ethical *theories* are merely standards of correctness and do not guide deliberation, but can anyone accept that the *reasons* which these theories are committed to do not figure in reasoning and deliberation?

However, just as consequentialists are often unimpressed by the challenge from action guidance, they need not be worried by the related challenge concerning reasons.

After arguing for this point, I will further examine my resulting view by applying it to an especially interesting problem case which will reveal different ways in which it could be developed. The case is a collective action problem which calls for mixed responses. The case highlights the distinction between right-making reasons and good deliberative reasons in an especially complex and interesting way.

I Consequentialism and Action Guidance

According to commonsense, right actions are right because they make the world a better place. This idea is captured by objective act consequentialism as follows: "Objective consequentialism is the view that the criterion of the rightness of an act or course of action is whether it in fact would most promote the good of those acts available to the agent." (Driver 2012, 98).

The "criterion of rightness" consists in the consequences which different courses of action would have. The problem is that at least some consequences are hard to predict, especially those in the far future. Agents are often unsure which of the actions available to them would most promote the good. Even if these consequences might make the action right or wrong they cannot serve as the considerations which guide our deliberation about what to do because we have limited epistemic access to them. Therefore, objective consequences provide the standard of rightness but often cannot be relied on as decision criteria in deliberation. This has been discussed as the problem of action guidance.

Objective consequentialists tend not to be troubled by this problem. They hold that there is an epistemic difficulty for agents, but that the consequentialist theory nevertheless gives true evaluations of which actions are right and wrong. Furthermore, the theory could in principle be supplemented with an account of considerations which are generally most helpful for guiding deliberation – even if these guiding considerations are distinct from the standard of rightness.

II The Challenge Concerning Reasons and Reasoning

If we apply the terminology of reasons to this problem, it might seem that the problem gets worse. It might be said that the standards of rightness constitute the *normative reasons*

that objective consequentialism posits, but that normative reasons, by definition, are guiding considerations for reasoning agents. So, objective consequentialism cannot introduce a distinction between standards of rightness and guiding considerations, it might seem.

Why think that objective consequences should be understood as normative reasons? What the agent ought to do is determined, according to objective consequentialism, by the totality of consequences, even if these are impossible to grasp. Since it is a truism that what the agent ought to do is determined by the totality of normative reasons, these consequences seem to play the role of normative reasons for objective consequentialism. It seems as if the normative reasons are constituted by objective consequences – or, at least, by something in the close neighborhood, maybe by present facts which determine which actions have which consequences. For instance, one might say that the fact *that the fish contains salmonella* is a normative reason not to eat it because if there are salmonella in the fish then the agent will suffer as a consequence of eating the fish. The normative reasons would then be facts which determine indirectly what an agent ought to do - by determining the consequences which more directly, in their role of standards of rightness, determine what the agent ought to do. But even if normative reasons might on this interpretation not be the consequences themselves but the present facts that determine the consequences, one would need to know how these present facts determine the consequences if one wanted these normative reasons to guide one's deliberation to the right action, and when one is unsure about the consequences of an action one does not possess this knowledge either.

So the normative reasons which determine which action ought to be performed, according to objective consequentialism, are either the consequences themselves or present facts in virtue of the role they play in determining the consequences. However, neither of these two candidates are helpful guides for deliberating agents when the consequences of an action are hard to predict. This is problematic, one might think, since reasons are for reasoning with.

Being a consideration which could be used to guide the agent in reasoning is often said to be the central mark of normative reasons (e.g., Kearns and Star 2008, 39). Depending on how this 'can' should be understood, it might be concluded from this that consequences which cannot be known – like non-obvious distant consequences – do not seem to bear the central mark of normative reasons. Nevertheless, according to objective consequentialism they are the normative reasons which determine what agents ought to do.

Similarly, many philosophers argue that there is a *deliberative constraint* on normative reasons, such that something cannot be a reason for an action unless the agent could perform that action for that reason (e.g., Kiesewetter 2017). But an agent cannot perform an action

for the reason that it has certain distant consequences if the agent cannot come to know these distant consequences at the time of action. So many consequences fail the constraint on normative reasons.

III A Distinction between Two Notions of Normative Reasons

Despite all this, I believe that introducing the notion of a normative reason does not make things worse for objective consequentialism. The distinction between a standard of rightness and a decision criterion which we find in objective consequentialism resembles a distinction which has recently been made between two notions of normative reasons, i.e., reasons that determine what is right or what ought to be done (e.g., Broome 2008) and reasons that, roughly, are premises in good reasoning (Setiya 2014, similarly Kearns and Star 2008).

In my terms, according to objective consequentialism consequences (or some facts in their neighborhood) are the *reasons which determine rightness (short: right-making reasons)*, but they need not be premises of good reasoning *(or good deliberative reasons).* Although we use one and the same term, i.e., "normative reasons", for the entities which are determinants of rightness and for the entities which guide deliberation, we may discover that there is an important difference between the entities which fulfill the first role and the entities which fulfill the second role. It is sometimes assumed that normative reasons play both these roles (e.g., Kearns and Star 2008, 39 and 49-51), but it may be doubted that this assumption can be defended, since these roles can come apart in various ways (Wedgwood 2015). When this distinction between two notions of normative reasons is made, it is not obvious that the deliberative constraint should be applied to right-making reasons, although it may be applied to good deliberative reasons. Objective consequentialists may simply say that all objective consequences (or all determinants of objective consequences) determine rightness and thus constitute right-making reasons, but that they need not meet constraints on reasoning like the one described by Kiesewetter, for these hold only for good deliberative reasons. If some right-making reasons like distant and non-obvious consequences do not constitute good deliberative reasons, the reasoning constraint does not seem to apply to them.

It might be objected that the claim that right-making reasons are consequences of actions implies that agents can *never* act for the right-making reasons when these right making reasons do not obtain prior to the action. How could they act for a reason which does not yet exist? But I think that agents can at least sometimes act for right-making reasons even if all

right making reasons are consequences of actions. Acting for a right-making reason does not presuppose that these reasons obtain prior to the action, but that the agent has some epistemic access to them prior to the action, even though they may obtain at a later time. Although some consequences are such that the agent has no epistemic access to them at the time of action, at least some consequences of actions are epistemically accessible. Sometimes agents are able to figure out many of the consequences of their actions (maybe by the use of indicators, as I argue in Mantel 2018). Under these circumstances, agents may act on their knowledge of the right-makers, even if these lie in the future. They may thus at least in some cases act for right-making reasons even if these are consequences.

IV Collective Action Problems

There are several problem cases for the view defended here. All I can do here is to sketch one interesting case involving collective actions and to discuss some of the questions and possibilities which it brings to mind, although a much more thorough treatment of this case and similar ones would be desirable which cannot be provided in this short paper.

In many collective action situations actual consequences are determined by many uncoordinated individual actions together. One class of interesting examples are cases which call for agents to act in different ways, as the following:[1]

> Going to Work
> Suppose that when people are going to work in a given city, they ought to use climate friendly means of transportation. Accordingly, it would be best if 60% used public transport and 40% used their bicycles – otherwise, either public transport or biking lanes would get crowded and would be slowed down or would eventually collapse. However, mobile communication just broke down and the agents are unable to coordinate who is to use which means of transport.

If consequences determine rightness, then we might think that the right-making or wrong-making reasons of each individual action are the consequences which each individual action has and compare them to the consequences which would obtain if the agent had not acted, presuming that everything else is held fixed.

[1] Compare, for example, Pinkert's (2014) 'The Concert Audience'. By contrast, most attention has so far been paid to scenarios where all agents intuitively ought to do the same thing, e.g., eat less meat. This is a challenge for consequentialism because it may nevertheless be thought that their individual actions made no difference to the harm that has been caused collectively (e.g. Kagan 2011).

Suppose everyone had used public transport and it broke down. According to the criterion just given, this would mean that everyone did the wrong thing, since for each agent it is true that it would have been better if *they* had used the bike while everyone else had still used public transport. But it might seem more plausible to say that 60% did the right thing, since 60% ought to use public transport, and only 40% ought to use bicycles (although, of course, there is no truth of the matter of who of those who used public transport belonged to the 60% who were right in doing so). After all, if everyone had used bikes, the biking lanes would have collapsed. It is counterintuitive that each individual should have used bikes if the consequences of everyone doing what they should would have been just as bad.[2]

When mixed behavior is called for but coordination is ruled out, the same decision situation applies to all agents, but best consequences are achieved only if a certain percentage of individuals acts differently than the others. The consequences of individual actions then depend on the combinations in which these individual actions stand. Therefore, one might suggest that objective consequentialism should be formulated not for individual actions (as it was in section I) but for combinations of actions. The primary bearers of rightness and wrongness might be combinations, where any combination which consists of 40% individual acts of taking the bike and 60% individual acts of using public transport is right.[3] This, however, means that there is no true answer to the question what a given individual should have done when all used public transport. According to this approach, there seem to be no right-making reasons for individual actions in a case such as this one. Instead, there would seem to be only right-making reasons for combinations of actions, such as a combination's efficiency in transportation.

What about good deliberative reasons, by contrast? Good deliberative reasons seem to apply primarily to individual actions, not to combinations of actions. Nevertheless, good deliberative reasons may guide agents towards those actions which are more likely to be parts of right combinations of actions. In the example, good deliberative reasons (under uncertainty of what the others will do) may thus favor using public transportation over using the bike, because there are more right combinations of actions in which the individual action is among the 60% which consist in taking public transport than right combinations of actions in which the individual action is among the 40% which consist of going by bike. The most

[2] This intuition is described by Portmore (2018) in the principle "moral harmony" and by Regan (1980) in his "adaptability".

[3] This view obviously raises the questions whether the rightness of a combination of actions requires joint agency or, at least, joint responsibility, which does not seem to be given without the possibility to coordinate (see e.g. Pinkert 2014, who draws on work by Virginia Held). However, maybe rightness and responsibility must be separated by a view along these lines. Good deliberative reasons might be closely related to responsibility even if right-making reasons were not.

troubling implication of this example seems to be that in cases like this one, although a mixed pattern of action would have the best consequences, the decision situation of all agents is stipulated to be identical, such that their good deliberative reasons all point towards the same action (taking public transport).

However, in principle it is possible to prevent this consequence. This is especially obvious if we keep in mind that good deliberative reasons are mere tools for decision making which need not be of any deep moral importance (by contrast to right-making reasons). Agents may simply invent new deliberative reasons when they create a lottery with a 60% chance of a blue ticket. They may then use the consideration that they drew a blue ticket as a good deliberative reason to use public transport, and the consideration that they did not draw a blue ticket as a good deliberative reason to use the bike.[4] Agents may know full well that the outcome of the lottery does not help them to predict the consequences of their action – especially not if other agents did not use a lottery as well. They may even hold that it is not morally prescribed to do what the lottery says. They need not believe that what they treat as a good deliberative reason is also a right-making reason with respect to their action, and may thus use the lottery ticket as a tool to guide them even if they don't ascribe any moral importance to the lottery.

V Conclusion

Objective consequentialism faces a challenge from action guidance. Consequentialists typically respond by drawing the distinction between a criterion of rightness and a decision criterion. Many objective consequentialists think that this challenge is thereby solved. However, if the challenge is reformulated as a challenge concerning the *normative reasons* which are implied by objective consequentialism, it might initially seem to be more troubling.

Normative reasons are typically thought to do both, determine rightness *and* serve as a decision criterion. A criterion of rightness could not amount to a normative reason, it is sometimes said, unless it can guide the agent in deliberation towards doing the right thing. However, there is a response available to this version of the challenge: We may distinguish between right-making reasons and good deliberative reasons. Plausibly, deliberative constraints hold for good deliberative reasons but need not hold for right-making reasons.

[4] This idea goes back to a manuscript by Kevin Baum and Eva Schmidt, the discussion of which has influenced many thoughts in this section. I owe Kevin and Eva my thanks.

Right-making reasons may thus consist in even those consequences of actions which cannot guide the agent in deliberation.

However, there remain various problems for objective consequentialism, for instance the challenge from uncoordinated collective action which helpfully illuminates how right-making reasons may come apart from good deliberative reasons. Thinking about similar problem cases with the distinction between different notions of normative reasons in mind therefore sheds new light on the debate about objective consequentialism.

References

[1] Broome, John. 2008. “Reply to Southwood, Kearns and Star, and Cullity.” *Ethics* 119 (1): 96-108.

[2] Driver, Julia. 2011. *Consequentialism*. New York: Routledge.

[3] Kagan, Shelly. 2011. “Do I Make a Difference?” *Philosophy & Public Affairs* 39 (2): 105-41.

[4] Kearns, Stephen, and Daniel Star. 2008. “Reasons: Explanation or Evidence?” *Ethics* 119: 31–56.

[5] Kiesewetter, Benjamin. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.

[6] Mantel, Susanne. 2018. *Determined by Reasons. A Competence Account of Acting for a Normative Reason*. New York: Routledge.

[7] Pinkert, Felix. 2014. “What We Together Can (Be Required to) Do.” *Midwest Studies in Philosophy* XXXVIII: 187-202.

[8] Portmore, Douglas. 2018. “Maximalism and Moral Harmony.” *Philosophy and Phenomenological Research* 96 (2): 381-41.

[9] Regan, Donald. 1980. *Utilitarianism and Co-operation*. New York: Oxford University Press.

[10] Schroeter, Laura, and Francois Schroeter. 2009. “Reasons as Right-Makers.” *Philosophical Explorations* 12 (3), 279-96.

[11] Setiya, Kieran. 2014. “What is a Reason to Act?” *Philosophical Studies* 167 (2): 221-35.

[12] Wedgwood, Ralph. 2015. “The Pitfalls of ‘Reasons’.” *Philosophical Issues* 25: 123-43.

A Shell Game Theory – Reconnect Mankind with Nature to Create Wealth

Vincent-Emmanuel Mathon, University of Rouen, France

Abstract

How to rethink financial transactions in order to create wealth by protecting nature? Wealth is the result of a transfer of utility (or of energy), conveyed through a currency. A currency requires an unalterable underlying asset, which is traditionally matter, like gold for instance. But with our modern economy becoming increasingly immaterial, the boundaries between matter and energy have been blurred. Utility must be re-assessed under that scope. Utility now depends simultaneously on two parameters – matter and energy – and should then be represented by a complex number, with both its real and imaginary components. Utility is a point on a complex plane, namely the energy/matter plane. Moreover, from now on, in our transaction, there is not just you and me around the table. There will be you, me, plus fictitious economic agents representing respectively nature, the effects of regulation, and people not necessarily known to us, but potentially influenced by our deal. During a transaction, all these agents will move on the energy / matter plane, transferring utility to each other, drawing their own convex space on that plane; all the convex planes from various transactions shall combine together to form a symbolic shell. This virtual energy / matter plane is then connected to reality through a blockchain, powered by a new type of hashing function, a one that would convert all the actions of the agents in the real world into energy and matter. That blockchain can be used as the unalterable underlying asset of a new crypto-currency a shell-type crypto-currency (for a blockchain is unalterable by definition). As this crypto-currency would involve agents that are not necessary human but that could stand for the interests of nature as well, it would make the very fact of protecting nature profitable.

I

This title is a play on words. I did not mean the "Theory of Shell Games" but the theory of games as applied to *shell*. The original purpose of this paper was to rethink financial transactions in order to create wealth *without* damaging nature and *by* protecting nature. It starts with a new enquiry on wealth – *where does it come from?* – leading to a new way of conceiving currency (and hence money), which, in turns, shapes a new type of economy where the very protection of nature can be profitable indeed. The initial step of this paper is to determine where wealth comes from. Wealth is the result of a transfer of utility. Shells (*seashells*) might have been the first means of exchanges, and the first way to transfer utility[1]. Symbolically speaking, a shell can indeed represent the safe repository in which I would

[1] See Wikipedia 2018, chapter "Currency": "Seashells have been used as a medium of exchange in various places, including many Indian Ocean and Pacific Ocean islands, also in North America, Africa and the Caribbean."

store the memories of all that I have been doing to / for the others and that I expect them to do for me in due time. All my actions are symbolically recorded in the shell. All the utility I have received from others and the utility I have been giving to others, is therefore symbolized by the shell. However, "*symbolically recorded*" is not sufficient in a complex society where corn can be traded for animals or artifacts. My actions, and more broadly, all my energy – for whenever human activity may be involved, it ends up with human energy – must be stored more effectively. My "safe repository" now becomes a stock of matter, for instance, gold. Gold has an ideal property: it is *unalterable*. Therefore, it can be a good asset to store the footprint of human energy. When I trade with you, I give you my energy and you give yours to me. This is the way I transfer my utility to you. This exchange of energy is printed onto matter, which is used as an intermediate; this is where the very concept of currency comes from. A currency uses unalterable matter – like gold – as an underlying asset.

Transfer of utility – which ultimately leads to wealth – has been conveyed through a currency fitted with an *unalterable* underlying asset. This can work if economy is essentially material, if all economic transactions end up in transforming matter. In other words, it works in a Newtonian world, where there is conservation of energy on one hand and, separately, conservation of matter on the other hand. It is then possible to create a direct and unique link between a given variation of energy and its corresponding variation of quantity of matter. In that case, utility is a one dimension entity – be it matter or energy – and can be represented by a real number. But with our modern economy becoming increasingly immaterial – *where wealth can be created through pure information not backed up by matter* –, the boundaries between matter and energy have been blurred. We then switch from a Newtonian world to an Einsteinian (*from Einstein's physics*) world where matter and energy are not separated anymore. Matter is potential energy and potential energy is matter; hence the famous formula $E=Mc^2$. It becomes therefore impossible to use matter as a fixed pattern to record the footprint of (human) energy because matter is potential energy and *vice versa*. This is why today currencies are "floating" and are deprived of any *unalterable* asset. This explains the collapse of the Bretton-Woods system in the early 1970s and the gap between the paces of financial and material exchanges, as conjectured by Tobin back in 1978 in his famous Essay *A Proposal for International Monetary Reform*.[2] The challenge would then be: how to create a new type of currency that would win back an unalterable

[2] See Tobin 1978: "The basic problems are these. Goods and labor move, in response to international price signals, much more sluggishly than fluid funds. Prices in goods and labor markets move much more sluggishly, in response to excess supply and demand, than the prices of financial assets, including exchange rates."

underlying asset in our Einsteinian world where energy and matter are not separated anymore? As a currency is basically what shall convey a transfer of utility, the very concept of *utility* must be re-assessed.

Utility *now* depends *simultaneously* on two parameters – matter and energy – and not on any of them *separately*, with potential energy being potential matter and vice versa. Philosophically speaking, this does make sense as the duet energy / potential matter and potential energy / matter may apply to the concept of pleasure as well. Pleasure can be actual or potential. I can feel it actually – as a *matter* of fact – or I can *anticipate* it. This concept of anticipation was developed by Epicurus[3], and more recently, it became the cornerstone of John Nash's bargaining theory (1950). This can also be compared to Aristotle's duet: "potential" (*Dynamis*) versus actual (*Energeia*).[4] Actual matter is potential energy and vice versa; actual pleasure versus anticipated – or potential – pleasure. An appropriate way to translate this potential / actual duet – be it through the energy / nature duet or through the concept of pleasure – is a complex number, with both its *real* and *imaginary* components. Utility is not to be modelized by a real number like in most micro economics literature – but by a complex one. Therefore utility is not a mere quantity but a point on a complex plane, defined by both the axes of real and imaginary components.

Another aspect of Einstein's theories is, that, contrary to Newton's physics, there are no *frames of reference*. In a Newtonian world, space is a three-dimension Euclidean vacuum in which motions of bodies come from their respective forces and can be traced from a *frame* of reference, with time considered as an independent value. As symbolically transferred in the economy, when I trade with you, or, in other words, when we transfer utility to each other, there is only you and me, and the frame of reference. Practically, this frame of reference can *simultaneously* be regulation, commodities, nature, environment, and *some* other people potentially influenced by the deal but not directly involved in it.

On the contrary, in an Einsteinian world, time is not an independent value. Space and time form a continuum. Material bodies do alter the shape of this continuum, thus creating motion. This is the law of gravity according to Einstein's theory. The very concept of *frame of reference* has no meaning anymore as even the metrics of our world can change. Instead of a frame of reference, there are only material bodies in motions. Which means that, for our financial transaction, our own frame of reference (regulation, commodities, nature, environment, other people potentially influenced by the deal) should now be replaced by an

[3] In Epicurean philosophy, pleasure is not only actual pleasure but also pleasure that is going to come or, even, to avert; hence the concept of *anticipation*.

[4] See Aristotle's physics in Aristote 1999.

array of material bodies, or, as translated in economics, of economic agents. Nature – commodities, natural resources – would be represented by economic agents. Regulation as well, or rather, the influx of regulation on our deal, should be modelized by an agent. Most of these agents would be fictitious as they would not correspond to an actual individual. But they would anyway, to some extent, refer to, or be linked to, some *material* body. Commodities, nature, and legal constraints are then represented by fictitious economic agents acting in their name. From now on, in our transaction, there is not just you and me around the table. There will be you, me, plus an agent acting in the name of some natural resources, of some animal species involved in our deal, plus an agent representing people potentially influenced by our deal, plus an agent standing for the effects of regulation on our deal, namely, regulation implemented.

All of this can be implemented through artificial intelligence which would modelize the behaviour of all those agents – fictitious or not –, in real time. Some of these agents are real and well *known* to us even though they need to be represented through mental images as they are fictitious agents (for instance, those representing nature). Some of them are *unknown* to us, yet they are influenced by our transaction and can potentially influence it. How can we discriminate agents *known* to us from agents *unknown* to us? My supplier is known to me, the supplier of my supplier may indirectly be known to me … the supplier of my supplier of my supplier … maybe not.

But does it matter to me if I know her? Maybe not. If knowing an agent better – i.e. having more information about her – enhances her utility in my transaction, I do consider this agent as *known* to me. Unknown agents do typically stand for those that are *not physically* around the table but that may though be influenced, or that may influence themselves our transaction. They do represent the "rest of the world". They should be represented by a fictitious agent acting in their name.

Our transaction is thus a *transfer of utility* involving *simultaneously* on one hand *"n"* actors (*including you and I*), either fictitious or real, that are *known* to us, i.e. that are trading directly with us, and, on the other hand, an *indefinite* number of actors unknown to us. A good method to model the way *"n"* actors do simultaneously transfer utility between each other is *Pagerank*, the algorithm powering Google's search engine, as described in Brin and Page's paper[5].

Pagerank assigns a rank to each web page. A rank is the equivalent of utility. When one page points to another one, the former transfers its rank – or part of it – to the latter. The

[5] Google's algorithm is being studied exclusively through the following paper: Brin and Page 1998.

very principle of Pagerank algorithm consists in modeling the transfer of utility between *n* given web pages whilst taking into account the "rest of the web".

The algorithm results in such an equation:

$R = AR + E$[6]

Where R is a *n x n* dimension matrix representing the pagerank – or utility – of our *n* given web pages, A is a *n x n* dimension matrix modelizing the losses in the transfer of utility, and E, a vector – with randomly chosen values – standing for the "rest of the web". A simultaneous transfer of utility between *n* given actors is thus – according to Pagerank – expressed through matrixes, i.e. linear applications.

As transcribed into our financial transaction, the R matrix would apply to the utilities of our *n* economic actors – including you and I – known to us, be they real or fictitious. The A matrix would feature the losses when those actors transfer utility to each other, just like for *Pagerank*. And, finally, the E vector would stand for the "rest of the world", i.e. all actors that are unknown to us and that yet either influence, or are influenced by, our transaction.

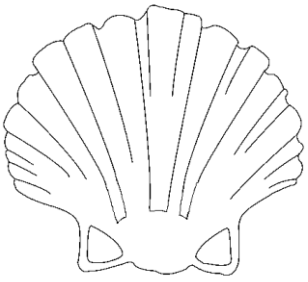
Being expressed through matrixes, thus through linear applications, utilities – by "utilities" I mean the *utility* of each and every agent involved in the transaction – are added up and multiplied each other simultaneously. Utility is here modelized by a complex number, not a real one. Hence, multiplying or adding up utilities means shifting them on the complex plane. Our *transfer of utility* is thus a simultaneous shift on a complex plane of our *n* actors known to us (processed through R matrix) and unknown actors (represented by vector E). They all tend to move towards the ideal location. The closer you get from that ideal location, the better, and the richer you are.

II The Ideal location

The ideal location can be determined using Nash bargaining solutions. All actors agree to have their respective utility functions displayed and they know that, after a finite number of iterations (namely, shifts on the plane), they shall reach an optimum point – the *ideal location* – where utility is maximized for all of them. Some of them may not have their own utility apparently maximized but if they divert from that point, the overall utility will diminish and, at the end of the day, they all shall lose.

[6] The actual equation from Page and Brin's paper is $R = c(AR + E)$ with $c<1$.

By applying Nash bargaining scheme, all actors will then define – through their various *shifts* – a convex space containing the optimum point inside. This convex space shall be labelled the *deal area* of our transaction. Each transaction will have its own *deal area* space. Such spaces are not totally isolated from each other as their respective E vectors, or source vectors, are fed by the transactions of others. There are combined, forming a shell–shape like this:[7]



Hence, the "shell returns" but not as a game conceived to deceive (a shell game) nor as a safe repository but as a shape symbol of all transactions combined with each other. Wealth is not expressed in terms of possession (of matter) but in terms of location. The better located you are (on that complex plane), the richer you become.

Still, that complex plane – which is basically a matter/ energy plane – on which all agents move remains fictitious at this stage. It must be connected to reality. The challenge is to throw all actual moves of economic agents in the real world – all their actions – onto this two-dimension energy/ matter plane. This could be done by assuming that all actions in the real world – be they abstract or concrete, intellectual or practical – result in changes in matter and energy. These changes could be processed through a function, let us call it *"function* $\mathscr{H}^{a}$". That function would admit a multi-dimension (or possibly infinite dimension) space – *to describe any action in the real world* – as input and the complex plane as output.

$$\mathscr{H}\mathbb{R}^{\infty} \rightarrow \mathbb{C}$$

Practically, the transfer of utility would then consist in converting all real actions into a complex number through the $\mathscr{H}$function. Transfer of utility, as implied by an Einsteinian world governed by the energy / matter duet, can then be implemented using this $\mathscr{H}$function. The

[7] https://www.hugolescargot.com/coloriages/3207-coquille-st-jacques/

goal is now to find a new type of currency, fitted with an unalterable underlying asset, able to convey such a transfer of utility.

A solution could consist in a crypto-currency based upon a blockchain. In a blockchain, information is dispatched and not alterable. *Dispatched* meaning information is stored on several occasions. *Unalterable* meaning, no matter where I find myself, the piece of information I see before me will never be altered. And that is because, any new piece of information has been specifically shaped into a new block (thanks to a complex algorithm); a block that is then ready to be inserted consistently into the already existing chain of blocks (namely, the *blockchain*). That complex algorithm is the *hashing function.*

The purpose of a hashing function is to make it very difficult to create a block and very easy to read it. The aim of a hashing function is thus to trigger complexity. In crypto-currencies like Bitcoin, complexity is artificial as it comes from an algorithm which is processed by computers, and which is hence time and energy consuming. A substantial amount of energy is thus removed from the *real* world – quite environmentally unfriendly – to create a *virtual* world of blockchains. A new type of hashing function should be imagined, a one that would, instead, dig its complexity not in any fixed pattern of algorithmic calculus but in nature itself. It would embrace the very complexity of nature, the way it evolves, including the behaviour of our economic agents (fictitious or real), human beings, animals and use that *natural* complexity to create blocks.

That hashing function would not be based on *artificial* complexity but on *natural* complexity. Instead of a fixed algorithm – like for current crypto-currencies – it would be a *liquid* algorithm embracing nature in all its components. It would admit the "real world" as input. Our above-mentioned $\mathscr{H}$function, with its infinite dimension space – i.e. the real world – as an input would be adequate.

$\mathscr{H}$would then become our *liquid* hashing function. $\mathscr{H}$would have a dual role: it would transform all the actions of agents into transfers of utility (on the complex energy/ matter plane) and, in so doing, it would also create blocks for the blockchain. And, this blockchain could be used as the underlying asset of a new currency, specifically shaped for our new type of economy where transfers of utility are moves on a complex plane; for, namely a *"shell-type economy".* Two challenges must though be overcome to implement this concept.

First, "everything" involved into our transaction – nature, regulation – must be turned into economic agents. It is practically possible by observing nature and using artificial intelligence but appropriate axiomatics are required. Second, our $\mathscr{H}$hashing function is still to be

found, and practically, all human and natural actions must be assessed in terms in changes of energy and matter. This can be easily done for natural phenomena by using sensors, it may be more difficult for intellectual operations. Still, with appropriate axiomatics this issue could be solved as well. Once those challenges overcome, once a shell-type economy and its associated crypto-currency implemented, two major pending issues could be solved: the Tobin issue and the profitability of the protection of nature.

First, the Tobin issue about de-correlation between financial and real exchanges would be solved as any actions in the real world would be automatically and mechanically transferred into the financial world, through our crypto-currency and its $\mathscr{H}$ashing function.

Second, as nature would *de facto* become a real economic agent, it would be inserted into the financial system, thus transferring wealth of its own through our energy / matter complex plane.

It would then become possible to earn money by protecting nature. That would create a genuine green economy, a one that would not be based on disguised subsidies but on actual financial gains from nature protection. Our new definition of utility – *a complex one based on the interactions between energy of matter* – thus implies a wide array of agents – either fictitious or real, some of them representing nature –, agents that will have all their actions symbolically traced as moves on a transaction plane. The very process – based on our $\mathscr{H}$ hashing function – through which their moves are traced on that plane also creates an underlying asset for a new type currency, powering a genuinely green economy, a "shell-type economy".

References

[1] Aristote. 1999. *La Physique*. Paris: Vrin.

[2] Brin S. and L. Page. 1998. *The PageRank Citation Ranking: Bringing Order to the Web*. Computer Science Department, Stanford University, Stanford, CA. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[3] Einstein A. 1956. *La relativité*, translated by Maurice Solovine, Paris: Editions Gauthier-Villars.

[4] Nash J. 1950. "The Bargaining Problem" *Econometrica* 18 (2): 155-62.

[5] Tobin J. 1978 (October 20). "A Proposal for International Monetary Reform." *Cowles Foundation for research in Economics Discussion Paper*, Yale University

[6] Wikipedia. 2018. *Seashell*, accessed July 1st, 2018. https://en.wikipedia.org/wiki/Seashell

Against Animal Replaceability: A Restriction on Consequences

Ricardo Miguel, University of Lisbon, Portugal

Abstract

Animal replaceability is supposed to be a feature of some consequentialist theories, like Utilitarianism. Roughly, an animal is replaceable if it is permissible to kill it because the disvalue thereby caused will be compensated by the value of a new animal's life. This is specially troubling since the conditions for such compensation seem easily attainable by improved forms of raising and killing animals. Thus, grounding a strong moral status of animals in such theories is somewhat compromised. As is, consequently, their position as an alternative to rights-based theories in animal ethics. Recognising this, some utilitarians tried to disassociate utilitarianism and replaceability. I will here add my voice to this project. However, instead of seeing the culprit in the usual suspects (hedonism, maximisation or the total view), I advance a new proposal. After identifying that the compensating value for a disvaluable action has to be its consequence, I present a restriction on consequences: consequences of sequences of actions cannot be consequences of the isolated actions in the sequences. Given this, the main argument is simple: killing an animal is permissible only if the the value of the new animal's life is a consequence of the killing; but this value is a consequence of a sequence of actions which involves the killing plus some additional actions; therefore, since, via the restriction, such value is not a consequence of the killing, it is irrelevant to its normative status. I then present two further motivations for the restriction: firstly, it prevents the value of conditional actions from trivially influencing the value of the actions on which they are conditional; secondly, it is useful – even if not a complete solution – to reply other objections to consequentialism: the accordion effect of action and the cluelessness problem. I finally consider a couple of objections.

I The Replaceability Argument

This is what I take to be the best available version of the replaceability argument (RA):[1]

(1) If killing animals[2] whose future life would have a positive value will lead to the creation of other animals which would not exist otherwise and whose lives will have at least the same value as the one lost with the killed animals, then such killing is permissible.
(2) Killing *this* animal exemplifies the antecedent of (1).

[1] See Miguel 2016 for a contrast between this and two other versions of the argument.

[2] I use 'animals' to abbreviate 'non-human animals'. Although the RA may apply to humans too, I choose to focus on animal replaceability mainly because of its greater practical importance.

(3) It is permissible to kill *this* animal.

A theory that implies this argument is not just incapable of a strong moral protection of animals' lives – like a right to life – but, in addition, allows killing them given conditions which seem rather easy to satisfy. Even if most animals presently raised for some purpose that implies killing them do not enjoy good lives and, as such, are not individuals which make (2) true, some present or improved forms of raising and killing animals may find support in the RA. For example, according to the RA, the conscientious small farmer who raises animals for food is killing them permissibly.

In general, moral theories that require promoting overall value above individual harms and benefits seem to imply the RA. As a standard example of such theories, Utilitarianism has been criticised for recognising an inadequate moral status of animals and, relatedly, for not grounding ethical vegetarianism. If this is sound, Utilitarianism falls behind competing views in animal ethics that tick those marks, like rights-based theories.

Nevertheless, I must note that some authors have defended non-standard utilitarian views that do not imply the RA, but I cannot assess their merits here. Instead, I will propose a novel way to cut the link between Utilitarianism and replaceability – one that restricts the notion of consequence and maintains the core properties of the standard theory.

II Necessary Conditions for the Compensating Value

Suppose that some value, *v*, is not a consequence of some action, ϕ. Then *v* does not determine, or contributes to determine, the normative status of ϕ. Therefore, when a value compensates a disvalue which is a consequence of ϕ, the former must be a consequence of ϕ too. This means that the success of the RA implies that the value of the new animal's life – the candidate to compensating value – must be a consequence of killing another animal – the action with a disvaluable consequence.[3] However, as I will argue, there are good reasons not to regard such value as a consequence of the killing. But before moving on to this, let me illustrate why the consequentialist is committed with this tight relation between the compensating value and the value it is to compensate.

[3] This action may have, and normally has, valuable consequences too. Throughout I use 'disvaluable action' just to mean the action with the relevant disvaluable consequence, which is compatible with it being, sometimes, permissible.

In November 2017 a lynx that had escaped from an animal park in Wales was shot dead. Imagine that right after this a new lynx was born at the park and that his life was at least as valuable as the killed lynx's future life would be if he had not been killed. Thus, the balance between the value of the new lynx's life and that lost with the killing would not be negative. Yet, the value of the new lynx's life does not compensate the killing in the required sense aimed at by the RA – it does not make it permissible. Why? Since the new lynx's existence is independent of the other lynx's death, whatever value his life has, it is not a consequence of the killing. Therefore, such value cannot determine the normative status of the killing. In addition, this example brings to light that, when the disvaluable action is independent of the alleged compensating value, its omission would have made things better.

Thus, besides the requisite of non-negative net value, we have two other necessary conditions for the compensating value: on the one hand, it has to be a consequence of the disvaluable action (consequentialism); on the other hand, the omission of the disvaluable action and the performance of that which leads to the compensating value has to be inaccessible to the agent (maximisation). In sum, utilitarian value compensation requires: (i) that an action ψ brings about a value at least as good as the one lost with a disvaluable action ϕ; (ii) that the value of ψ is a consequence of ϕ; and (iii) that performing ϕ and ψ maximises the good.

To my knowledge, everyone discussing this matter has been accepting that the RA satisfies (ii).[4] I think that this is wrong and will argue for a restriction according to which (ii) fails.

III A Restriction on Consequences

III.a Blocking the Replaceability Argument

Consider the following restriction on consequences:

(R) Consequences of sequences of actions cannot be consequences of the isolated actions in the sequences.[5]

[4] To name a few, see Singer (2011), Regan (2004), Višak (2013, 2016), Chappell (2015) and Delon (2016).

[5] I am shamelessly applying to my needs Diogo Santos' "Non-disaggregation Principle" (ms.), which he uses to deal with the cluelessness problem (see the end of III.b). After reading Bratman (2006) on the connections between the accordion effect and Hart and Honoré's (1959) Voluntary Intervention Principle, I realised that (R) also has some connections with that principle, but I cannot explore them here.

I will now show that (R) blocks the RA. Recall premiss (1): If killing animals whose future life would have a positive value will *lead to* the creation of other animals which would not exist otherwise and whose lives will have at least the same value as the one lost with the killed animals, then such killing is permissible. For this to be true and, as I have argued, faithful to Utilitarianism, 'lead to' must relate the killing with its consequences. However, killing animals, *by itself*, does not "lead to the creation of other animals which …" Some additional actions are required, like making animals reproduce, taking good care of the newborn, and so on. Thus, the plausible sense in which killing animals leads to such and such is by being one action among a sequence of actions which has that consequence. Yet, in this sense, (R) tells us that the value of the new animal's life is not a consequence of the killing (nor of the other isolated actions). According to consequentialism, then, the new animal's life is irrelevant to the normative status of the killing. Therefore, (1) is false, for although the killing leads (in the specified sense) to the valuable state of affairs, this has no bearing on its permissibility.[6]

An obvious question now arises: why should a consequentialist accept (R)? Well, if one cares about stopping the RA, then this already counts in its favour. But of course that this alone will seem rather *ad hoc*. Moreover, without any further support, (R) is also too strong a claim just to deal with a problem for utilitarians concerned with the ethics of killing animals. Nevertheless, I believe that we can say more in favour of (R).

III.b Two Further Motivations

Firstly, without a restriction like (R) the consequentialist allows the value of conditional actions to trivially influence the value of actions on which they are conditional. And I think that this is untenable. Consider an example of *value sabotage*. You did an intuitively permissible action like saving a person's life. Now suppose that someone killed another person on the condition that your saving was successful. Then your saving may not be permissible after all, for its normative status depends on the overall value of those two actions. It is odd that the value of an action depends not just on the things that it brings about, but also on the things *chosen* to be brought about by it (*mutatis mutandis* for *value improvement*, where the conditional action allegedly improves the condition action). To be clear, in these

[6] Were the argument stated with the consequence relation, (R) implies that the antecedent of (1) is false, making premiss (2) false. Interestingly, Persson (2017, 78-9) agrees that raising good lives cannot compensate killing good lives, "for while the latter could be done by means of a single act, the former cannot." But Persson leaves unclear why performing various acts cannot compensate a single one. My proposal is a step to explain this.

cases, conditionalising is itself a result of agency. Therefore, contrarily to non-agential conditional events, the conditional action can occur without the condition action.

This way of influencing the normative status of actions is too trivial to be acceptable. Even if every non-agential consequences of one's action would be good/bad, conditional actions could always overturn the balance. So, unless one is prepared to abandon a view of agency as being tightly connected to individual responsibility, consequences of other actions should not be treated like an action's non-agential consequences.

Consequentialists, then, can make a relevant distinction between consequences of sequences of actions and consequence-related events tracing back to a single action.[7] (R) does just this by preventing that consequences of sequences of actions bear on the normative status of the isolated actions. Still, (R) does not depart from the basic idea that, to evaluate an action, consequences are all that matter. In this way, although (R) restricts the received view about what counts as consequences of an action, we remain on consequentialist ground.[8]

Secondly, (R) has other useful applications for consequentialists. I will point out two.[9] The first regards the so called "accordion effect" of action. In brief, the worry is that the same set of events can be appropriately described in various ways that are such that the action in one description contains some of its consequences in another description. Adapting an example from Miller (1987), consider these two descriptions of what Jones did:

(a) Jones tells a lie.

(b) Jones saves a life.

If (a) and (b) are correct descriptions of Jones' action, then, assuming that the relevant value is in (b), consequentialists can only account for the normative status of the action via description (a); in contrast, non-consequentialists will care if (b) follows (or not) some rule. If

[7] E.g. pushing a person on the street is not permissible because someone decides to benefit that person if you push her; however, it would be if, say, by pushing her, a bullet happens to miss her. In the latter case, but not in the former, the valuable consequence is a result of your action alone. Note also that all I said is compatible with both single or multiple agent sequences of actions.

[8] Smart (1956) distinguished "extreme" and "restricted" Utilitarianism by, respectively, having a focus on single actions or on classes of actions. My suggestion is similar, but I am distinguishing single actions from sequences thereof and disregarding their being subsumed under a rule. Thus, in Smart's sense, Utilitarianism with (R) is still extreme.

[9] The aim here is just to motivate (R)'s acceptance beyond the RA and not to exhaust its usefulness. But I also envisage other applications, e.g. to a more commonsensical consequentialist account of blameworthiness.

this is sound, then, as Oldenquist (1966, 183) puts the problem, "whether we appeal to rules or to consequences to determine the rightness or wrongness of a particular action is of no moral significance." That is, the notion of consequence is left without distinctive normative relevance.

Given (R), however, the accordion can only be stretched so much: although we can agree that, say,

(*) Jones deceives the intending murderer

is also a correct description of what he did, we cannot say the same of description (b). The reason is that (b), but not (*), forces us to recognise multiple actions – whether or not a life is saved also depends on the intending murderer's action. Thus, (R) prevents the accordion from stretching beyond descriptions involving single actions.

Finally, another useful application of (R) pertains the cluelessness objection (Lenman 2000). In brief, the objection is the following: since the consequences of our actions are normally spread in time and space in a way that surpasses our knowledge, then we have no clue about what we ought to do. What seems to be a perfectly permissible action, like sparing the life of a pregnant woman, might actually be impermissible because such action happens to have the consequence of not preventing the birth of a future terrible dictator and all his atrocities.

Again, with (R) at disposal, the consequentialist has a line of response: consequences of the dictator's actions are not consequences of sparing his ancestor. We remain clueless about the consequences of sequences of actions that contain our actions as parts. Yet, given that such consequences are not consequences of our actions alone, we are not required to know them (we could not). And since they have no bearing on the normative status of our actions, ignoring them does not imply that we are in the dark about what we alone ought to do.[10]

[10] There is a reply if we can be clueless even if there are no sequences of actions involved. But the burden of proof is with those who think that single actions can have massive causal ramifications and that most of our actions are like that. But note that the claim here is modest: if (R) can mitigate this problem (as well as the accordion effect), then its acceptability goes beyond its stopping of the RA.

IV Objections

One tempting objection to my way of blocking the RA is that, somehow, we can automate the sequence of actions that together lead to the new animal's valuable life. In this way, it seems that there would be a single action, e.g. the press of a button, that leads to the killing of one animal and to the raising of another satisfying the relevant conditions. Thus, the value of what would otherwise be a sequence of actions is, in the automation case, the value of a single action. Since this, apparently, would not involve a sequence of actions, (R) would not apply and, therefore, it seems that the killing would be permissible (given that the press of the button would).[11]

This objection fails because it overlooks one crucial action (or sequence), namely, setting up the automation, making it seem that (R) would not apply when in fact it does. Hence, the valuable state of affairs would still be a consequence of a sequence of actions.

Perhaps one serious objection is that the RA can be restated in a way that bypasses (R). One might say that it does not matter whether or not the valuable state of affairs is a consequence of the killing, for as long as the whole sequence brings about such state of affairs, then, replacing an animal, that is, the whole sequence, is permissible. In other words, we shift the evaluation focus from actions to sequences of actions. And since I do not deny that the valuable state of affairs is a consequence of the sequence, then it seems that I have to agree that it determines (or contributes to) the normative status of the sequence.

But is this an objection to my proposal? The goal was to argue that, contrarily to widespread agreement, standard act-Utilitarianism does not imply the RA. After all, this was the target of those who used the RA against Utilitarianism (e.g. Pluhar 1982; Regan 2004). To achieve that goal I proposed a novel way, via (R), to stop the RA. But I did not claim that every utilitarian view with (R) stops the RA. It may well be the case that a global utilitarian view, that is, one which allows every sort of thing as evaluative focus, implies the RA. At the very least, the objector has to argue that a utilitarian should accept sequences of actions as evaluative focus. This comes with difficulties.

The said shift of evaluative focus requires completing and making sense of the new, reformulated principle:

[11] I had thought of this objection before, but I thank Melinda Roberts for mentioning it to me and thereby confirming my intuition that it was something I had to address.

> (C*) A sequence of actions is permissible iff it brings about more value than any other alternative _____ available to the agent.

The natural move is to fill the blank with 'sequence', but do agents have alternative sequences of actions to choose from? Maybe just in single agent sequences, for an agent cannot choose a sequence that involves other people's actions (otherwise he would know how others would act).[12] And while single agent sequences are enough to formulate the RA (but seriously limiting its application), we would still need a systematic account of the normative relation between sequences and the actions composing them. Without such account, that the consequences of a sequence are good overall is not enough for its permissibility, since it may be the case that a single impermissible act stains the sequence of which it is part.

V Conclusion

The value of the new animal's life should be a consequence of killing another animal if the RA is to be successful. Yet, I argued that such value, given the restriction on consequences I presented, is not a consequence of the killing. Therefore, the first premiss of the RA is false. Since that restriction is quite strong and, apparently, *ad hoc*, I offered two distinct motivations for it: one axiological and one of usefulness. I then considered and replied two plausible objections, the last of which hints at further work on coordinated actions and on the normative relation between sequences and the actions composing them.

I should conclude by stressing that even though I could not assess here the relative merits of others ways to disassociate Utilitarianism and replaceability, my proposal does not give up of any of the usual suspects like those other ways do – hedonism, maximisation or the total view. And while I am sure that other objections might be raised, I think that this utilitarian proposal against the RA is worthy of being discussed in more detail.[13]

[12] What about coordinated actions? Here seems possible to choose a sequence involving other people's actions because everyone agreed to act in such and such manner and so the agent seems reasonably informed in a way that does not preclude the sequence from being an alternative action. I have no answer to this.

[13] Meanwhile, following comments from Theron Pummer, Bruno Jacinto, José Mestre and Pedro Galvão, to all of whom I am thankful, I became aware of other difficulties, and also possible developments, of the view presented here. I hope I can address them in the future.

Acknowledgement

I am grateful to Diogo Santos for many discussions about the proposal advanced here. I thank Eze Paez, Tomi Francis and Simon Rosenqvist for their helpful questions at the ISUS 2018 conference. This work was done with the support of FCT studentship SFRH/BD/107907/2015, cofinanced by POCH/FSE and MCTES.

References

[1] Bratman, Michael. 2006. “What is the Accordion Effect?” *The Journal of Ethics* 10 (1-2): 5-19. DOI: 10.1007/s10892-005-4589-3.

[2] Chappell, Richard. 2015. “Value Receptacles.” *Noûs* 49 (2): 322-32. DOI: 10.1111/nous.12023.

[3] Delon, Nicolas. 2016. “The Replaceability Argument in the Ethics of Animal Husbandry.” In *Encyclopedia of Food and Agricultural Ethics*, edited by Paul Thompson, and David Kaplan. Dordrecht: Springer. DOI: 10.1007/978-94-007-6167-4_512-1.

[4] Hart, H. L. A., and A. M. Honoré. 1967. *Causation in the Law*. Oxford: Oxford University Press.

[5] Lenman, James. 2000. “Consequentialism and Cluelessness.” *Philosophy & Public Affairs* 29 (4): 342-70. DOI: 10.1111/j.1088-4963.2000.00342.x.

[6] Miguel, Ricardo. 2016. “What is the problem of replaceability?” In *Food futures: ethics, science and culture*, edited by I. Anna S. Olsson, Sofia M. Araújo, and M. Fátima Vieira, 52-8. Wageningen: Wageningen Academic Publishers. DOI: 10.3920/978-90-8686-834-6_6.

[7] Miller, Arthur. 1987. “Acts and Consequences: Squeezing the Accordion.” *Metaphilosophy* 18 (3-4): 200-7. DOI: 10.1111/j.1467-9973.1987.tb00853.x.

[8] Oldenquist, Andrew. 1966. “Rules and Consequences.” *Mind* 75: 180-92. DOI: 10.1093/mind/LXXV.298.180.

[9] Persson, Ingmar. 2017. *Inclusive Ethics: Extending Beneficence and Egalitarian Justice*. Oxford: Oxford University Press.

[10] Pluhar, Evelyn. 1982. “On Replaceability.” *Ethics and Animals* 3 (4): 96-105.

[11] Regan, Tom. 1983/2004. *The Case for Animal Rights*. Berkeley: University of California Press.

[12] Singer, Peter. 1979/2011. *Practical Ethics*. 3rd edition. Cambridge: Cambridge University Press.

[13] Smart, J. J. C. 1956. "Extreme and Restricted Utilitarianism." *The Philosophical Quarterly* 6 (25): 344-54.

[14] Višak, Tatjana. 2013. *Killing happy animals: explorations in utilitarian ethics*. Basingstoke/Hampshire: Palgrave Macmillan.

[15] _____. 2016. "Do Utilitarians Need to Accept the Replaceability Argument?" In *The Ethics of Killing Animals*, edited by Idem, and Robert Garner, 117-35. Oxford: Oxford University Press.

What Exactly Is Wrong with Human Extinction?

Tim Mulgan, University of Auckland, New Zealand

Abstract

In this paper, I explore the impact of risks of human extinction on the debate between consequentialism and contractualism. To get clear intuitions, I consider an imaginary case where scientists discover that deadly cosmic rays will hit the Earth in two hundred years, instantly and painlessly killing all living things and rendering the Earth uninhabitable. We can only avoid total human extinction by constructing interstellar 'generation starships' where a small population and their descendants will continue the human story in space in the hope that *their* distant descendants will re-establish human civilization on some distant exoplanet.

This is a fruitful thought experiment for moral philosophers, because competing moral theories that often go together in real-life come apart very radically, thus forcing us to choose between them. I contrast two moral theories: Scanlonian Contractualism and Rule Utilitarianism. These represent two broad approaches to moral theory: Consequentialism and non-Consequentialism. I argue that even the most plausible forms of Scanlonian Contractualism and Rule Utilitarianism find it difficult to make sense of our choice in Cosmic Rays. I also conclude that thinking about extinction puts pressure on Derek Parfit's recent argument that Contractualism and Consequentialism can be reconciled, because even the the most moderate Rule Utilitarianism gives much greater importance to extinction risks than even the most future-oriented Contractualism.

Introduction

Most people agree that human extinction would be bad. But moral theories disagree about *why* it would be bad and *how* bad it would be. These differences don't emerge in implausible tales where one option leads to certain extinction. But they come to the fore in more mundane cases involving small risks of extinction.

I The Generation Starship Tale

To get clear intuitions, I consider one imaginary case:

Cosmic Rays: Scientists discover that deadly cosmic rays will hit the Earth in two hundred years, instantly and painlessly killing all living things and rendering the Earth uninhabitable. We cannot prevent this. But we can avoid total human extinction by constructing interstellar 'generation starships' where a small population and their descendants will continue the

human story in space in the hope that *their* distant descendants will re-establish human civilization on some distant exoplanet.

We have only two options.

1. *Remain*: Accept imminent extinction, and try to make the lives of the last people as pleasant and worthwhile as possible. When the cosmic rays hit, humanity becomes extinct.

2. *Starship*: Invest heavily in starships. The cosmic rays still eliminate life on Earth, but no longer bring human *extinction*. Unfortunately, to have any realistic chance of success, this option must be *very* expensive – greatly depleting non-renewable resources, causing enormous environmental damage, and making life much less pleasant for those remaining on Earth.

My Cosmic Rays tale has the following ethically salient features:

1. The Starship option *adds future people*. Remain is clearly better for all actual or necessary present or future people (i.e., everyone who will exist whatever we do), but Starship *maximises* expected future well-being.[1]

2. The Starship option involves deliberately creating a very limited quality of life for some future people – namely, those living on starships – in the hope that much later future people – namely, those living in a flourishing human society on our target exoplanet – will enjoy much better lives. The starship people are thus effectively used as a means to an end. The deficiencies of starship life include:

    a. Lower quality of life.

    b. Restricted liberty, freedom, or leisure time.

    c. Potential moral tragedy. (Apart from tragic choices onboard a starship, when resources are insufficient to meet everyone's basic needs, our target exoplanet may already contain indigenous life that cannot coexist with humans.

[1] I assume here that, while it is very harsh, life in the post-starship future is still (on average, on balance) worth living. This optimistic assumption is controversial, but it is very common in the consequentialist literature on human extinction (e.g., Beckstead 2013; Kaczmarek 2017). A full treatment of starship cases would need to accommodate more pessimistic alternatives – especially the staple fictional situation where some future catastrophe leads to a fallen starship community whose members live in Stone Age poverty without any awareness that they are living on a spaceship (e.g., Aldiss 1958; Heinlein 1963).

The last starship generation must then choose between voluntary human extinction and the destruction of an entire extra-terrestrial ecosystem.)

3. The original choice is iterated. Each Starship generation faces the same two options: Remain and Continuation.

4. Procreative ethics is especially salient, for several reasons.

    a. Starship volunteers impose starship life *on their own descendants.* Is this a permissible exercise of their procreative freedom?

    b. Long-term starship survival demands very tight population control. Procreative freedom will be virtually unknown.

    c. Bearing and raising children in outer space are hazardous activities. (As is growing up in outer space.) Does anyone have a right to impose such hazards on others?

    d. Because the burdens of bearing children in space are borne by women, Starship life thus reintroduces gendered hazards, risks, and obligations which modern affluent liberal societies had hoped to consign to the past.

Why we should talk about such far-fetched examples? Here are some reasons.

1. Similar thought experiments are explored in detail in speculative fiction.[2] So we have many existing imaginative resources to draw on.

2. This case is not necessarily entirely fictional. Generation starships are one future possibility. (At least, some influential and wealthy people *think* this is a future possibility.)

3. Even if it isn't possible, this imaginary case shares salient features with mundane futures that are definitely possible. Other extinction avoidance strategies – colonising the solar system, enduring a temporarily broken terrestrial future, escaping to virtual worlds, uploading ourselves into computers, colonising the galaxy

[2] A good overview of the Generation Starship sub-genre is Caroti 2011. Influential and/or philosophically interesting contributions include Aldiss 1958, Heinlein 1941/1963, Ballard 1962, Delaney 1965, Wolfe 1993-1996, LeGuin 2002, MacLeod 2005, Bear 2007-2010, Robinson 2015.

with inanimate machines, or even making very significant sacrifices now to prevent truly catastrophic future climate change – are less dramatic than generation starships. But they raise very similar ethical issues.

4. This is a good case for moral philosophers, because competing moral theories that often go together in real-life come apart very radically, thus forcing us to choose between them. This is my focus today.

I contrast two moral theories: Scanlonian Contractualism and Rule Utilitarianism, which represent broader approaches: Consequentialism and non-Consequentialism.

I assume that Contractualism can accommodate non-identity, obligations to future people, and the imposition of risk (Kumar 2003, 2015; Weinberg 2015; Frick 2015). Instead of fixating on what actually happens to the worst-off actual person, our Contractualist asks whether some representative future person could object to our present behaviour. For instance, by storing nuclear waste negligently where it might leak radiation in a thousand year's time, I display an objectionable lack of respect for any future person who might suffer as a result.

I also assume that Rule Utilitarianism avoids collapse into Act Utilitarianism, and provides a plausible, moderate, liberal alternative within the utilitarian tradition.

These are, I believe, the most plausible forms of Contractualism and Consequentialism. They are also the points where the two traditions are closest together, and therefore provide the most interesting contrast. My question is whether either theory can make sense of our choice in Cosmic Rays.

II Contractualism

For the Scanlonian Contractualist, an act is wrong if and only if it is forbidden by a set of principles that no one can reasonably reject as a basis for our common life together (Scanlon 1999). Contractualism balances the complaints of representative individuals. In Cosmic Rays, there are several possible salient groups:

- *Present Decider*s: These people make the initial choice between Starship and Remain. They want to keep both options open. So they want to reject any principle that prohibits either Starship or Remain.

- *People who remain on Earth between now and when the cosmic rays hit*: These people much prefer Remain, because the Starship option makes their lives go much worse in clearly non-trivial ways. So they want to reject any principle that permits Starship.

- *Future people living on Starships*: These people only exist at all if we choose the Starship option. So they could not possibly object to Remain. (If you never exist, you have no complaint!) However, if their lives are sufficiently bad, or if they otherwise object to our using them as a means to ensure human survival, they might well reject any principle that permits Starship.

- *Future people flourishing in some distant future exoplanet civilization*: These people would obviously prefer Starship. But, as they only ever exist if that option is chosen, they cannot reasonably reject any principle that permits or requires Remain.

- *Sentient non-human animals who live on Earth between now and when the cosmic rays hit*: If their voices are heard, these animals will favour a version of Remain where (at least some of) the resources that would otherwise be consumed by the Starship programme are devoted to their welfare.

- *Sentient non-human extra-terrestrial beings (who may be persons)*: As the Starship option may impact very negatively on any extra-terrestrial life we encounter, extra-terrestrial sentient or rational beings will want to reject any principle that permits that option.

Even setting aside the possible complaints of future starship people – let alone those of non-humans on Earth or elsewhere – the complaints of the people left behind on Earth alone seem to clearly outweigh those of the present deciders. Surely our desire to preserve our own freedom of choice cannot outweigh their desire not to endure a very impoverished life?

It seems that Contractualism could not *require* the Starship option. At most, it might *permit* it. Unless it is extremely demanding, Contractualism presumably permits *some* sub-optimal projects. But would that be enough? Is human-extinction-avoidance just another morally optional sub-optimal personal project that is permitted but cannot ever be recommended, let alone required?

Here are some strategies Contractualists might use to support the Starship option:

1. *Only present people*: We only consider the complaints and perspectives of *present* people who already exist. (We must then argue that the starship programme can be delivered in a way that *no* present person could reasonably reject – perhaps by giving all present people some stake in the project.)

2. *Only necessary people*: We only consider the complaints and perspectives of present people *and those future people who will exist whatever we do (i.e., necessary future people).* We then argue that any future people who suffer on Earth because of Starship would *not* have existed otherwise *and therefore have no right to complain.* (Perhaps, like Parfit's Risky Policy (Parfit 1984), Starship involves widespread social upheaval that is identity-determining for *all* future people.)

3. *Compensation and meaning*: We argue that the very fact that their existence is necessary to avoid human extinction adds *meaning* to the lives of both those who suffer on Earth and those who endure restricted lives on starships. And this extra meaning outweighs the undesirability of their lives. (We might find this move more plausible for starship people, who are essential to the Starship plan, rather than for people left on Earth who are merely collateral damage.)

4. *Contingent people*: We allow contingent future people to reject principles that prevent them from coming into existence. (We might appeal to Nils Holtug's (2010) suggestion that bringing someone into existence *benefits* that person.)

5. *Impersonal values*: We strengthen present people's reasons for rejection by allowing appeals to *impersonal values* – either because impersonal values are generally admissible, or because there is something special about the value of avoiding human extinction. The basis of such appeals could be either impersonal value simpliciter (as Parfit does in *On What Matters*) or the importance *for the agent herself* of responding to impersonal values (as Scanlon does himself).

6. *Pluralism*: Contractualism only captures one part of morality. Other moral reasons may compete with 'what we owe to each other'. Extinction-avoidance might enter our overall theory alongside other non-Contractualist obligations to the environment or to non-human animals.

7. *Precondition*: The continued existence of human society is a precondition for the applicability of Contractualist morality. Perhaps we should therefore introduce a hitherto neglected background obligation to do *whatever we can to avoid imminent human extinction or avoid civilizational collapse*.

III Consequentialism

Consequentialists object that Contractualism misses the worst thing about human extinction: the *loss* of all that future human happiness, the *absence* of all those happy future people. The challenge for Consequentialism is to avoid the *dominance* argument, which concludes that the Starship option is always obligatory (e.g., Beckstead 2013). Any negative impact over the next two centuries is dwarfed by the potential *loss of billions of extra happy lives*. The far distant future must dominate our present ethical thinking. Any reduction in extinction risk justifies any present cost.

The basic Dominance Argument is simple:

1. If we avoid imminent human extinction, then humanity could continue for billions of years.

2. The expected value of possible futures where humanity continues for billions of years is astronomically large.

3. Therefore, the expected value of *any reduction in the probability of* imminent human extinction is also astronomically large.

4. Therefore, any reduction in the probability of imminent human extinction outweighs any present or near future cost.

Rule Utilitarians argue that the right thing to do in any situation is the act that follows from the ideal moral code – the code whose widespread acceptance would have the best consequences relative to other possible moral codes (Hooker 2000; Mulgan 2006, 2015, 2017). Prima facie, the dominance argument applies to rule utilitarianism as much as to act consequentialism (Kaczmarek 2017).

Rule Utilitarians have several strategies to weaken the dominance argument.

1. *Uncertainty about long-term survival*. Avoiding imminent human extinction only dominates if it raises the probability that humanity will survive for *billions* of years. But how much faith should we put in any future prediction whose conclusion so far outstrips any possible evidence base? What if the probability that humanity will survive for billions of years (even if we avoid imminent extinction) is itself *astronomically small*?

2. *Uncertainty about far distant future well-being*. Avoiding imminent human extinction is only good if (most) far distant future lives are worth living. But why think that? Are most *present* lives worth living? Have most *past* human lives been worth living? If not, why assume the future will be better? Can we reasonably project current levels of happiness or current (upward) trends indefinitely into the future?

3. *Rejecting Total Utilitarianism*. The standard dominance argument assumes total utilitarianism. Most people reject total utilitarianism. Some alternatives (such as Larry Temkin's 2015 suggestion that what matters is the number of *times* that are inhabited by happy humans) seem to give extinction avoidance an even higher priority. But others – notably average utilitarianism, limited quantity views, or diminishing marginal value views – may counsel against extreme extinction avoidance measures such as Starship.

4. *Rejecting anthropocentrism*. The Starship option poses an unknown threat to undiscovered extra-terrestrial life. If (some of) that life is sentient, then the potential loss of extra-terrestrial wellbeing may outweigh any value added by ensuring that there are some happy humans – especially if we attach diminishing marginal value to each species or kind of life.

5. *Priority to the Worst-off*. If we give priority to the fate of the worst-off future people, then the costs imposed on those who suffer on Earth or on starships may outweigh the (possible) benefits enjoyed by distant future people who might flourish on a colonised exoplanet.

6. *Rejecting expected value maximisation*. If we give priority to avoiding catastrophic outcomes, then the Starship option may be ruled out by the many ways it could go very badly wrong – which have long been *the* central theme in generation starship fiction.

7. *Limits on demandingness*. Rule Utilitarians argue that their theory is less demanding than Act Utilitarianism, because (a) human beings cannot internalise an overly demanding moral code, and (b) once we factor-in the costs of compliance borne by every successive generation, such a code won't maximise long-term well-being anyway. If these argument carry over to the special case of human extinction (where one possible future has no inhabitants in the far distant future), then they suggest that Rule Utilitarianism will not make extreme demands regarding extinction avoidance.

8. *Consequences in other possible futures*: Even if the Starship option would produce the best consequences in this particular case, it doesn't automatically follow that someone who had internalised the optimific moral outlook would feel free to choose this option. When rule utilitarians select an ideal code of rules, they must assess it against a wide range of possible futures. Would a willingness to abandon a ruined Earth and unleash our not-entirely-admirable-consumer-society on an undeserving galaxy have negative effects in other (more plausible) scenarios? (In the real world: Does the fantasy of escaping the consequences of their own environmental destruction insulate the super-rich from recognising the need to mend their ways?)

9. *Pluralism*: We might acknowledge that Consequentialism only captures one part of morality, and that other moral reasons may compete with our reasons to promote the good (whether individually or collectively). Consequentialist reasons to avoid human extinction might compete with – or are constrained by – a Contractualist story about what we owe to each other.

IV Beyond (or between) Contractualism and Utilitarianism

Intuitively, Contractualism seems to give too little weight to human extinction and Utilitarianism too much. It is hard to accept that avoiding human extinction is merely (at best!) one optional but sup-optimal project among many. But it is equally hard to accept that threats of extinction should always and completely dominate our ethical thinking. I have presented several ways for Contractualists to take extinction more seriously, and for Utilitarians to take extinction less seriously. Perhaps one or other of these succeeds. Perhaps the two theories can – as Parfit hoped – meet again in the middle. If not, we must look elsewhere.

One option is to introduce a new normative source – attaching direct importance to the continuation of humanity itself and/or the avoidance of extinction (e.g., Frick 2017). I think this is a mistake. Consider two puzzle cases:

Leave or Remain: Humanity faces an existential threat. We have only two options. We can either (a) maximise the quality of life on Earth for a short time, or (b) pour all our resources into a very Spartan short-lived generation starship programme. If we *Remain*, then a hundred million people will enjoy good lives for one century, and then humanity becomes extinct. If we *Leave*, then during each of a thousand centuries, one hundred people will enjoy

less good lives, and then humanity becomes extinct. This is a *Partial Same People Choice*. Everyone who exists under Leave would also have existed under Remain. (Perhaps Leave involves cryogenic storage in outer space, with a hundred people thawed each century.) If we Leave, then *every individual* is worse-off than *she* would otherwise have been, and there are many fewer happy people, but *Humanity* will endure for an extra nine-hundred-and-ninety-nine centuries. Could Leave possibly be preferable?

Multiple Escape: At t1, the human community embarks on an ambitious multi-pronged plan to avoid human extinction: generation starships are dispatched across the galaxy, human minds are uploaded into virtual worlds stored safely in the asteroid belt, the rest of the solar system is colonised, and so on. Once established, these human colonies cannot interact with one another or with the people left on Earth. At t2, we *are* those people left on Earth. An approaching meteor poses a (very) small risk to *all human life on Earth*. How seriously should we take this threat? What sacrifices should we endure to avoid it? And, in particular: Would the failure of the various extra-terrestrial human communities established at t1 make the risk of meteor strike *more important*?

Contractualism and Utilitarianism both answer 'No' in both cases. Remain is (much) better than Leave, and the fate of other communities should not affect our reasoning about the meteor threat. These verdicts seem very plausible to me.

A final option is dualism. If Contractualism errs in one direction, and Utilitarianism in the other, then why not simply combine them? Perhaps our ethical theory should balance Contractualist obligations to particular people with Consequentialist obligations to promote future well-being. I suggest that this dualist option is worth pursuing.[3]

References

Academic References

[1] Beckstead, Nick. 2013. *On the overwhelming importance of shaping the far future*. PhD thesis, Rutgers.

[2] Caroti, Simone. 2011. *The Generation Starship in Science Fiction: A Critical History 1934-2001*. Jefferson, NC: McFarland and Co.

[3] This is the slightly edited text of a talk presented to the International Society for Utilitarian Studies conference in Karlsruhe in July 2018. I am very grateful to the ISUS audience for helpful comments, and to Christoph Schmidt-Petri and Michael Schefczyk for organising such a thought-provoking and enjoyable conference.

[3] Frick, Johann. 2015. “Contractualism and Social Risk?” *Philosophy and Public Affairs* 43: 175–223.

[4] _____. 2017. “On the survival of humanity.” *Canadian Journal of Philosophy* 47: 344-67.

[5] Holtug, Nils. 2010. *Persons, Interests, and Justice*. Oxford: Oxford University Press.

[6] Hooker, Brad. 2000. *Ideal Code, Real World*. Oxford: Oxford University Press.

[7] Kaczmarek, Patrick. 2017. “How much is rule-consequentialism really willing to give up to save the future of humanity?” *Utilitas* 29 (2). DOI: 10.1017/S0953820816000352.

[8] Kumar, Rahul. 2003. “Who can be wronged?” *Philosophy and Public Affairs* 31: 99–118.

[9] _____. 2015. “Risking and wronging.” *Philosophy and Public Affairs* 43: 27–51.

[10] Mulgan, Tim. 2006. *Future People*. Oxford: Oxford University Press.

[11] _____. 2015. “Utilitarianism for a Broken World.” *Utilitas* 27: 92-114.

[12] _____. 2017. “How should utilitarians think about the future?” *Canadian Journal of Philosophy* 47: 290-312. DOI: 10.1080/00455091.2017.1279517.

[13] Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

[14] _____. 2011. *On What Matters*. Oxford: Oxford University Press.

[15] Scanlon, Tim. 1999. *What we owe to each other*. Cambridge, MA: Harvard University Press.

[16] Temkin, L. S. 2015. “Rationality with respect to people, places, and times.” *Canadian Journal of Philosophy* 45 (5-6): 576-608. DOI: 10.1080/00455091.2015.1122386.

[17] Weinberg, Rivka. 2015. *The Risk of a Lifetime: How, When, and Why Procreation May Be Permissible*. Oxford: Oxford University Press.

Generation Starship Fiction

[18] Aldiss, Brian. 1958. *Non-Stop*. London: Faber and Faber.

[19] Ballard, J. D. 1962. “Thirteen to Centauris.” *Amazing Stories* (April).

[20] Bear, Elizabeth. 2007-2010. *Jacob’s Ladder*. 3 Volumes. New York: Bantam Spectra.

[21] Delaney, Samuel. 1965. *The Ballard of Beta-2*. New York: Ace.

[22] Heinlein, Robert. 1963. *Orphans of the Sky*. London: Gollancz. [originally published as two separate short stories in 1941].

[23] LeGuin, Ursula K. 2002. “Paradises Lost.” In *The Birthday of the World and Other Stories*. New York: Harper Collins.

[24] MacLeod, Ken. 2005. *Learning the World*. London: Orbit.

[25] Robinson, Kim Stanley. 2015. *Aurora*. London: Orbit.

[26] Wolfe, Gene. 1993-1996. *The Book of the Long Sun*. 4 Volumes. New York: Tor.

Hegels Begriff der Nützlichkeit: Zum Zusammenhang von Religionskritik und Terror

Ryu Okazaki, Humboldt University Berlin, Germany

Abstract

Some of the literature on Georg Wilhelm Friedrich Hegel's reception of the utilitarian philosophy of the French enlightenment, especially Claude Adrien Helvétius and Paul-Henri Thiry d'Holbach contends that Hegel misunderstands their concept of utility (Nützlichkeit), because he oversees the potential of that concept and reduces it to the isolation of human subjects, claiming it resulted in the terrorism of the French revolution. Against such criticism, this paper analyses the arguments within the *Phenomenology of Spirit*, in order to clarify the conceptual and logical connection between the criticism of religion by the French enlightenment and the terror of the French revolution. The first part deals with Hegel's analysis on the enlightenment's criticism of religion according to the three moments of superstitious consciousness, namely what it believes, why it believes and how it believes, in order to highlight not only the weak point of superstitious consciousness but also the fault of enlightenment. The second part analyses the emergence of the French Revolution and its relationship to the terrorism within it, from the perspective of the concept of utility, which the consciousness of enlightenment offers as the alternative to the religious relationship between human subjects, which was destroyed by the enlightenment. Through the reading of these passages it should become clear that the terror was caused by the French Revolution which try to eradicate all kinds of institutions. Our task now is to establish anew these institutions on the basis of the concept of utility as the reciprocity of need and its satisfaction between human beings.

Einleitung

Im Abschnitt über "den sich entfremdeten Geist: die Bildung" in der *Phänomenologie des Geistes* hat Hegel bekanntlich den Begriff der "Nützlichkeit" entfaltet. Wenn auch Hegels Einfluss im Werke der klassischen Vertreter des Utilitarismus, wie Jeremy Bentham oder John Stuart Mill kaum spürbar ist, weisen einige Literaturen darauf hin, dass Hegel dabei den Nützlichkeitsgedenken von den Denkern des französischen Materialismus, etwa von Claude Adrien Helvétius und Paul-Henri Thiry d'Holbach mitberücksichtigt, die als Vorläufer des utilitaristischen Denkens gelten. Blickt man auf den systematischen Stellenwert dieses Begriffs in der *Phänomenologie des Geistes*, dann ergibt sich, dass dieser Abschnitt, in dem der Begriff thematisiert wird, Ausgang vom Rechtszustand des römischen Reichs nimmt und dann die Geistesgeschichte bis hin zur Französischen Revolution darstellt. Dabei bedeutet die Entfremdung nicht nur den Verlust, sondern auch die Bildung, weil der Entfremdung eine Funktion "der Aufhebung des natürlichen Seins" (Hegel 1980, 267) zugeschrieben wird.

In diesem Bildungsprozess versucht das Bewusstsein, sich selbst zu verallgemeinern, weil die Allgemeinheit des Rechts zuerst etwas nur unmittelbar von außen her Gegebenes ist. Die Gestalt des Bewusstseins der Aufklärung tritt im zweiten Abschnitt auf, wobei sich die Aufklärung mit dem (Aber-)Glauben auseinandersetzt. Die Aufklärung versucht dort ihre Allgemeinheit dadurch zu beweisen, dass sie im Kampf mit dem Glauben "das Absolute" des Glaubens als kein Absolutes, sondern etwas vom Selbstbewusstsein Hervorgebrachtes aufzeigt. Hierbei analysiert Hegel, wie bereits gesagt, die aufklärerische Religionskritik des französischen Materialismus, der anhand der menschlichen Sinnlichkeit die Scheinhaftigkeit des Absoluten bzw. Gottes zu entlarven versucht und dabei zugleich den Begriff der Nützlichkeit darstellt.

Im Folgenden soll die Bedeutung des Nützlichkeitsbegriffs besonders hinsichtlich der zwei Begriffe, der Religionskritik und des Terrors etwas genauer analysiert werden. Wie wir sehen werden, weist Hegel einerseits auf die Mangelhaftigkeit der aufklärerischen Religionskritik hin, wobei es sich um eine Zerstörung der Bindung der Menschen untereinander handelt, die erst der Glauben ermöglicht hatte. Hegel sieht nämlich–einerseits das unerwünschte Resultat der Aufklärung in der Atomisierung des Menschen. Andererseits sieht er im Begriff der Nützlichkeit die Möglichkeit, eine neue Konzeption der Vergesellschaftung bzw. der Wiederherstellung des verlorenen menschlichen Zusammenlebens darzustellen, die ja nach dem Untergang der religiösen Bindungskraft eine moderne Weise der Vergesellschaftung gewährleisten soll. Gerade hierin besteht jedoch der kritische Ansatzpunkt der Literaturen z. B. die von Günther Mensching, der etwa betont, Hegel habe die Tragweite der Sozialphilosophie der französischen Aufklärung übersehen, indem Hegel die Relevanz ihrer utilitaristisch- materialistischen Religionskritik auf die Atomisierung reduziere, die ja letztendlich zum Terror führt (Mensching 1971). Um diese Kritik kritisch zu hinterfragen, soll angesichts der Nützlichkeit der logische, begriffliche Zusammenhang zwischen der Religionskritik und dem Terror geklärt werden. Dazu werde ich im ersten Schritt Hegels Analyse der aufklärerischen Religionskritik skizzieren, um das Argument für die Atomisierung durch die Aufklärung zu rekonstruieren. Sodann werde ich den Begriff der Nützlichkeit in Bezug auf die darauffolgende Erfahrung des Terrors betrachten, um die Bedeutung und Grenze der Nützlichkeitskonzeption als eines modernen Vergesellschaftungsprinzips zu ermessen.

I Hegels Analyse der aufklärerischen Religionskritik

Im zweiten Abschnitt über "den sich entfremdeten Geist" behandelt Hegel die Gestalten des Bewusstseins als Aufklärung. Dabei zielt die Aufklärung darauf ab, gegenüber der Beziehung des Glaubens zu seinem Gegenstand des Glaubens, nämlich zum absoluten Wesen, zu zeigen, dass das angeblich absolute Wesen kein absolutes ist. In diesem Zusammenhang ist jedoch zuerst auf die Bestimmung des glaubenden und aufklärerischen Bewusstseins zu achten. Beide Gestalten des Bewusstseins haben nämlich gemeinsam, dass jedes sich auf das absolute Wesen bezieht. Der Unterschied beider besteht hingegen darin, dass sich das glaubende Bewusstsein nicht als *Selbst*bewusstsein auf jenes Wesen bezieht, während das aufklärerische Bewusstsein das absolute Wesen als etwas von ihm Hervorgebrachtes sieht, sich nämlich als *Selbst*bewusstsein verhält.

Hegel erwähnt bei der Analyse die drei Momente des Glaubens, mit denen sich die Aufklärung auseinandersetzt: *das absolute Wesen*, *der Grund des Glaubens*, und *das Tun bzw. der Dienst desselben*. Diese Momente lassen sich auch so formulieren: *was* man glaubt, *warum* man glaubt, und *wie* man glaubt.

Das erste Moment ist das Absolute, oder der Gegenstand des Glaubens, also das, *was* man glaubt. Hierbei führt Hegel ein Beispiel vom Aberglauben an, dessen Objekt des Glaubens beispielsweise ein natürliches Ding ist: "Die Aufklärung sagt hiernach über den Glauben, dass sein absolutes Wesen ein Steinstück, ein Holzblock sei, der Augen habe und nicht sehe, oder etwas Brotteig, der auf dem Acker gewachsen, von Menschen verwandelt darauf zurückgeschickt werde" (Hegel 1980, 300). Dabei scheint folgerichtig gesagt zu werden, dass das Objekt des Aberglaubens nichts anderes als ein sinnliches Ding ist, und das ist ganz anders als das, was ein Absolutes sein soll. Dieser scheinbar rationalen Argumentation der Aufklärung widerspricht Hegel. Wenn die Aufklärung den Gegenstand des Glaubens als einen bloß sinnlichen nennt, versteht die Aufklärung den Glauben falsch. Denn ein solcher Gegenstand des Glaubens kann für den Glauben nur insofern ein Gegenstand des Glaubens sein, als dieser Gegenstand auf das Absolute bezogen wird. Trotzdem löst die Aufklärung die Gebundenheit des sinnlichen Dinges mit dem Absoluten auf und isoliert das sinnliche Moment als solches. Für den Glauben ist es ohne weiteres verständlich, dass der Gegenstand ein sinnliches Ding ist. Aber der Glauben weiß zugleich, dass diese sinnlichen Dinge ohne die Beziehung auf das Absolute nichts sind, zumal er nicht aufgrund der Sinnlichkeit jenes Dinges daran glaubt. Hegels Kritik am aufklärerischen Angriff auf das erste Moment lautet deshalb, dass die Aufklärung die Beziehung auf das Absolute einseitig übersieht, indem sie die Sinnlichkeit des Gegenstandes des Glaubens zu entlarven meint.

Nun komme ich zum zweiten Moment, dem Grund des Glaubens, d.h. *warum* man glaubt. Dazu betrachtet Hegel die Überlieferung einer beliebigen Religion als ein Beispiel des Grundes des Glaubens, wobei die Aufklärung dem Glauben zeigt, dass es immer zufällig ist, ob diese Überlieferung wahrhaft oder gefälscht ist. Wenn der Aufklärer den gläubigen Menschen nach der Wahrhaftigkeit fragt und dementsprechend der Glauben dem Aufklärer einen rationalen Grund zu geben versucht, dann ergibt sich unabsichtlich, dass der Glauben nicht aufgrund der Absolutheit seiner Religion, sondern aufgrund der Rationalität der Begründung daran glaubt, was aber dem unmittelbaren Glauben widerspricht. Er zeigt nämlich entgegen seiner Absicht seinen Glauben nicht an das Absolute, sondern an die Rationalität. Dazu merkt Hegel trotz der unbewussten Bekenntnis des Glaubens als Glauben an die Rationalität wiedermal an, dass diese kritische Verführung der Aufklärung so einseitig ist, dass sie den Glauben selbst außer Acht lässt. Die Aufklärung abstrahiert nämlich wieder die Bezogenheit irgendeiner religiösen Überlieferung auf das Absolute und isoliert das diesseitige zufällige Moment.

Und zuletzt komme ich zum dritten Moment des Glaubens, nämlich das Tun bzw. der Dienst des Glaubens, d.h. *wie* man glaubt. Dieses Moment veranschaulicht Hegel an der diesseitigen Aufopferung des Individuellen: “Dies Tun ist das Aufheben der Besonderheit des Individuums oder der natürlichen Weise seines Fürsichseins, woraus ihm die Gewißheit hervorgeht, reines Selbstbewußtsein nach seinem Tun, d.h. als fürsichseiendes einzelnes Bewußtsein eins mit dem Wesen zu sein” (Hegel 1980, 301). Hier ist wohl ein religiöser Dienst wie beispielsweise das Fasten gemeint, wodurch das Individuum auf seinen Genuss, nämlich das diesseitige Bedürfnis Verzicht tut, um den jenseitigen Genuss zu garantieren. Die Aufklärung merkt demgegenüber an, dass dieses Tun bzw. der religiöse Dienst unsinnig ist, weil es auf den Genuss Verzicht tut, um denselben zu bekommen. Hegel zufolge ist diese Kritik am Dienst jedoch ebenfalls einseitig. Sobald nämlich die Aufklärung den jenseitigen Genuss, den Zweck des Dienstes, mit dem diesseitigen Genuss identifiziert, übersieht sie, dass die diesseitige Aufopferung nur zwecks der jenseitigen Wohltat durchgeführt wird und nicht zwecks des Genusses im Allgemeinen.

Zusammenfassend lässt sich sagen, dass die Kritik Hegels an der aufklärerischen Religionskritik darin liegt, dass ihre Religionskritik die Bezogenheit des einzelnen, diesseitigen, zufälligen Moments auf das Absolute abstrahiert und das diesseitige sinnliche Moment als solches isoliert. Da sich jedoch der Glauben zugleich auf die Unmittelbarkeit des Absoluten stützt, kann der Glauben trotz der Einseitigkeit der aufklärerischen Religionskritik nicht mehr bestehen und wird nunmehr von der Aufklärung besiegt. Durch die drei Angriffe geht nämlich der Glauben an das Jenseits bzw. das Absolute verloren. Dies ist deshalb wichtig, weil die Beziehung der Menschen, die der Glauben, wenn auch auf prekärer Weise, gar-

antiert, verloren geht. Die Aufklärung befreit zwar den Menschen von seinen abergläubischen Ketten, aber aus dieser Befreiung resultiert zugleich in die Atomisierung der Menschen, weil die verbindende Kraft der Religion dadurch verloren gegangen ist.

II Hegels Begriff der Nützlichkeit und dessen Bezug zum Terror

Wenn auch das Resultat der aufklärerischen Religionskritik so zu sein scheint, dass durch die Atomisierung die Beziehung der Menschen irreversibel zerstört ist, so entsteht doch die Möglichkeit neuer Strukturen des Zusammenseins der Menschen: das Prinzip der Nützlichkeit. Im Folgenden werde ich nun versuchen zu klären, ob und wenn ja, inwiefern das Prinzip der Nützlichkeit mit dem Terror zu tun hat.

Hegel thematisiert den Begriff der Nützlichkeit zweimal: zuerst in Bezug auf den Angriff der Aufklärung auf das dritte Moment des Glaubens, nämlich das Tun und den Dienst des Glaubens und dann im darauffolgenden Abschnitt mit dem Titel "Die Wahrheit der Aufklärung", wo nach dem Untergang der religiösen Vergesellschaftung eine neue Konzeption der Beziehung der Menschen zuinander dargestellt wird. Im letzten Teil des Entfremdungsabschnitts geht es dann darum, wie die Nützlichkeit zum Terror übergeht.

Wie wir gesehen haben, basiert die aufklärerische Religionskritik gegenüber dem ersten sowie dritten Moment darauf, dass das Absolute des Glaubens tatsächlich mit der Sinnlichkeit des einzelnen individuellen Menschen verbunden ist. Hierbei findet die Aufklärung ein neues Prinzip der Vergesellschaftung der Nützlichkeit, die das religiöse Prinzip ersetzen soll. Die Wechselseitigkeit der Nützlichkeit beschreibt Hegel folgendermaßen:

> Wie dem Menschen alles nützlich ist, so ist er es ebenfalls, und seine Bestimmung ebensosehr, sich zum gemeinnützlichen und allgemein brauchbaren Mitgliede des Trupps zu machen. So viel er für sich sorgt, gerade so viel muss er sich hergibt, so viel sorgt er für sich selbst, eine Hand wäscht die andere. Wo er aber sich befindet, ist er recht daran und wird genützt. (Hegel 1980, 305)

In diesem Zitat geht es Hegel darum, dass, indem man etwas gebraucht und es genießt, immer eine Wechselseitigkeit des Menschen vorausgesetzt und aktiviert wird, gerade worin die Nützlichkeit besteht. Wenn auch das religiöse Band der Menschen zerstört wurde, so bleibt oder entsteht noch die Möglichkeit, basierend auf dem essentiellen Moment des Menschen, dem Genuss, ein Vergesellschaftungsprinzip zu konzipieren, was Hegel mit der im Begriff der Nützlichkeit angelegten Wechselseitigkeit darzustellen versucht. Etwas

Nützliches ist nur insofern nützlich, als dessen Genuss zugleich einen Genuss für den anderen Menschen schafft. Hegels Definition der Nützlichkeit lautet deshalb: "der *nicht in sich zurückkehrende* Wechsel der Momente des *ansich* und des *für ein anderes* und des *fürsich* Seins" (Hegel 1980, 314). In dieser Definition des Begriffs ist es besonders wichtig, dass die Nützlichkeit "nicht in sich zurückkehrt", weil diese Rückkehrlosigkeit keineswegs eine Feststellung des isolierten Individuums bzw. dessen Genuss bedeutet. Die Nützlichkeit kehrt nicht in irgendein Subjekt zurück. Die wesentliche Bestimmung der Nützlichkeit ist somit die Wechselseitigkeit.

Aus dieser Bestimmung der Nützlichkeit als der Wechselseitigkeit ergibt sich nun, dass Hegel die Nützlichkeit nicht bloß auf das isolierte Individuum reduziert, wie bereits erwähnte Literatur von Mensching, die den Terror auf Hegels angeblich falsches Verständnis zurückführen, meinen. Trotzdem macht das Bewusstsein im darauffolgenden Abschnitt die Terror-Erfahrung in der französischen Revolution. Die Kritik an Hegel, dass nach Hegel wegen des egoistischen Nützlichkeitsdenkens jedes Individuum die anderen nur zum eigenen Zweck als ein Mittel benutze und dadurch den Terror entstehen ließe, ist somit als falsch anzusehen, weil wie bereits gesagt, die Nützlichkeit keineswegs das Benutzen des anderen für eigene Zwecke bedeutet, sondern die Wechselseitigkeit voraussetzt. Wenn dem aber so ist, stellt sich doch die Frage, wie der Terror entsteht.

Hierbei ist es zuerst darauf zu achten, dass die Konzeption der Nützlichkeit noch nicht als solche realisiert ist. Die Nützlichkeit soll als Prinzip fungieren, mit welchem man eine neue moderne Vergesellschaftung konzipiert und auf dem basierend man die Gesellschaft institutionell umbauen kann. Trotzdem bleibt der Begriff der Nützlichkeit bloß ein Prinzip, d.h. es ist noch nicht verwirklicht. Dies verdeutlicht Hegel folgendermaßen:

> Die Nützlichkeit ist noch Prädikat des Gegenstandes, nicht Subjekt, oder seine unmittelbare und einzige *Wirklichkeit*. Es ist dasselbe, was vorhin so erchien; dass das *Fürsichsein* noch nicht sich als Substanz der übrigen Momente erwiesen, wodurch das Nützliche unmittelbar nichts anderes als das Selbst des Bewusstseins und hiedurch in seinem Besitz wäre. – Diese Rücknahme der Form der Gegenständlichkeit des Nützlichen ist aber *an sich* schon geschehen, und aus dieser innern Umwälzung tritt die wirkliche Umwälzung der Wirklichkeit, die neue Gestalt des Bewusstseins, *die absolute Freiheit* hervor. (Hegel 1980, 316)

Dieses Zitat befindet sich am Anfang vom letzten Abschnitt, "Die absolute Freiheit und der Schrecken". Hierbei ist auffällig, dass die Nützlichkeit noch nicht ein Subjekt, sondern ein Prädikat ist. Dies besagt, dass die Nützlichkeit noch nicht durch die Subjektivität verwirklicht ist, weil sich das Individuum der Vergesellschaftung noch nicht bewusst ist. In diesem Zusammenhang kann man sich daran erinnern, dass die Nützlichkeit wesentlich nicht in ein Subjekt zurückkehrt. Die Totalisierung der Individualität durch die Nützlichkeit erfolgt

dagegen erst durch “das Bewusstsein der absoluten Freiheit”, womit Hegel das Bewusstsein der Französischen Revolution beschreibt.

Was diese neue Gestalt versucht, ist Hegel zufolge somit, dass sich das Individuum ausgehend von seinem sinnlichen individuellen Sein das Ganze der Gesellschaft aneignet. Weil durch die aufklärerische Religionskritik alle andere Macht als das sinnliche Sein des Menschen liquidiert ist, muss das Individuum darauf abzielen, dass “jedes Individuum immer ungeteilt Alles tut” (Hegel 1980, 317). Diese “ungeteilte” Tun charakterisiert passend das Verwirklichungsprinzip der absoluten Freiheit. Es müssen nämlich alle Institutionen ausgelöscht werden, insofern sie nicht von dem fürsichseienden einzelnen Bewusstsein selbst ausgerichtet sind. In diesem Versuch der Verwirklichung der absoluten Freiheit ausschließlich durch sich selbst muss allerdings, laut Hegel, die Nützlichkeit verloren gehen und nur noch der Terror bleibt übrig.

Da die Französische Revolution alle vorherige Institution vertilgt und sodann nur noch die vereinzelten Menschen übrig sind, kann keine Institution mehr aufgebaut werden, die die Nützlichkeit konkretisieren soll. Die Nützlichkeit an sich soll zwar dem Menschen eine neue Vergesellschaftung nach dem Untergang der religiösen Gesellschaft ermöglichen. Insofern es aber an Institutionen fehlt, muss dieses Prinzip der Nützlichkeit notwendig versagen.

III Fazit

Hegels Deutung nach verwendet die Aufklärung das Moment der Sinnlichkeit bzw. des Genusses dazu, die Scheinhaftigkeit des Absoluten des Glaubens zu entlarven. Die Aufklärung hat zum Begriff gebracht, dass der Mensch nicht wegen der Absolutheit des Absoluten an das Absolute glaubt, sondern der Mensch glaubt an das Absolute, um den Genuss zu vergewissern oder zu rechtfertigen. Wenn auch durch die aufklärerische Verkehrung des Absoluten trotz der Einseitigkeit derselben die Beziehung der Menschen verloren geht und alle Menschen nur noch isoliert leben können, so tritt doch ein neues Prinzip des Zusammenlebens, das Nützlichkeitsprinzip, auf, weil in der unbewussten Handlung des Menschen eine Wechselseitigkeit der Nützlichkeit ausgedrückt ist. Hegel sieht, mit der Französischen Aufklärung bzw. dem Materialismus die Möglichkeit, dass sich durch die Nützlichkeit anstelle eines vormodernen religiösen Vergesellschaftungsprinzips ein modernes Konzept des menschlichen Zusammenlebens darstellen lässt. Ferner sieht Hegel den Terror nicht als eine notwendige Konsequenz des Nützlichkeitsdenkens. Hegel betrachtet den Zusammenhang von der Nützlichkeit mit dem Terror nicht als notwendig, sondern es geht ihm vor allem

darum, dass die Konkretisierung oder Vergewisserung der Nützlichkeit ohne eine institutionelle Vermittlung unmöglich ist. Die Nützlichkeit fungiert immer noch als ein wichtiges Prinzip, mit dem sich die Sittlichkeit nach der Zerstörung durch den Terror institutionell wiederherstellen lässt.

Literatur

[1] Hegel, Georg Wilhelm Friedrich. 1980. *Phänomenologie des Geistes*. Hamburg: Meiner.

[2] Mensching, Günther. 1971. *Totalität und Autonomie. Untersuchungen zur philosophischen Gesellschaftstheorie des französischen Materialismus*. Frankfurt am Main: Suhrkamp.

James Mill on Offences Committed by the Press

Filimon Peonidis, Aristotle University of Thessaloniki, Greece

Abstract

I critically discuss James Mill's rather neglected essay "Liberty of the Press" (1823). Mill embarks upon a normative inquiry to define justifiable exceptions to the liberty of the press in view of the British government's onslaught on dissenting political speech through a harsher libel law. He maintains that newspapers should not publish false accusations concerning private individuals or public officials, true statements we would classify as "hate speech" as well as incitements to obstruct lawful state procedures. On the contrary, the publication of criticism against the government and its officials, even if it is expressed in "a passionate language", is absolutely necessary for the proper functioning of representative democracy. I argue that, despite their shortcomings, Mill's arguments and recommendations bear heavily upon contemporary free speech debates.

Introduction

There is no doubt that James Mill's mature essay "Liberty of the Press" (Mill 1992, 95-135), first published as a supplement to the *Encyclopedia Britannica* in 1821 and then included in a volume of Mill's essays that appeared in 1823 has not drawn much attention, in stark contrast to his son's "Liberty of Thought and Discussion", the famous second chapter of *On Liberty*.[1] Whether this neglect is justified or not remains to be seen.

Mill is interested in determining anew which content-based legal restrictions on newspaper publications could be justified in a parliamentary democracy, given his conviction that the existing legislation is vague, unfair and utterly hostile to a robust conception of the liberty

[1] See Hamburger (1977, chapter 2), O' Rourke (2001, 9-15), Niesen (2015, 295-6) and Grint (2017). The last article focuses on Mill's unpublished commonplace books, and it is a valuable source for understanding the development of his thought on libel, censorship and the value of a free press.

of the press.[2] Thus, he does not aim at constructing a general theory of free speech or expression, but he, unavoidably, relies on moral considerations as well as on his overall understanding of the political function of a free press to justify his *de lege ferenda* conclusions.

The first conceptual distinction he makes is between "offences capable of being committed by the press" and "offences in the commission of which the press had an instrumental role" (Mill 1992, 99-101). An example will clarify the second term of the distinction. Suppose that a man publishes an ad in the personals section of a newspaper seeking the company of women who like gardening and the opera. In fact, this person is a psychotic serial killer who uses the services offered by newspapers to lure potential victims. It would be absurd to hold the press accountable for such a publication, since there was no indication of foul play. However, there are cases in which the offence, if any, *is* a publication in the press. Mill chooses to focus on two types of relevant cases, defamation of private individuals and politically-minded speech, and embarks upon an inquiry about the scope of the freedom of the press as far as these two types of speech acts are involved.

I Defamation of Private Individuals

For Mill, everyone has a right to be publicly portrayed as she really is. By defaming a particular individual through the press, we tarnish her reputation causing her considerable harm. Can this harm be morally unjustified to such an extent as to attract the legislator's attention? Mill's answer is that the law should take defamation seriously, and he distinguishes three separate instances of it that can be reconstructed as follows:

Case A. The press is subject to penalties if:

i. It has been proven in court that a certain publication had falsely imputed to A an action or a disposition to action.

ii. The imputed act or disposition "brings the evil of dislike or disrepute" upon A.

[2] In the aftermath of the Peterloo Massacre (1819), the British government passed six acts, two of which aimed at more effectively suppressing the dissemination of radical and dissenting political writings, a task that in practice proved to be difficult, given the number of radical authors that were being acquitted by juries. In fact, the acquittal rate in prosecutions for libel between 1817 and 1822 was sixty-two percent. See Harling (2001, 110). Cf. O' Rourke (2001, 12-3) and Grint (2017, 363-8). For the rather limited impact of "Liberty of the Press", see Hamburger (1977, 32-3).

He makes clear that it is profoundly wrong to publicly accuse someone for something she has not done or to attribute to her a character flaw she does not possess. In particular, the ensuing harm "affects a man in two ways": the victim suffers either a monetary loss or is made worse off from the "lessening [of] the marks of respect and affection which he would otherwise have received" (Mill 1992, 101). The exact nature of the afflicted harm is something that can in principle be decided in court, albeit lack of evidence or the complexity of the issues involved might aggravate the establishment of the relevant facts.[3] As far as the remedies for defamation are concerned, those who have lost money should receive adequate compensation, and those who have seen their reputation tarnished are entitled to a retraction, which will include the publication of the sentence of the court and whatever else is deemed necessary for the restoration of their good name. These penalties will also deter publishers from attacking the reputation of ordinary individuals. Moreover, Mill examines the likelihood that all these measures will prove insufficient in the sense that the general public will not be convinced that the defamed person has been wrongly accused. This might happen if the public is aware of the existence of evidence that was withheld from court, or it is incapable of seeing the truth of the matter. In the second case, the government is held accountable for not cultivating the epistemic virtues of its citizens.

Case B. The press is not subject to penalties if:

i. It has been proven in court that a certain publication had imputed to A an action or a disposition to action that are unquestionably true.

ii. The imputed act or disposition "brings the evil of dislike or disrepute" upon A.

iii. The ensuing loss of A's good reputation is morally justified, since it makes A recognize her fault and it deters members of society from behaving in a similar manner.

Here Mill argues that the press should not be held answerable to the law for exposing someone's true character or for revealing the morally repugnant actions someone has performed, even if the publicity she will receive will make her fall into disrepute. On the contrary, there are good utilitarian reasons for doing so: "The advantage which would be derived from the true exposure of any man's actions of any sort, would exceed beyond calculation the attendant evil" (Mill 1992, 106). Mill is closely following Bentham (2005, 106)

[3] It is noteworthy that, contrary to Bentham, in this essay Mill does not consider the possibility of a miscarriage of justice due to juries that have been selected to reach the verdict magistrates want. Moreover, he does not distinguish between telling a lie and simply making an erroneous statement. Perhaps, this could be explained by the emphasis he places on the harm suffered by the defamed person, which is not affected by the defamer's false belief that she was right in her accusations. For his earlier views see Grint (2017, 372, 374-5).

in attributing to people a strong motive to seek the pleasures and avoid the pains resulting from the moral opinion other people have about them. This makes them fear dishonor, disgrace, infamy and shame. If all agents (or most of them) realize that morally reprehensible conduct, no matter what legal penalties it might bring forth, will be made public knowledge through the press, they will have an additional incentive to abide by the established moral rules and this will work to everyone's benefit.

Case C. The press is subject to penalties if:

i. It has been proven in court that a certain publication had imputed a true fact or action to A.

ii. The imputed fact or action "brings the evil of dislike or disrepute" upon A.

iii. The ensuing loss of A's good reputation is morally unjustifiable, since it is undeserved, and it is generated by a crooked system of social or religious prejudice.

Mill admits that his recommendations concerning the press's liability in case B are not absolutely valid. It is likely for someone to be met with disapproval, contempt or even derision for her actions, her character traits or her manners for the wrong reasons. In these rather rare cases, where the moral sentiments of the many are "perverted" and "corrupted," individuals should be protected from "the declaration of truth by the press" (Mill 1992, 108). When it is written that someone is of humble origins and this statement is beyond dispute, this person will feel the "antipathy" of society not because of her own fault but because of the domination of an aristocratic class-system, which puts an unjustifiably high premium on what it defines as noble ancestry and has managed to command widespread acceptance. In this case, this particular truth (which of course makes sense only within a system of aristocratic values) should not be reported by the press to keep this person out of harm's way.

There is an important point undergirding Mill's discussion. Defamation is not only about the truth or falsity of what is said, but also about endorsing a normative framework that bestows – in a justifiable or unjustifiable manner – disvalue on what is said. There is nothing wrong with having "slit eyes" or a "red neck". These descriptions become derogatory because those who use them express through them a long-established bias against people of Asian origin or southern American farmers. This is not the case when you call someone who has been convicted for appropriating money she did not own an "embezzler". When it comes to the legal evaluation of defamatory speech, moral judgments are unavoidable.

II Politically-Minded Speech

Up to this point, the discussion of free-press issues has been associated with their positive or negative consequences for private individuals as well as for society's moral health. The next topic Mill discuses – politically-minded speech expressed through the press – provides him with the opportunity to highlight the general significance of the liberty of the press in a democratic society. The establishment of this liberty allows ordinary individuals (at least most of them) to undertake and successfully perform the role of responsible and competent citizens, thus securing the proper functioning of representative democracy. His basic argument (Mill 1992, 115-30) can be summarized as follows:

a. The people need adequate information to choose the right representatives and to assess their performance, when they come into power.

b. Given the tendency of those in power to care more about their own interests, the people must be able to express their discontent with them, which is the only means for removing the evils of bad government.

c. There is no epistemic authority that can tell the body politic what is right and wrong in matters of government.

d. Therefore, all views and reports (positive or negative) concerning matters of government and the performance of politicians and other state officials should be freely published to enable the majority of citizens to weigh the available evidence, to make up their own minds and to act accordingly.

This highly idealized argument, if it is valid,[4] undoubtedly justifies a general presumption in favor of the liberty of the press, but, in my opinion, Mill's originality lies in his treatment of three particular cases involving the expression of politically-minded speech. General endorsements of the free press as the oxygen of democracy are common from the eighteenth century onwards, but the devil is hiding in the (legal) details.

[4] It is difficult to understand why "there is moral certainty … that the greater number of [the people] will judge aright" (Mill 1992, 121), if no one is epistemically qualified to distinguish the right political opinions from the mistaken. In my view, Mill has either to produce a skeptical argument – if we can never tell with certainty that an opinion is false, no restrictions in their publication are allowed – or to invoke something like Bentham's public opinion tribunal.

II.a Subversive Advocacy

According to a contemporary definition (Tassopoulos 1993, 13), "subversive advocacy is the inciting of other citizens to undertake the violent overthrow of governmental institutions as a means of political change". Although Mill does not use this term, he discusses print exhortations to the "people in general to take arms against the government, for the purpose of altering it against the consent of its rulers" (Mill 1992, 112). Surprisingly, he argues that subversive advocacy should not be made an offence. On the one hand, if the people are determined to revolt and only a spark is needed for the fire to start, it is pointless to punish the inciter. In such cases, no one is to be deterred from the existence of penalties. On the other hand, if the inciter is not to be taken seriously by her readers, no harm is done, and therefore it would be wrong to criminalize harmless speech.

II.b Incitement to Obstruct Lawful State Procedures

Mill has more to say on what he calls "exhortations to obstruct the operations of government in detail." Here one does not wish the overthrow of the government but instead objects to a particular established procedure related to the general functioning of the state and/or to the outcome that is expected to come out of it. Thus, she encourages the public to intervene and forcefully stop the above lawful procedure (cf. Mill's example in 1992, 113).

Not all exhortations fall within the same category. There are "direct" and "explicit" exhortations, which leave no doubt about the author's intentions. In addition, there are "implied" and "constructive" exhortations from which the author's position cannot be inferred with certainty. For Mill, there is a great difference between writing, "let's storm the parliament to stop the government from passing this onerous new income tax bill" and "I wonder how the government will react, when the people storm the parliament to demand the withdrawal of this onerous new income tax bill" (the examples are mine). Statements of the second type are to be interpreted as expressions of harsh political criticism and not as encouragement to commit crime and therefore they enjoy the protection of the law.

On the contrary, statements of the first type should be subject to legal sanctions, when they lead to the use of force against the state apparatus. Mill realizes that explicit exhortations addressed to small groups under the right circumstances are likely to be effective, thus contributing to a blunt and forceful obstruction of lawful proceedings. The legal system cannot allow individuals to have any role whatsoever in the violation of particular laws. Citizens

should be deterred from showing disrespect for the normal operations of government, especially when they are about to disagree with decisions resulting from these operations. Nevertheless, Mill insists that the sanctions imposed should be of "moderate severity" and not be motivated by "vengeance". Given the tendency of the bearers of power to "multiply the list of offences against governments," we should take steps to prevent penalties that prescribe some sort of retribution for insults given to state officials. That is why it is a mistake to retain offences like contempt of court. The unjustifiable harm done here is strictly restricted to the obstruction of lawful state procedures.

II.c Criticism of Public Officials

Finally, when it comes to the criticism of public officials in general, anything is allowed with one exception.

> If, in supporting his opinion of the inaptitude of any public functionary, [an individual] imputes to him actions which there is not even an appearance of his having performed, that limited prohibition … will strictly apply. With this exception, freedom should be unimpaired. (Mill 1992, 126)

Mill would be inconsistent in allowing public officials to be accused of things they have not done. This is ruled out by his endorsement of the value of factual truth (with the exception noted above). When someone has only suspicions of foul play on the part of a public functionary, she should state so. Otherwise, she should provide adequate evidence. What about the language used in attacking public officials? Should we demand certain standards of decency and good manners to apply? In the last section of his essay (Mill 1992, 130-35), Mill reflects upon the prohibition of "indecent discussion". He maintains that criticism of public officials is unavoidably associated with certain strong sentiments like contempt, anger, sympathy, admiration and hatred caused to third parties by their acts and omissions. Politics is not like mathematics. The public performance of an office-holder is expected to trigger positive and negative feelings in any citizen, feelings she is entitled to convey to her fellow citizens to convince them of the rightness of her views. The obligation of critics to provide evidence does not include an obligation to express it in a "calm and gentle language". Moreover, any official attempt to set standards of decency regarding the language of political criticism runs the following risks: (a) to impose the relevant *subjective* views of the legislators on the public and (b) to give judicial authorities a pretext to prosecute views they dislike

as libelous. Hence, Mill concludes, the evils of punishing language "to which the name *passionate* could be applied" far outweigh the goods arising from not punishing it.[5]

Conclusion

I have tried to show that James Mill has remarkable practical insights to offer, which could be charitably interpreted to become relevant to contemporary free speech discourse, even if one disagrees with certain of his arguments and recommendations.[6] The issues that concerned him continue and will continue to occupy scholars, legislatures and courts alike. Thus, we have at least one good reason for a more thorough study of his work, one that is more detached from his son's legacy.

References

[1] Bentham, Jeremy. 2005. *An Introduction to the Principles of Morals and Legislation*. Edited by J. H. Burns, and H. L. A. Hart, with a new introduction by F. Rosen. Oxford: Clarendon Press.

[2] Grint, Kris. 2017. "The Freedom of the Press in James Mill's Political Thought." *The Historical Journal* 60: 363-83.

[3] Hamburger, Joseph. 1977. *James Mill and the Art of Revolution*. Westport, CT: Greenwood Press.

[4] Harling, Philip. 2001. "The Law of Libel and the Limits of Repression, 1790-1832." *The Historical Journal* 44: 107-34.

[5] Mill, James. 1992. *Political Writings*. Edited by Terence Ball. Cambridge: Cambridge University Press.

[5] It is a hard for a modern reader to fail to notice a certain affinity between Mill's views on the freedom to criticize public officials and the ruling of the U.S. Supreme Court in *New York Times Co. V. Sullivan* (1964) according to which even a false defamatory statement related to the official conduct of a public official enjoys constitutional protection unless "he proves that the statement was made with 'actual malice' – that is, with knowledge that it was false or with reckless disregard of whether it was false or true".

[6] For some people it is improper to call someone a "murderer" publicly, even if she has been convicted of murder, and there is a wide consensus that contemporary mass media have a tendency to deceive, disorientate and manipulate the public in ways Mill could not have thought of.

[6] Niesen, Peter. 2015. “Parole, vérité et liberté de Jeremy Bentham à John Stuart Mill.” *Archives de Philosophie* 78: 291-308.

[7] O’Rourke, K. C. 2001. *John Stuart Mill and Freedom of Expression*. London: Routledge.

[8] Tassopoulos, Ioannis A. 1993. *The Constitutional Problem of Subversive Advocacy in the United States of America and Greece: A Comparison of the Legal Guarantees of Political Speech in Times of Crisis*. Athens-Komotini: Ant. N. Sakkoulas.

Parfit's Reorientation between *Reasons and Persons* and *On What Matters*

Ingmar Persson, University of Gothenburg, Sweden

Abstract

This paper aims to show that between *Reasons and Persons* and *On What Matters* the orientation of Derek Parfit's philosophy underwent a significant change. The approach of *Reasons and Persons* is largely *revisionist*, which is exemplified by his reductionist account of personal identity. This account is suppressed in *On What Matters* apparently because it does not fit in with the *conciliationalist* project of this work. The aim of the first two volumes of this work is to show that, on the basis of a non-naturalist theory of normative reasons, rule-consequentialism, Kantian and Scanlonian contractualism could converge into a Triple Theory. In the third volume, the conciliationalist approach is carried further by Parfit's attempt to show both that his metaethical position is in essential agreement with rivals, like Allan Gibbard's expressivism, and to reconcile parts of common-sense morality and consequentialism in order to bring them together in the Triple Theory. However, there isn't space here to pursue the problems with aspects of Parfit's conciliatory project other than those of personal identity.

Introduction: Parfit's Reorientation from Revisionism to Conciliationalism

A few years after the publication of *Reasons and Persons*, Derek Parfit said to me as we walked along High Street in Oxford: 'I don't want to become like Hare'. He went on to explain that what he had done in his book was to present various arguments rather than trying to defend any particular ethical position. By contrast, what Richard Hare was known for was precisely to defend vigorously certain ethical positions, both in meta-ethics and normative ethics: a form of rule-utilitarianism based on his universal prescriptivism.

The normative framework of *Reasons and Persons* is overall consequentialist, and already at the time of writing it Parfit's favoured meta-ethical view was a non-naturalist theory of reasons, according to which normative reasons are irreducible to natural facts. But he was anxious to stress that many of his arguments didn't presuppose this objectivist theory as opposed to subjectivist theories of reasons. And in this book he didn't *defend* consequentialism, though he's obviously most at home in this tradition, a tradition that he inherited from Henry Sidgwick, just as he inherited non-naturalism about normative reasons from him. Thus, at the time of writing *Reasons and Persons*, Parfit differed from Hare in that he

didn't try to argue for any particular normative position on the basis of a particular meta-ethical position. He didn't take what he calls the 'High Road' (1987, 447).

At that time, Hare hadn't yet argued in print that Kant could have been a utilitarian, but he was to do so in 'Could Kant have been a Utilitarian?' (Hare 1997). There's an obvious similarity between Hare's idea of the universalizability of moral judgments and Kant's first formulation of the categorical imperative to the effect that we should act only on maxims that we could simultaneously will to be universal laws (Hare 1997, 153-54, 161). So, Kant eventually came to occupy a place of some prominence in Hare's moral philosophy – just as he did in Parfit's moral philosophy.

Whilst Kant is barely mentioned in *Reasons and Persons*, he undeniably looms large in *On What Matters*. But the only positive claim Parfit succeeds in extracting from Kant seems to come from his first formulation of the categorical imperative, precisely what Hare also zoomed in on. It's the centrepiece of what Parfit calls Kantian contractualism. A main objective of the first two volumes of *On What Matters* is to show that Kantian contractualism could be aligned not only with rule-consequentialism, as in Hare's case, but with Scanlonian contractualism as well, making up what he calls the 'Triple Theory'.

This normative unification project is carried further in the third volume of *On What Matters*, where Parfit also argues (2017, ch. 58) that his Triple Theory supports the acceptance of an improved version of common-sense morality. This, too, is in line with Hare's 'two-level' moral theory in which common-sense morality corresponds to an 'intuitive' level which is underpinned by a consequentialist 'critical' level (1981). Moreover, Parfit's normative unification project, like Hare's, is based on a certain meta-ethical position, albeit of a non-naturalist kind, which is diametrically opposed to Hare's prescriptivism. Nevertheless, central for both of them was that there are moral judgments that are objective to the extent that all rational subjects in possession of all relevant empirical facts would agree about them.

Consequently, Parfit ended up doing moral philosophy much more like Hare than you could have anticipated 30 years ago. They were both hoping to show that there's a *'single true morality*' (2011, II, 155), which could take the shape of rule-consequentialism. This is of course compatible with there being huge differences with respect to both their ways of proceeding and their conclusions.

In the Introduction to *Reasons and Persons*, Parfit refers to Peter Strawson's distinction between 'two kinds of philosophy, descriptive and revisionary' (x). Parfit describes himself as a revisionist 'by temperament', and the tenor of RP is indeed revisionist. In these terms, *On What Matters* is by contrast 'descriptive' or conservative as regards matters of normative

substance. Its aim isn't as much to challenge commonsensical moral claims as to defend them by showing that, with a bit of revision, they can be supported by leading ethical theories once we get the meta-ethics right. His fear is that 'if we cannot resolve our disagreements … morality might be an illusion' (2011, II, 155).

Parfit is celebrated for developing revisionist views about personal identity in *Reasons and Persons*, and there's evidence that he never repudiated this revisionism. But it receded into the background in *On What Matters*, as I'll try demonstrate in section I. The reason for this recession is probably that it's in tension with the 'conciliationalist' project of this book. Of course, revisionism and conciliationalism aren't *inconsistent*; it's just that a combination of them is unlikely to succeed, since revisionist views are likely to stir up disagreement. In section II I'll briefly comment on some other aspects of his conciliationalist project.

I The Suppression of Reductionism about Personal Identity

In *On What Matters* (I, section 19) Parfit develops a version of Sidgwick's 'dualism of practical reason' which consists in there being two kinds of reasons to care about the well-being of individuals: *self-interested* or *personal* reasons to care about our own well-being, and *impartial* reasons to care about everyone else's well-being. He concedes to Sidgwick that he 'rightly claims that we have reasons to be specially concerned about our own future well-being' but, he goes on, many of these reasons 'are provided, not by the fact that this future will be *ours*, but by various psychological relations between ourselves as we are now and our future selves' (I, 136). Notice that this implies that *some* of our reasons to care about our future are provided by the fact that it's *ours*. This contradicts his famous claim in part III of *Reasons and Persons* that 'personal identity is not what matters'.

It isn't that he has *abandoned* this claim about personal identity, for in his paper 'We are not Human Beings', published the year after *On What Matters* (2011), he re-affirms his allegiance to it. After having argued that the animalist or biological view of our identity – according to which are identical to our human organisms – isn't true, he confesses that he has 'a reason to *wish* that Animalism were true' (2012, 27), since this would make it easier for him to vindicate his claim that our identity doesn't matter. Contradicting what's implied in *On What Matters*, he writes that when 'we have reasons for special concern about our future, these reasons are not given … by the fact that this will be *our* future' (2012, 27). So, some sort of double-thinking or ambivalence about the importance of personal identity is present.

As regards the psychological relations that provide us with reasons to care about ourselves, we have 'partly similar relations to some other people, such as our close relatives, and those we love' (2011, I, 136). Thus, these relations provide us with '*personal* and *partial* reasons to care about the well-being of ourselves and those to whom we have close ties' (2011, I, 136). These reasons are 'only *very imprecisely* comparable' (2011, I, 137) to impartial reasons. They're only very imprecisely comparable in the sense that, though we can tell, for instance, that we're permitted to save our own lives rather than the lives of at least two strangers, we can't give anything like a precise answer to how many strangers we're allowed to sacrifice to save ourselves. According to Parfit, this imprecision is due to the fact that, whereas impartial reasons are *person-neutral*, self-interested and partial reasons are *person-relative* in the sense that they 'are provided by facts whose description must refer to us' (2011, I, 138), either because these facts concern *our own* well-being, or the well-being of people to whom *we* have close ties.

Now although self-interested reasons *permit* us to give somewhat greater weight to our own well-being in comparison to the well-being of strangers, Parfit thinks – surely rightly – that it would be 'too egoistic' (2011, I, 139) to maintain that they *require* us to give greater weight to our own well-being: we're permitted to 'give equal or even greater weight to some stranger's well-being' (2011, I, 139) than our own. He notes, however, a difference in this respect between reasons to care about our own well-being and the well-being of others to whom we have close ties. For in a case in which I could save either my own child or the child of some stranger 'I ought morally to give priority to my child' (2011, I, 141).

This difference seems sufficient for holding our self-interested reasons to care about our own well-being to be a different kind of reason than partial reasons to care about the well-being of others to whom we have close ties. Thus, whereas Parfit lumps together self-interested and partial reasons and talks about *two* kinds of reasons (e.g. 2011, I, 138), a case can be made for distinguishing *three* different categories of reasons: *self-interested, partial and impartial reasons*.

I'll now try to show that both this alleged difference between self-interested and partial reasons and the claim that both of them are only very imprecisely comparable to impartial reasons conflict with his reductionism about personal identity in *Reasons and Persons*. According to this reductionism, our identity consists in the holding of psychological and physical relations that can hold to a greater or lesser degree, and there are cases in which it's *indeterminate* whether or not they hold to such a degree that it can truly be said that we persist. This is a claim about the *analysis* of our identity. But his reductionism also features a claim about its *importance*, expressed by the slogan that personal identity is not what

matters. What matters in identity is rather 'psychological connectedness and/or psychological continuity, with the right kind of cause' (Parfit 1987, 214). The right kind of cause is normally the persistence of one and the same brain.

Psychological *connectedness* consists in the holding of psychological connections, like memories and interests, and psychological *continuity* in chains of such connectedness. Strong enough psychological continuity is necessary for personal identity, but not sufficient. Suppose that each hemispheres of the brain of a person is capable of underpinning the psychology of the person and that they're separated and transplanted to two different bodies (Parfit 1987, ch. 12). In such a case of *branching* psychological continuity, personal identity is disrupted, since the original person can't be identical to both of the resulting persons, who are clearly distinct from each other, and it would arbitrary to identify the original person with any one of them. Personal identity, then, consists in *non-branching* psychological continuity, with the right kind of cause. But Parfit claims that the occurrence of such a division isn't *worse* for us than survival as the same person with the same degree of psychological connectedness and continuity. Thus, it's the latter relations that matter for us, not personal identity.

If this is correct, it raises the question how we could be morally permitted to sacrifice ourselves for a smaller benefit to some stranger, but not sacrifice someone else who is closely related to us. For surely we must be morally permitted to sacrifice the people who come into being when our psychological continuity branches no less than ourselves. The relations that matter are in both instances the same. This conclusion undermines the alleged difference between self-interested and partial reasons.

Furthermore, Parfit's reductionist campaign in *Reasons and Persons* encompasses a proposal to extend morality into the intrapersonal sphere which implies that we're *not* morally permitted to treat ourselves differently than others. He considers 'a boy who starts to smoke, knowing and hardly caring that this may cause him to suffer greatly fifty years later' (1987, 319-20). In such cases in which there's a considerable loss of psychological connectedness, but enough psychological continuity to preserve identity, he proposes that we may outlaw great imprudence by importing moral reasons into the intrapersonal sphere, with the result that 'we ought not to do to our future selves what it would be wrong to do to others' (1987, 320). But this is incompatible with the claim that we're morally permitted to give 'even greater weight to some stranger's well-being' than our own. This would permit us, for instance, to die a more painful death to save some stranger from a less painful death, though we're hardly permitted to save one stranger from a less painful death at the expense of another stranger dying a more painful death. Still less are we permitted to let someone close to us die a more painful death.

Consider now the proposed very imprecise comparability between person-relative reasons and impartial reasons. Imagine a spectrum of cases in which the reduction of psychological continuity successively increases, by successively greater parts of the underlying brain being replaced by parts supporting different memories and so on. It would seem that the difference in strength between a case in which there's definitely enough (non-branching) psychological continuity for identity and a case in which there's too little for identity, and barely enough for someone being close to us, could be as great as between the latter and a case in which psychological continuity is just about so weak that there would somebody else who isn't even close to us. If so, and reasons for concern are based on psychological connectedness and/or continuity, it might be wondered why their comparability must be less precise in the latter case when person-relative reasons are compared to impartial reasons than in the former case when only person-relative reasons are involved.

Suppose, however, that we instead adopt the non-reductionist view that Parfit in *Reasons and Persons* attributes to common sense and Sidgwick, to the effect that the difference between ourselves and others involves 'a further fact' beyond psycho-physical continuities, a fact that is either–or rather than a matter of degrees (1987, 138-39, 329). Then it appears more comprehensible how self-interested and partial reasons, on the one hand, and impartial reasons, on the other hand, could be only very imprecisely comparable given that the former are person-relative in the sense that they're provided at least partly by facts that refer to what's irreducibly *ourselves* (though it's harder to see why self-interested and partial reasons should differ in the way he thinks). This is to insinuate that Parfit suppresses his reductionist view of personal identity when he propounds a dualism of practical reason in *On What Matters*.

The reasons for this suppression probably have to do with the fact, as he confesses, he is 'deeply worried by disagreements with people who seem as likely as I am to be getting things right' (2017, xiii). Samuel Scheffler concurs with this diagnosis in his Introduction to *On What Matters*: "The drive to eliminate disagreement … is a defining feature of Parfit's work" (2011, I, xxxi). If Parfit was a revisionist by temperament, he was also a conciliationalist. The more embedded he became in the academic establishment, the stronger he may have felt the pressure to include more people in the agreement. Radical revisionary ideas inevitably kick up dust, as he had experienced not least in the case of his views about personal identity. Feeling that he couldn't convince adversaries of the truth of these ideas, he might have felt that they better be covered up in a unification project.

II Concluding Remarks about Other Aspects of Parfit's Conciliatory Project

It's also important to Parfit's pivotal belief to the effect that we have an intuitive ability to recognize irreducibly normative – rational and moral – truths that it doesn't become apparent that we have 'deeply conflicting normative beliefs' (2011, II, 546). By contrast to the two first volumes of *On What Matters* in which Parfit tries to *refute* rival meta-ethical theories, in volume III he rather extends the conciliationalist strategy from normative ethics into meta-ethics. He argues, for instance, that his and Allan Gibbard's "main claims don't conflict" (2017, 225). But Gibbard seems in fact to stick to his expressivism. He's prepared to say that normative claims can be true, but this is truth only of "a minimalist sort" (2017, 221), according to which "'It's true that suffering is bad' just means that suffering is bad" (2017, 205). Gibbard adds: "As for truth of a more robust sort, I suspend judgment pending some satisfactory explanation of what this more-than-minimal truth consists in" (2017, 221-22). So, contrary to Parfit's wishful belief (2017, 226), Gibbard explicitly refrains from committing himself to truth in Parfit's more robust "descriptive sense" (2017, 226) which underlies his non-naturalism. The fact that there's no agreement about there being reasons in Parfit's sense obviously reduces the significance of building an agreement in normative ethics by appeal to these putative reasons.

A leitmotif of part IV of *Reasons and Persons* is the failure to come up with "a new theory of beneficence" that can cope with problems in population ethics such as the non-identity problem and the repugnant conclusion (1987, 443). In Parfit's view, in order to show that any moral "theory could be *objectively* the best theory", "we must find a theory which resolves our disagreements' about these matters" (1987, 452). But in *On What Matters* there's no attempt to show how a morality underpinned by the Triple Theory could resolve these disagreements. The task of finding "a new theory of beneficence", "Theory X", has disappeared from the horizon. The non-identity problem and the repugnant conclusion are together with reductionism about personal identity the topics in *Reasons and Persons* that have generated most controversy, but they have symptomatically almost vanished in *On What Matters*.

To conclude, between these books there's a shift in the orientation of Parfit's moral philosophy, from revisionism to conciliationalism. As my remarks indicate, I'm skeptical of the success of his conciliatory project, but I don't think this is due to any shortcoming on Parfit's part. I regard him as the greatest philosopher I've met, and I'm glad I got around to telling him so before he died. It's reasonable for moral philosophers to aim to establish a rational

consensus about what's morally right and wrong, and what the ground for this is. I'm however strongly inclined to think that this goal is unattainable because we have fundamentally conflicting intuitions both in normative ethics and in meta-ethics.

References

[1] Gibbard, Allen. 2017. "Gibbard's Commentary." In Parfit 2017. 205-24.

[2] Hare, R. M. 1981. *Moral Thinking*. Oxford: Clarendon Press.

[3] _____. 1997. *Sorting out Ethics*. Oxford: Clarendon Press, 147-65.

[4] Parfit, Derek. 1987. *Reasons and Persons.* Reprinted with corrections. Oxford: Clarendon Press.

[5] _____. 2011. *On What Matters*. vols. I & II. Oxford: Oxford University Press.

[6] _____. 2012. "We are not Human Beings". *Philosophy* 87 (1): 5-28.

[7] _____. 2017. *On What Matters*. vol. III. Oxford: Oxford University Press.

Prioritarianism and the Moral Negativity Bias

Ingmar Persson, University of Gothenburg, Sweden

Abstract

The moral negativity bias is an intuition to the effect that there's more of a moral reason to reduce what's bad for individuals than to increase what's equally good for them. This intuition, if sound, supports a negatively weighted utilitarianism to the effect that what's bad for individuals has greater negative moral weight than what's equally good for them has positive moral weight. But it's here argued that the moral negativity bias can be given a debunking explanation that undermines its soundness. This explanation refers to the psychological fact that negative feelings are usually stronger than positive feelings because the badness of the deteriorations with which they are associated is generally greater than the goodness of the improvements associated with positive feelings. Appeals to intuitions which could be accounted for by the moral negativity bias seem also to have been made in support of prioritarianism in opposition to egalitarianism. If this bias is given a debunking explanation, this support is undercut.

I Negatively Weighted Utilitarianism, Prioritarianism and the Negativity Bias

Intuitively, it seems there's more of a moral reason to reduce – that is, to remove, prevent or avoid producing – what's intrinsically bad for individuals than to increase what's equally good for them – more precisely, to increase goodness *directly*, as opposed to doing it by reducing what's bad. This is *the moral negativity bias*.

Negative utilitarianism – championed e.g. by Karl Popper (1966, ch. 9, note 2) – is commonly understood as the doctrine that there's *only* moral reason to reduce what's bad for individuals, and *no* moral reason to increase what's good for them; in other words, only what's bad for individuals has moral weight or value. A less extreme view – *negatively weighted utilitarianism*, *nw-utilitarianism* – is that there's *stronger* moral reason to reduce what's bad for individuals than to increase what's good for them to a corresponding degree; that what's good for individuals has *some*, but smaller, moral weight or value than what's equally bad for them. Nw-utilitarians claim e.g. that there's stronger moral reason – or that it has greater moral weight – to reduce the intrinsic badness that suffering due to physical pain has for us than to increase the intrinsic goodness that happiness due to physical pleasure to a corresponding degree has for us.

The practical implications of nw-utilitarianism converge to some extent with those of teleological prioritarianism. According to this version of prioritarianism, the positive (negative) moral weight of benefits (burdens) to the worse-off is proportionally greater than the moral weight of the same benefits (burdens) to those who are better off. Accordingly, it's morally irrelevant whether a benefit – i.e. something that makes recipients better off – consists in an increase of what's intrinsically good or a decrease of what's intrinsically bad. What morally matters is the positions of recipients on a scale from being better off to being worse off: the worse off they are, the greater the moral weight of benefiting them is. But it's reasonable to surmise that benefits to those lower down the welfare scale are likely to consist more in the reduction of intrinsic badness than benefits to those higher up, which will consist more in increases of intrinsic goodness. Benefiting those who are worse off will then be morally more important not only on prioritarianism, but also on nw-utilitarianism.

All the same, nw-utilitarianism and prioritarianism clearly differ. Compare a choice between either preventing a pain of somebody who's better off or causing somebody who's worse off to feel a pleasure of the same magnitude, which would do more to put the two individuals on the same welfare level. In contrast to prioritarianism – and egalitarianism – nw-utilitarianism would recommend preventing the pain, which might seem intuitively plausible.

Due to the fact that their practical implications sometimes converge, not only nw-utilitarianism but also prioritarianism can be nurtured by the moral negativity bias, as will transpire in section III. I believe this bias to be an element of common-sense morality, but shall in section II present a debunking explanation of it which is compatible with equal amounts of goodness and badness in fact having equal moral weight, in accordance with traditional utilitarianism.

II A Debunking Explanation of the Moral Negativity Bias

Let's begin by surveying some psychological facts. It seems clear that the signal of something's being harmful for our bodies, physical pain, can be more intense and bring more suffering than the signal of something's being beneficial for our bodies, pleasure, can bring enjoyment. For instance, there's surely no pleasure so intense that having it is worth undergoing the most painful torture of equal duration.

Some philosophers, like Arthur Schopenhauer, have claimed that, in contrast to pain, pleasure is nothing positive; it's merely the cessation of pain (1995, 146). This is certainly false, but it's symptomatic that nobody has been crazy enough to uphold the corresponding view about pain, that it's just the absence of pleasure, a neutral state. Pain can simply be so intense that its positive existence is undeniable: nobody could seriously think that the most painful torture is on a par with being unconscious!

Speaking of happiness and suffering of 'equal intensity', Jamie Mayerfeld maintains that 'the intense suffering would not be compensated by an episode of the intense happiness lasting for a considerably *longer* amount of time.' (1999, 133).[1] This is explicitly a claim about a *moral* asymmetry between suffering and happiness in the intrapersonal domain, so it apparently entails that we could be acting morally wrongly if we intentionally suffer a pain in order to enjoy an equally intense pleasure that lasts longer. I find this baffling.

On the other hand, we've seen evidence of an asymmetry in *inter*personal cases: of there being more of a moral reason to prevent the pain of somebody who's better off than to cause somebody who's worse off to feel a pleasure of the same magnitude, even though the latter would do more to put them on the same welfare level. But it seems there's *no* moral asymmetry *within* lives, *no* moral objection to either undergoing oneself, or letting somebody else undergo, a pain in order to experience an equally great pleasure, let alone a greater pleasure.[2]

I can't find any satisfactory explanation of such a discrepancy between interpersonal and intrapersonal domains. And the simplest view is that, irrespective of whether experiences are positive or negative, the moral (or prudential) value of them parallels the value they have for their subjects in virtue of their intensity and duration. We can stick to this view if we can find a debunking explanation of the intuitive moral negativity bias.

My hypothesis is that, because there's a *psychological asymmetry* to the effect that pains are as a rule more intense than pleasures, evolution could have equipped us with a general tendency to give priority to reducing pain to producing pleasure which we illicitly carry over to cases in which the pleasure we can produce is not only equal, but somewhat greater than the pain we could relieve. Such illicit transferrals of affective and/ or conative reactions against our better judgment are known to occur. For instance, if someone suffers from

[1] G. E. Moore seems to espouse a similar view (1903, 212), and Thomas Hurka develops more sophisticated forms of it (2010, 203-6).

[2] Since I believe such intrapersonal matters belong to the domain of *prudence* rather than morality, I would rather say that there's no *prudential* objection to undergoing the pain.

arachnophobia, this might make them reluctant to touch even spiders they know are made out of some harmless material like rubber.

We should expect that pains can be more intense than pleasures because, first, they're generally signs of bodily damage which could be *irreversible*. This is true of loss of limbs, not to mention death through which all the value of life is lost for good. The bodily well-functioning which pleasure signifies can't bring comparable gains because we'll eventually lose everything we could gain when we die, if not earlier.

Irreversibility is, then, one reason why the extrinsic badness of the harmful conditions behind pain is generally greater than the beneficial conditions behind pleasure, and so why it's more important to reduce pain than to increase pleasure. A second reason for the greater extrinsic badness of harm is that, while loss of capacities *excludes* benefits that could accrue from exercising them, acquisition of capacities doesn't by itself *guarantee* such benefits. This requires in addition advantageous external circumstances, such as good books in the case of the ability to read.

Therefore, it seems evolutionary advantageous for us to be equipped with receptors that enable us to feel pains more acutely than pleasures. But if we're used to pains being more intense and having causes of greater badness than the goodness of the causes of pleasure, this attitude could be transferred to situations in which we compare pains and pleasures that are stipulated to be equal, so that we erroneously judge it to be better to reduce the pains than to induce the pleasures. If this is the origin of our intuitive moral negativity bias, it doesn't support nw-utilitarianism. We could stick to the straightforward, traditional utilitarian view that the intrinsic moral weight of pleasure and pain alike matches their intrinsic goodness and badness for subjects.

Notice next that the psychological asymmetry between positive and negative feelings extends beyond painful and pleasant sensations and the suffering and enjoyment they bring. The negative emotion of *fear* is more widespread and could be considerably more intense than its positive counterpart of *hope* or *longing*: fear could be intensified to *terror* and *horror*, to which there's no counterpart in the case of hope or longing. This isn't surprising for life presents more grave dangers than golden opportunities. For instance, we could die at any moment and lose *everything* life has to offer, or be seriously crippled and lose a major part of it, but there are no comparable gains in store for us.

For the same reason, *sadness* and *sorrow* can be more intense and long-lasting than their positive counterparts, gladness and joy. *Depression* can be paralyzing and debilitating, to which elation can put up no counterpart.

Furthermore, in a world in which most of the time we risk losing more than we could reasonably hope to gain, and in which we compete with each other over scarce resources, it promotes our reproductive fitness if the negative reaction of *anger* is more widespread and stronger than its positive counterpart of *gratitude*, since it's more important to scare off attackers than to return favours rendered by do-gooders. Consequently, it isn't surprising that anger can be stoked up to *fury* and *rage*, but gratitude can't grow correspondingly intense.

Finally, *compassion* with the suffering of others – the emotion that Schopenhauer took to be the basis of morality – is stronger than *sympathy* with the happiness of others. This is precisely what we should expect if we are subject to the moral negativity bias.

Emotions differ from sensations in that they have *propositional objects*: we fear that we'll die, we're angry because we've been insulted, and so on. The propositional objects of emotions could be any state of affairs that's beneficial or harmful for us in some way. We can ask whether emotions are rational since, in virtue of having such objects, they involve beliefs that can be rational. Our emotion of fear is often irrationally strong, as exemplified by phobias like arachnophobia and agoraphobia. Daniel Kahneman provides a further relevant example by drawing attention to *loss aversion* (2011, 282-86) which might manifest itself in our fear of losing something valuable being greater than our hope of gaining something equivalent. A simple illustration of loss aversion is that people generally demand a significantly higher price to sell an item they own than they offer to buy something of the same kind.

The explanation of loss aversion might be that we're disposed to grow attached to things we get to know intimately. It might seem irrational to prefer these well-known things to seemingly indistinguishable things that we don't know intimately, but there might be some justification for this preference: in practice, we seldom know new specimen as well as those that are familiar, so it could reasonably be feared that the new specimen will be inferior in unobvious ways.

Facts already explored provide an evolutionary explanation of why we generally fear losses more than we are attracted by gains. It's a scaring fact about life that losses are often irreversible, while benefits never are. For instance, when we die, we'll be dead forever, forever excluded from the goods of life. When we lose a limb, it can scarcely be restored, so we've lost for good the benefits to which it was a necessary means. A disease, if it doesn't kill us, may mean that we never recover our former good health, but are left disabled, with chronic pain and a lower life-quality in general. These misfortunes may afflict us anytime, so it's of great importance that we're on our guard. In contrast, gains will eventually be 'reversed',

lost or consumed: whatever assets our genes or fortunate external circumstances enable us to collect, we're destined to lose eventually.

It was also observed that the loss of an ability *rules out* that we'll enjoy the benefits its exercise could bring, while acquisition of an ability normally *doesn't ensure* that we'll enjoy the benefits that its exercise could bring: advantageous external circumstances are usually also necessary for such enjoyment. These two factors imply that the extrinsic badness of losing an ability is usually greater than the extrinsic goodness of gaining the ability. Further, the loss of an ability will often occasion pain, whereas the acquisition of an ability, e.g. by training, will typically not occasion pleasure. So no wonder if evolution has wired us up to be in general more strongly averse to losses than attracted to gains. This could spill over to cases in which we have every reason to believe that there's no evaluative difference.

Summing up this discussion of the moral negativity bias, the debunking explanation proposed is that it's a tendency with which we've been equipped because there's a psychological asymmetry to the effect that negative feelings are normally more intense than positive feelings. Generally, this asymmetry is justifiable, since negative feelings mark conditions whose extrinsic badness is greater than the goodness of conditions marked by positive feelings, though there are exceptions to this rule, illustrated by loss aversion. Since our tendency to prioritize reducing what's bad, because it's generally greater, is so deeply ingrained, it's extended to situations in which the good is stipulated to be as great as the bad.

However, I'm far from certain that this debunking explanation is complete and correct in all details, though it seems certain that an explanation should refer to the psychological asymmetry highlighted. Nonetheless, the fact that there's at least in outline a debunking explanation of the moral negativity bias is reassuring, since it's hard to accept it at face value as a foundation for nw-utilitarianism because of the intuitive discrepancy between intrapersonal and interpersonal cases.

III The Moral Negativity Bias, and Prioritarianism vs. Egalitarianism

Our inclination to exhibit the moral negativity bias can be suspected of having served to support not only nw-utilitarianism, but also prioritarianism. For example, Derek Parfit (1995, note 35) praises Joseph Raz for putting the difference between egalitarianism and prioritarianism well in the following passage:

> What makes us care about various inequalities is not the inequality but … the hunger of the hungry, the need of the needy, the suffering of the ill, and so on. The fact that they are worse off in the relevant respect than their neighbours is relevant. But it is relevant not as an independent evil of inequality. Its relevance is in showing that their hunger is greater, their need more pressing, their suffering more hurtful, and therefore our concern for the hungry, the needy, the suffering, and not our concern for equality, makes us give them priority. (Raz 1986, 240)

Notice that Raz talks about bad states, states that arouse our compassion: "the hunger of the hungry, the need of the needy, the suffering of the ill". He claims it's the fact that "their hunger is greater, their need more pressing, their suffering more hurtful" instead of the fact that they may be worse off than somebody else that "makes us give them priority". Consequently, Raz could be read as appealing to the moral negativity bias: the greater moral urgency of reducing what's bad for individuals instead of the prioritarian idea of attaching greater moral weight to improving the situation of the worse-off, regardless of whether this improvement consists in increasing goodness or reducing badness. Benefits to the worse-off could be thought to have greater moral weight because they're more likely to consist in reducing what's intrinsically bad for individuals than boosting what's intrinsically good.

So, when prioritarians argue against the egalitarian reference to worseness relative to others, we should check whether they're relying on intuitions that could be accounted for by the moral negativity bias. Contrast the following two kinds of case. If there are individuals who are very badly off, e.g. who are very hungry, there's clearly a strong moral reason to relieve their hunger, which is indeed bad for them. Now it mightn't be obvious that this reason is strengthened if it's added that there are other individuals who, unjustly, are less hungry, though egalitarianism implies this. Thus, this kind of case – Case 1 – appears to offer comfort to prioritarians against egalitarians.

Consider instead – Case 2 – individuals who are well off but not very well off, say, they have enough wine, but of a rather mediocre reserva sort. The moral reason to increase their enjoyment by providing them with gran reserva wine seems relatively weak. In opposition to egalitarians, prioritarians maintain that this reason isn't strengthened by the addition of another population who's unjustly better off by having access to gran reserva wine. But, at least according to my intuition, the egalitarian view that this additional population strengthens the case for providing the reserva people with better wine is more plausible. Of course, prioritarians could simply reject this intuition, but it does seem that when recipients are quite well off and benefiting them consists in injecting intrinsic goodness, it's more difficult to deny that the presence of individuals who are unjustly even better off strengthens the moral reason to benefit those less well off.

Prioritarians have a hard time explaining why the addition of the better-off apparently amplifies our moral reason to benefit the worse-off in Case 2. On the other hand, egalitarians have an explanation why this egalitarian reason is seemingly absent in Case 1 when the worse-off of the two populations is very badly off and benefiting them consists in reducing what's bad, namely the presence of a strong reason deriving from the moral negativity bias which 'drowns' it. This bias isn't at work in Case 2 because there's nothing intrinsically bad about the condition of the worse-off population here. As this bias has been debunked, egalitarians could happily appeal to it to undercut support for prioritarianism. By contrast, it would be awkward for egalitarians as well as prioritarians to accept this bias as evidence for nw-utilitarianism because it counteracts both views, for instance, by advising us to alleviate the pain of the better-off rather than augmenting the pleasure of the worse-off. For adherents of egalitarianism, like myself, it's important both that the moral negativity bias can be debunked and that it can still be employed to undermine support for prioritarianism.

References

[1] Hurka, Thomas. 2010. "Asymmetries in Value." *Noûs* 44: 199-223.

[2] Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. London: Allen Lane.

[3] Mayerfeld, Jamie. 1999. *Suffering and Moral Responsibility*. New York: Oxford University Press.

[4] Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

[5] Parfit, Derek. 1995. *Equality or Priority? The Lindley Lecture, 1991.* Lawrence, KA: University of Kansas.

[6] Popper, Karl. 1966. *The Open Society and Its Enemies*. 5th edition. London: Routledge & Kegan Paul.

[7] Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

[8] Schopenhauer, Arthur. 1995. *On the Basis of Morality*. Translated by E. F. J. Payne. Providence, RI: Berghahn Books.

About the Badness of Existence and the Prospect of Extinction

Giuseppe Rocché, University of Palermo, Italy

Abstract

In this paper I consider the case of people who find The Mere Addition Principle counterintuitive. Their particular intuitions may be understood as instances of The Principle of Intrinsic Disvalue of Existence (PIDA). Following this idea, Contractualism seems to be an appropriate method to solve population ethics dilemmas. Still, I show that their rejection of The Mere Addition Principle – if understood as an instance of PIDA – is not enough to avoid these dilemmas and to reach a stable equilibrium among their intuitions. In fact, if their denial of The Mere Addition Principle is grounded on PIDA, the consequences which would follow are likely to be unacceptable for many of them. In particular, either they hold that we have a duty in favor of extinction, or they cannot take PIDA seriously enough. Rejecting the Mere Addition Principle without endorsing PIDA seems the best they can do in order to reach a stable equilibrium among their intuitions.

According to the *Mere Addition Paradox* (Parfit 1984, 419-40; 2004) we cannot consistently hold both

(i) *The Mere Addition Principle*: if a number of people with positive wellbeing is added without affecting the original people's wellbeing, the resulting population (A+) is at least not worse than the original population (A) (Arrhenius 2000, 250)

(ii) *Non-Antiegalitarianism:* considered two populations of the same size, if in the first (B) there is both a higher average wellbeing and more equality than in the second (A+), the first is better than the second (Ng 1989, 238)

(iii) *The Principle of Transitivity and Substitution* (Temkin 1987, 143-44)

and

(iv) *The Denial of the Repugnant Conclusion* – a population (A), of at least ten billion people, in which all its members have a high level of wellbeing is better than a much bigger population (Z) in which all its members have lives that are barely worth living.

Just what a solution to the Mere Addition Paradox consists in is a contentious issue. Speaking of our intuitions about A+ and A, Parfit said that

> To avoid the paradox we must believe, without considering the rest of the argument, that A+ is worse than A. […] To the extent that we find this hard to believe, we still face a paradox (Parfit 1984, 428; 2004, 16)

These words may be taken to mean that we look for a *psychological* solution: solving the puzzle is finding a principle that – given our present attitudes – is *intuitive* (a substantive methodological choice as far as it rules out the use of *bullet biting* – a strategy that Parfit himself endorsed, see Parfit 2016, 120). In this framework, given a clash among the aforementioned four judgments, we settle the issue by showing that, despite the appearances, we do not find one of them really intuitive. In this paper, I propose to wonder whether we feel psychologically compelled by The Mere Addition Principle. Is The Mere Addition Principle really engraved in our psychological outlook? Parfit exposed some natural properties of a Mere Addition – namely that additional people's lives are worth living and that they do not lessen other people's wellbeing. Now we should ask whether we happen to care about these natural properties or not.

Some authors take on this challenge arguing that we attach *intrinsic* value to the existence of additional people. Hints of this axiological attitude are – for example – that we think it would have been bad if the happiest share of humanity of the past had never been born (Rachels 1998, 103); that we think would have been a terrible loss if some land on earth had never been populated (Ord 2014, 51); that we would regret being alone in the universe (Tännsjö 2004, 231-32; Rachels 1998, 103). Without discussing these arguments at length, I point out that in some cases our alleged recognition of the intrinsic value of existence is affected by what we may call – in the absence of anything better – *aesthetic* features of lives (Sumner 1996, 21-23). Insofar as we aim at a comprehensive ranking of possible populations, aesthetic features should be taken into consideration, but if we are interested just in *welfarist* axiology, we should neglect them. My proposal is to ask whether we attach value to the creation of people with worthwhile but totally anonymous – devoid of aesthetic value – lives.

Imagine a technological or biological machine capable to create additional people with positive wellbeing without affecting anyone. The machine is currently turned off, but we can easily turn it on so that it will start creating new anonymous lives. These will be long lives with many pleasures and some peaks of sheer bliss, interrupted by some pains, pain will increase as the end gets closer. What to do in this case? I conjecture that I would not turn the machine on. Without holding that this would be the reaction of many people, I assume

that there is a moral tribe[1] which would share my same reaction. The question is whether the members of this tribe have a set of intuitions which avoid population paradoxes.

The refusal to activate the machine could be explained on the basis of different principles, here I shall understand it as an instance of the idea that existence is bad – call it *The Principle of Intrinsic Disvalue of Additions* (PIDA). According to PIDA mere additions are all things considered bad, whereas additions are *prima facie* bad but their badness can be counterbalanced by other goods, above all the *instrumental* value of additions for existing people. Now a problem arises, i.e. that the Repugnant Conclusion is implied even though we recognize that additions of people with a positive wellbeing are intrinsically bad. For we can easily imagine additions, dubbed *Benign Additions* (Huemer 2008), which raise the wellbeing of the original population but lower the average wellbeing of the total population.

Contractualism – choice under a veil of ignorance – offers an answer to this predicament. Imagine (Case 1) you have to choose in self-interested terms whether to be member of a world in which people have a welfare level of 100 or of a world in which those people have a welfare level of 105 and many more people have been added whose lives are barely worth living (level 5). It may be argued (Tännsjö 2004, 211-12) that a self-interested decision-maker in different-number choices has to consider the risk not to exist at all and, therefore, could have reasons to choose the bigger population with a lower average wellbeing. Still, if PIDA is our starting point, parties rather than being averse to the risk to not exist, would be averse to the risk to exist. This idea does not imply the very radical conclusion that (Case 2) given two populations the first with a very high wellbeing, the second with a very low wellbeing but slightly smaller than the first, parties would choose the second world. Parties consider existence as a risk in itself, but not every existence is equally risky and they may be more averse to a low risk of existing and having a life which is barely worth living, than to a high risk of existing and having a life which is well worth living. Then in Case 1 parties would choose the less populated world with higher average wellbeing, whereas in Case 2 they would go for the more populated world.

How much parties – who have been constructed on the basis of PIDA – want to avoid the risk of existing? Different answers to this question shape different conceptions of Contractualism in population ethics. A possible conception of Contractualism is the use of what we may call *"The Same-Number Restriction"* (SNR). According to SNR, parties who are about to choose between different-number scenarios – alternative populations of different sizes – ignore this feature of their choice, so that they think to be choosing between same-number scenarios. In Case 1 parties are facing the choice between two worlds one of them much

[1] I borrow this expression from Greene 2013.

more populated than the other. Anyway, they ignore this fact and know just that in the first world everyone has a very high level of wellbeing, whereas in the second a tiny share of people is even better-off than in the first world but the vast majority has barely worth-living lives.

SNR gives horrible results when negative well-being is concerned. This can be easily proved through Parfit's *Hell 1* and *Hell 2* thought experiment (1984, 392): a world in which a handful of people is suffering hellish torments for fifty years would be worse than a world in which billions of people are suffering the very same torments for fifty years minus a day. In fact, under SNR parties would ignore the fact that in the second world many more miserable people exist. Still, these problems may be avoided by introducing some exceptions into SNR. I do not discuss this point here.

Other cases pose problems which are harder to be solved by means of exceptions. Imagine that – Case 3 – you have to choose between a world in which everyone has a very high welfare level – say 100 – and a world in which the welfare level of the people of the first world has been raised – to, say, 110 – and some people with a welfare level even higher – say, 111 – have been added. We can call these cases *Fair Benign Additions* (FBA), additions in which original people's well-being is raised, so it is the average well-being, and – to exclude cases like Case 1 – additional people have a positive welfare level higher than original people's welfare level after the addition. For many people, intuitively, FBA are never bad, moreover thinking that they are never bad does not entail The Repugnant Conclusion – insofar as the result of FBA is an increase of average well-being. For many people their intuitions about FBA are unproblematic because they do not yield counterintuitive results – as Mere Additions do – when they are put together with their other intuitions.

Still, if we think that existence is somehow bad – see PIDA –, we may also think that some FBA are bad. Imagine – in Case 3 – that the number of additional people with a well-being level of 111 is huge. If – following SNR – we think that the addition should be performed, our endorsement of PIDA is somehow shaky. Actually, it seems one of the weakest endorsement possible. We would recognize that it would be good adding every number of people in order to slightly improve the well-being of an already existing person – by, say, giving him an additional lollipop –, when the additional people are better-off than him. In other words, SNR mirrors the psychological outlook of those who are averse to the risk of existing but are *lexically* more averse to the risk of existing and living a life which is worse than the life they

could have lived[2]. If you endorse PIDA but not this lexical priority, then you should drop SNR.

Hence, if you think that you would refrain from adding billions of people – whose lives would be full of pleasures but also with some pains, especially at their end – when by doing so the well-being of an existent person would be slightly improved – an additional lollipop –, then you feel the need to take the badness of existence – PIDA – more seriously than how SNR implies. Now a major risk is that, once we have found a principle which takes more seriously the idea of the badness of existence, we are bound to accept that we have the duty to stop procreating and to cooperate to realize the extinction of mankind. Some people think we have this duty. This duty is for them intuitive (Benatar 2006, 207), even though they pretend to prove it by means of considerations other than intuitiveness – namely, *bullet biting* and *evolutionary debunking arguments* (Benatar 2006, 202-7). Other people – who endorse PIDA – lack the intuition that we have a duty to bring about the extinction of mankind. A relevant question is whether these people's intuitions are unstable like those of people who accept The Mere Addition Principle but want to avoid the Repugnant Conclusion.

Perhaps a solution could be provided by a plausible description of what would happen if we head towards extinction. The process realistically would determine that most of last people's lives would have negative well-being: the prospect of being the last people in the universe fills many of us with anguish. According to a *moderate* lexical account, the badness of FBA can overwhelm the disvalue of a decrease in existing people's well-being when their well-being remains anyway high and positive, but the badness of FBA cannot overwhelm the disvalue of the production of tormented lives – as we postulate last people's lives would be.

Even this proposal may be unsatisfying for those who have the intuition that existence is intrinsically bad. Imagine a case – Case 4 – in which just a handful of people exists. If they do not reproduce, then they will have tormented lives. If they have children, their lives will be worth living but they will create billions of people whose lives will be barely worth living. In a case like this, taking PIDA seriously seems to imply that the relevant gain in existing people's wellbeing cannot make up for the huge disvalue of the addition of billions of people with lives barely worth-living.

[2] Their attitude must not be confused with *leximin.* We can imagine three welfare level W1 slightly higher than W2, and W2 slightly higher than W3 and two populations P1 (W2) and P2 (W1,W3). According to the lexical principle expressed by SNR, parties could have reasons to choose P2 if the risk to have welfare level W3 is very low – because the corresponding subset of population is very small.

As a last option, perhaps we could cling to a peculiar feature of our condition. In general, we may grant that for every population with a certain welfare level, there is a number of possible people with a certain welfare level whose addition would be optimal. The more we add beyond that number – or we fail to add under that number – the worse is the outcome. In Case 4 there is a large relevant[3] disproportion between the number of additional people and the number of existing people whose well-being is positively affected by the addition. But it may be argued, this is not our condition. To avoid to live tormented lives we are not required to create thousands of billions of people. We need just to add roughly as many people as we are: we need just to secure our replacement. Then, in cases like Case 4, where there is a relevant disproportion between additional people and existing people, extinction ought to be chosen; if, on the contrary, a generation can avoid a painful extinction by just "replacing" itself, this is what that population should do. Because our condition resembles this latter case, we have no duty to cause our extinction – on the contrary we should have children.

Many people would find this reasoning nothing but a cunning casuistry. It is somehow true that a generation is directly responsible only for the addition of the next generation, but as far as we can foresee that our successors will have the same bitter alternative between a painful extinction and the addition of another generation, we are indirectly responsible for the addition of that generation as well. Because we will be indirectly responsible for their successors, and for the successors of their successors – and so forth –, our condition does resemble Case 5 in which there is a stark disproportion between the number of additional people and the number of existing people whose well-being is positively affected by the addition.

Concluding, the members of the moral tribe, who share the intuition that would be bad to turn the machine on, may have discovered that their set of intuitions is as shaky as that of the supporters of The Mere Addition Principle. In fact, *if* they understand their response to the machine thought experiment in terms of the endorsement of PIDA, they are forced to accept principles like the lexical account, the moderate lexical account, or the restriction of our moral responsibility to our direct responsibility – the aforementioned "cunning casuistry". If, on one hand, they are not glad to accept any of these principles, and, on the other hand, they are not psychologically led to revise their judgment in the machine case, they seem to be in a deadlock. My suggestion is that they should look upstream for an understanding of their response in that case different from PIDA. What about the case in which

[3] I.e. adjusted considering the welfare levels involved.

the machine would be operative and we could turn it off? If they conjecture that they would leave it working, can they make any sense of this answer?

References

[1] Arrhenius, Gustaf. 2000. "An Impossibility Theorem for Welfarist Axiologies." *Economics & Philosophy* 16: 247-66.

[2] Benatar, David. 2006. *Better Never to Have Been.* Oxford: Clarendon Press.

[3] Greene, Joshua. 2013. *Moral Tribes.* New York: The Penguin Press.

[4] Huemer, Michael. 2008. "In Defence of Repugnance." *Mind* 117 (468): 899-933.

[5] Ng, Yew-Kwang. 1989. "What Should We Do About Future Generations? Impossibility of Parfit's Theory X." *Economics & Philosophy* 5 (2): 235-53.

[6] Ord, Toby. 2014. "Overpopulation or Underpopulation?" In *Is the Planet Full?,* edited by Ian Goldin, 46-60. Oxford: Oxford University Press.

[7] Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

[8] _____. 2004. "Overpopulation and the Quality of Life." In *The Repugnant Conclusion,* edited by Jasper Ryberg, and Torbjörn Tännsjö, 7-22. Dordrecht: Kluwer Academic Publishers.

[9] _____. 2016. "Can We Avoid the Repugnant Conclusion?" *Theoria* 82 (2):110-27.

[10] Rachels, Stuart. 1998. "Is it Good to Make Happy People?" *Bioethics* 12 (2): 93-110.

[11] Sumner, Leonard W. 1996. *Welfare, Happiness, and Ethics.* Oxford: Clarendon Press.

[12] Tännsjö, Torbjörn. 2004. "Why We Ought to Accept the Repugnant Conclusion." In *The Repugnant Conclusion*, edited by Jasper Ryberg, and Torbjörn Tännsjö, 219–38. Dordrecht: Kluwer Academic Publishers.

[13] Temkin, Larry S. 1987. "Intransitivity and the Mere Addition Paradox." *Philosophy and Public Affairs* 16 (2): 138-87.

Sidgwick, Reflective Equilibrium and the Triviality Charge

Michael W. Schmidt, Karlsruhe Institute of Technology, Germany

Abstract

I argue against the claim that it is trivial to state that Sidgwick used the method of wide reflective equilibrium. This claim is based on what could be called the Triviality Charge, which is pressed against the method of wide reflective equilibrium by Peter Singer. According to this charge, there is no alternative to using the method if it is interpreted as involving all relevant philosophical background arguments. The main argument against the Triviality Charge is that although the method of wide reflective equilibrium is compatible with coherentism (understood as a form of weak foundationalism) as well as moderate foundationalism, it is not compatible with strong foundationalism. Hence, the claim that a philosopher uses the method of wide reflective equilibrium is informative. In particular, this is true with regard to Sidgwick.

Introduction

This paper contributes to the debate whether Sidgwick used the method of reflective equilibrium. "Reflective equilibrium" is the name of the method of justification which John Rawls suggests in his *A Theory of Justice* (1971). He claims that many other philosophers have used the method even before: Most prominently Nelson Goodman in *Fact, Fiction, and Forecast* (1955) and Henry Sidgwick in *The Methods of Ethics* (1907). Whereas Catherine Elgin – a scholar sympathetic to Goodman – endorsed Rawls's suggestion and even worked out a better understanding of the methodology, from the very beginning there was a dispute over the claim that Sidgwick employed such a method. I argue that Rawls's claim that Sidgwick used the method can be – in some sense – defended against certain strong criticisms.

The following is divided into three parts. In the first part of this paper, I will provide some important background information: I will do that by referring to an article by Peter Singer published 1974 entitled "Sidgwick and Reflective Equilibrium" in which he argued that reflective equilibrium is neither an adequate method of justification nor the method Sidgwick did employ. I will reconstruct Singer's interpretation of Sidgwick and his argumentation against Rawls's claim.

This sets the ground for the second part. I will begin by sketching the reasons why Singer slightly revised his earlier criticism of the method of reflective equilibrium in his latest works. After considering the now predominant wide interpretation of the method Singer

now holds that it is indeed possible to claim that Sidgwick used it – but he still would refrain from saying so, because of one more charge he presses against the method and the claim: the charge that if it is understood in the wide sense, it is simply trivial to state that a philosopher makes use of it. According to the charge there is no alternative to the method of reflective equilibrium, if it is interpreted in such a wide way that its use involves all relevant philosophical background arguments, because then it includes all other rival methods. So, to state that someone uses the method settles nothing and is pointless. This is what I dubbed the *Triviality Charge*.[1]

In part three I will assume that the wide interpretation of the method of reflective equilibrium, which Singer considers compatible with the method of Sidgwick, is the only plausible interpretation. I will argue that it is – even in the wide interpretation – not trivial to state that some philosopher and especially that Sidgwick did use it. Hence, by refuting the claim of the *Triviality Charge*, I will argue that it is informative and justified to state that Sidgwick used the method of reflective equilibrium, even if one takes granted that Singers interpretation of Sidgwick's *Methods of Ethics* is the correct one or grasps the important methodological points adequately.

I Different Interpretations: Sidgwick and Reflective Equilibrium

According to Singer's interpretation in "Sidgwick and Reflective Equilibrium", Sidgwick proposes a top-down approach in the realm of normative ethics to justify moral propositions: one has to start with self-evident axioms (in the form of universal principles) and to see what follows from these. If our everyday moral judgments cohere with the ethical theories derived from the axioms, this can be used as an argument to convince common people to adopt the justified ethical theories – but this doesn't show the justification of these theories, since they themselves are only justified by their status as being inferentially connected to the self-evident axioms, which have a privileged epistemic status. Yet, still one can be mistaken in holding an apparent axiom to be a real axiom. In terms of Laurence Bonjour widely used in epistemology one can call this a moderate foundationalism, where some basic beliefs – here the self-evident axioms – are themselves justified without being inferred

[1] I am borrwoing the name "Trivility Charge" from Julia Langkau. Langkau argues contra the charge against reflective equilibrium in a different (non-moral but epistemic) context and takes a different line of argument, though one could say, we share the same strategy. Cf. Langkau 2013.

from other beliefs. They can pass on by inference relation the justification to other beliefs and this suffices, given that we have true beliefs, that we also have knowledge – in our case moral knowledge (cf. Singer 1974, 498-501, 503-5, 507-8; cf. BonJour 1985, 26-30).

So, according to Singer, the basic beliefs in Sidgwick's moderate foundationalism are on the most abstract level of moral entities: Sidgwick's basic beliefs concern ultimate ethical principles – namely the *principle of justice*, the *principle of prudence* and *the principle of benevolence* –from which the morally right theories are deducible. So, in order to choose the right moral theory, it is essential to have an intuitive insight in the self-evidence of the axioms and to check if the self-evidence was merely apparent self-evidence by reflecting if there are any other self-evident axioms that conflict with the one under investigation, if there is an consensus on the axioms and if the principles corresponding to the axioms are ambivalent or precise. As I see it, this interpretation of Sidgwick – which I tried to reconstruct here in a condensed form – remains Singer's interpretation and hasn't changed substantially in the other works I will refer to (cf. Singer 1974, 503, 507-8; cf. Sidgwick 1907, Book III, esp. Chapter XI, 2, and Chapter XIII).[2]

He contrasts this moderate foundationalism in the realm of normative ethics with an interpretation of the method of reflective equilibrium. The basic idea of reflective equilibrium is that a theory and our common sense considered judgments should be brought into agreement. And if they both support each other in the best available way the judgments as well as the corresponding theory are justified. Both are also open to revision in the process of adjustment.

Although sometimes the method of reflective equilibrium is accused of being a form of disguised common sense-intuitionism – which means that it is a bottom-up moderate foundationalist approach that presupposes that one can do ethics analogue to (some common interpretations of) empirical inquiry or science – mostly it is recognized as a form of coherentism, as it is by Singer. Importantly Singer seems to imply that the use of the method of reflective equilibrium would result not only in a coherence account of justification but also a coherence account of validity or truth (cf. Singer 1974, 492-5, esp. 493-4).

Rawls, according to Singer, thus misinterprets Sidgwick when he suggests that they share the same method of justification. He thinks that this misinterpretation rests on the passages where Sidgwick tries to show that the utilitarian theory, which can be derived from the

[2] There are, of course, other interpretations: Rawls himself refers to Schneewind 1963. Skelton 2010 backs some of the points that lead to a rejection of the claim that Sidgwick used the method of reflective equilibrium, Crisp 2002 is as well critical on the suggestion that he used this method, but on different grounds, Sverdlik 1985 and Brink 1994 have interpretations that would in contrary back the claim of the direct use of the method.

principle of benevolence together with *the principle of justice*, is fitting best to the judgments of our common-sense morality But this *ad hominem* argument – as Singer calls it – (respectively the coherence with our common sense judgments) is not what justifies this axiom (or any other). What justifies all possible axioms is, that they are self-evident and remain self-evident after due reflection.

Now if the method of reflective equilibrium were to be understood in the way Singer suggested, and one accepts Singer's interpretation of the *Methods of Ethics* it would clearly be inadequate to state that Sidgwick used the method of reflective equilibrium. Sidgwick, according to Singer, would have used a top-down moderate foundationalism and believed in objective moral truths, whereas Rawls would have used a coherentist approach, that includes not only a coherence account for justification but also for truth and thus he would be a subjectivist or cultural relativist concerning moral truth, so that their methodology is not consistent at all.

II Singer's Revised Position

This brings us to part two of this paper. We can begin by asking the question: Was Singer right with this interpretation of reflective equilibrium? Most often reflective equilibrium is – as Singer rightly suggested – indeed considered a coherentist method of justification. But typically, this involves a form of weak foundationalism, which means that while there are no beliefs which themselves are already justified without inferential backing, there are some which have an initial credibility, because they are what we in fact believe before we start to scrutinize and criticize our system of beliefs. This initial credibility is – according the weak foundationalist interpretation of coherentism (which I want to presuppose henceforth) – not enough to grant an inquirer knowledge. Justification thus arises only if beliefs – initially held or not – can be incorporated in a system of held beliefs in the most coherent way such that they are mutually backed by inferential relations better than in any alternative system of beliefs that one could accept (cf. Rawls 1974, 8).

But although reflective equilibrium is widely understood as a coherence method of justification, that doesn't mean that a proponent of it must embrace a coherence account of truth: There can be objective moral truth that is not created nor secured by a coherent

system of moral beliefs even if one thinks that this is what justifies these beliefs.[3] Most proponents of the method of reflective equilibrium take coherence (combined with initial credibility of the beliefs that we actually accept) as a criterion of justification but not the criterion of truth and thus admit, contrary to Singers earlier assumption, that there is (moral or non-moral) truth or objectivity independent of its subjective justification.

One of the most detailed accounts of reflective equilibrium originates from Norman Daniels (1996). He builds his account on the distinction between a narrow reflective equilibrium and a wide reflective equilibrium.[4] If one is trying only to achieve a narrow reflective equilibrium one just tries to achieve coherence only between our considered judgments and theories. But according to the method of reflective equilibrium in the wide interpretation – which we should use in philosophy – one has to incorporate in the weighing process all relevant background theories and arguments. These background theories are scientific or philosophic theories or arguments that would have an impact on the narrow reflective equilibrium, were they to be considered.

Thus, the method of reflective equilibrium in the wide sense is a method that enables us to be critical of our judgments and scrutinize our biased system of belief. It is also wide enough for background theories, that mandate for special areas of investigations special sub-methods. If, for example, a plausible theory casts doubt on our common sense moral judgments, it could be possible to discredit these judgments systematically in moral inquiry – just as Singer himself holds – with the possible result (if it also can be argued for the remaining beliefs to be basic) that one establishes for the realm of normative ethics a moderate foundationalism. In this way, wide reflective equilibrium is indeed compatible with the method of Sidgwick the way Singer interpreted it.

Weak foundationalism – according to this interpretation of reflective equilibrium – remains the "default setting" for inquiry unless a different sub-method is vindicated for certain areas of investigation.

Also, Singer himself, who, of course, did follow the debate on reflective equilibrium, now explicitly accepts in his 2005 article "Ethics and Intuitionism" and his book *The Point of View*

[3] Indeed, Rawls himself leaves room for the idea that if we use the method of reflective equilibrium it might result in a convergence of out ethical belief systems what could indicate that we are getting closer to moral truths. That implies on the other hand, that we still can go wrong, even if we have reached a reflective equilibrium. Cf. Rawls 1974, 9-10, 21. Cf. Daniels 1996, 33-40.

[4] According to Rawls' terminology in his article "The Independence of Moral Philosophy" (1974) – Rawls suggested the use of wide reflective equilibrium already in *A Theory of Justice*. Cf. Rawls 1971, 49; 1993, 8-9.

of the Universe – jointly written with Katarzyna de Lazari-Radek (2014) – that the method of wide reflective equilibrium is indeed compatible with Sidgwick's method.[5]

> Admittedly, it is possible to interpret the model of reflective equilibrium so that it takes into account any grounds for objecting to our intuitions, including those that I have put forward. Norman Daniels has argued persuasively for this 'wide' interpretation of reflective equilibrium. If the interpretation is truly wide enough to countenance the rejection of all our ordinary moral beliefs, then I have no objection to it. (Singer 2005, 347)[6]

II.a The Triviality Charge

So, what is stopping us – granted that we do agree with Singer's Sidgwick interpretation – from simply stating that Sidgwick indeed used something like the method of wide reflective equilibrium? According to Singer there is a price to be paid, if the method of reflective equilibrium is understood as wide, as Daniels suggests:

> The price for avoiding the inbuilt conservatism of the narrow interpretation, however, is that reflective equilibrium ceases to be a distinctive method of doing normative ethics. Where previously there was a contrast between the method of reflective equilibrium and ''foundationalist" attempts to build an ethical system outward from some indubitable starting point, now foundationalism simply becomes the limiting case of a wide reflective equilibrium. (Singer 2005, 347)

Singer claims that if the use of the method of reflective equilibrium is no means anymore to distinguish a priori, that some moral philosopher proposes rather a coherentism than a moderate foundationalism for ethical inquiries, then stating that the philosopher used the method of reflective equilibrium becomes meaningless. It's just trivial to state, that a philosopher used the method of reflective equilibrium understood in this wide sense, that includes the possibility of an ethical moderate foundationalism – so to say that Sidgwick used the method of reflective equilibrium is pointless and we shouldn't state pointless utterances (at least, this seems to be implied). This is the core of the triviality charge:

> […] whether wide reflective equilibrium and foundationalism can be distinguished depends on the substance of 'the acceptable moral theory' and on what the philosophical arguments allow us to conclude. Without knowing which moral theory is acceptable and whether there are

[5] He accredits the notion of wide reflective equilibrium to Daniels and not to Rawls – and I think he is mistaken here, Cf. Singer and de Lazari-Radek 2014, 11-114; cf. Singer 2005, 347.

[6] One could interpret the cited statement in a way that Singer now accepts a weak foundationalism in which certain beliefs are epistemically devalued, but I suggest that Singer's position is still a moderate foundationalism.

> philosophical arguments that reveal which moral judgments are objectively true, we cannot exclude the possibility that, once we have found the soundest moral theory and the best philosophical arguments, we will be able to demonstrate that none, or virtually none, of our existing moral judgments are credible; such that we can confidently reject all, or virtually all, of our current moral judgments, and replace them with the judgments that follow from the moral theory. […] In that case, the distinction between wide reflective equilibrium and foundationalism has narrowed to a vanishing point. It would then be true, but trivial, that when we do normative ethics, there is no alternative to the method of reflective equilibrium. There would be no alternative because wide reflective equilibrium is so wide that it includes all possible methods, including foundationalism. (Singer and de Lazari-Radek 2014, 112-3)[7]

III Arguments against the Triviality Charge

I will argue that it is not trivial to state that Sidgwick used the method of reflective equilibrium (understood as the attempt to achieve a wide reflective equilibrium).

To support this thesis, I will advance one main argument and two further arguments.

1. The main argument takes that *fallibilism* is a necessary condition for being able to support the method of reflective equilibrium. Fallibilism is roughly defined here as the presupposition that all beliefs without exception which we could use in our philosophical argumentation are in principle questionable and open to scrutiny since the epistemic agent cannot be certain of their truth. Why is fallibilism a necessary condition for the method of reflective equilibrium? If there were some beliefs which were not open to scrutiny, they would be totally fixed points in the process of adjustment when conflicts occur in our system of belief. And everything that could be brought in deductive inferential relation with them were as well as fixed (given that we accept some set of rules of logic as one of these fixed points). But fixed points are not open to weighing considerations if they conflict with other beliefs in a system that is to be brought in a reflective equilibrium. All the balancing process that lies at the core of the process of achieving reflective equilibrium is only possible if we accept fallibilism. Yet, fallibilism isn't the only condition for the method of reflective

[7] A similar position is held by Sem de Maagt: "One problem is that by including just about any possible disagreement related to the justification of our moral beliefs into its methodology, reflective equilibrium runs the risk of becoming vacuous as a method of moral justification, because ultimately reflective equilibrium will simply be reduced to reasoning about ethics in general. That is, if any kind of disagreement is included in the search for reflective equilibrium it is not evident that it still can function as a method of moral justification." (2017, 458).

equilibrium: A second further necessary condition is a (weak) holistic perspective of justification which includes inferential interdependence. A third condition consist in criteria of rational belief revision in case of inconsistent beliefs in a holistic system of beliefs – namely that a belief succeeds over another conflicting one in the case it has (in light of all supportive evidence) a higher degree of credence than the other belief.[8] A fourth condition would be the weak foundationalist assumption that all the beliefs we think to be true – the "considered judgments" or "commitments" – have an initial weak credibility that suffices to distinguish them from merely possible beliefs but that is not strong enough for granting us knowledge without further inferential backing.

But to challenge the triviality charge, I will only rely on fallibilism.

Now there are, at least, two reasons why it is informative – rather than trivial – to state that Sidgwick was a fallibilist:

i. The first reason is obvious: There are philosophers who are infallibilists and think that they have reached some unquestionable truths. Singer himself points to Descartes to illustrate this position from which he wants to distinguish Sidgwick's position. This means obviously that Singer also thinks that pointing to Sidgwick's fallibilism is informative. In fact, if one wanted to classify that kind of infallibilism that Descartes seems to present, one would call it – according to BonJour – strong foundationalism.[9] So it is literally false to simply state that foundationalism might be included in the method of reflective equilibrium: Some form of foundationalism might not be included, which shows us again, that to assign the method of reflective equilibrium to some someone is informative. Granted – nowadays most philosophers seem committed to some sort of fallibilism but there still might be some who argue for infallible truths – in ethics as well as in other areas of philosophy. So even with the majority of contemporary philosophers being fallibilists it is not completely trivial to state that some philosopher is a fallibilist. This already counts for contemporary philosophers, but Sidgwick is a historic philosopher, so the claim that he used the

[8] This should normally lead to a maximal coherent set of beliefs, what could in itself be counted as another condition for reflective equilibrium.

[9] One could argue that Descartes only seems to present a strong foundationalism and if we were to reconstruct his position with the principle of charity in mind it would turn out that he too was a fallibilist. This might hold for other philosophers as well. If so, the claim that the method of reflective equilibrium is the method of philosophy could be true (generally speaking) – I leave this open to further investigation. But even if this were correct, it still would be informative to state that these philosophers were fallibilists because one had to debunk their apparent infallibilism and theoretically infallibilism would still be an option.

method of reflective equilibrium is even more informative and interesting, if we rightly can assume that there have been more infallibilists in the past (Crisp 2002, 60-63).

ii. The second reason is an exegetic one: In Sidgwick's *Methods* the notion of self-evident axioms plays – as it is well known – a fundamental role. But how is the notion of self-evidence understood in Sidgwick's *Methods of Ethics*? To state that the axioms he proposes are based on a self-evident intuition (plus reflection) *but* still are fallible is an informative interpretation.[10]

This is my main argument and I would like to turn to two additional arguments against the Triviality Charge

2. The second argument concerns the meta-level on which the methodological design of a subdomain, i.e. certain areas of investigation with specific features, is justified. On this meta-level we are operating on coherentist or weak foundationalist standards even if there is a moderate foundationalist standard (or any other fallible standard) justified in the sub-domain. A sub-domain could be for example the area of normative ethics. If one does assume that Sidgwick proposed a moderate foundational method for normative ethics but was using a reflective equilibrium to justify this method on the meta-level of – let's say – metaethics, then on this meta-level he was arguing in a coherentist or weak foundationalist way. I hold that the same pattern is true for externalist epistemologists. Since these claims could be more controversial than the precedent, I would like to build upon it only an additional argument against the *Triviality Charge*. Yet I think, it might be the philosophically more interesting claim. There seem to be – at least – two non-trivial statements connected with this claim:

   i. The arguments by which a moderate foundationalism in a subdomain like normative ethics is justified are establishing the methodological design of the subdomain in the first place and are thus superordinate. In other words, the "default position" of inquiry is coherentism or weak foundationalism and a change from the default position must be justified and held justified over time in coherentist or weak foundationalist terms.

[10] Singer himself emphasizes that Sidgwick is a fallibilist: Cf. 1974, 508.

ii. It follows then that the methodological design of a subdomain that changed from the "default position" is open to ongoing scrutiny in the always dynamic remaining process of the method of reflective equilibrium. To change the method in a subdomain is always provisional, as every justification in reflective equilibrium is provisional and open to further scrutiny. Trying to achieve a reflective equilibrium remains always an ongoing task because there always can be new beliefs – for example through new experiences or evidence – that would have to be incorporated in the holistic system of beliefs. And there is always the possibility that one overlooked relevant background theories or arguments. There could be new inconsistencies at any time which we did or could not anticipate, such that trying to achieve a reflective equilibrium is an ongoing dynamic process. It seems that a perfect reflective equilibrium is a philosophical ideal, but even this ideal state – at a certain time – would be provisional.

If this is all true, then it is clearly non-trivial.

3. The third argument points to the fact that fallibilism is an ideal as well as an important attitude in important domains of civil society: for example, in science or in the political sphere of liberal democracies. This renders information about the use of the method of reflective equilibrium informative and valuable in a social sense – it is useful for us as citizens and epistemic agents if it is stressed that an important theory is fallibilistic. As this is a claim that must be explained and argued for in in detail (which I cannot do extensively here), I consider it only an additional possible argument, which I want to point to at the end of my argumentation.

IV Conclusion

To conclude: Is it misleading or inadequate to state that Sidgwick used the method of reflective equilibrium as his method of justification? Since the method is capable of justifying in the realm of ethics coherentism as well as moderate foundationalism, it is at least important to qualify how Sidgwick used the method exactly: Did he use it to establish a moderate foundationalism for the area of normative ethics with his abstract universal axioms as fallible basic beliefs? Or did he treat his axioms merely as provisionally fixed points (like Rawls treats the judgment about the injustice of slavery as a provisionally fixed point) but not as basic beliefs in the sense that is needed for a moderate foundationalism? To decide the correct answer is a goal for experts on Sidwick's philosophy. So far Singer's position might be vindicated.

But the claim that he used the method of reflective equilibrium in general as his method of justification gives no preliminary decision to this question and is nevertheless quite informative.[11]

And as Sidgwick wasn't merely a fallibilist, but also tried to argue in a coherent way for his position with respect to personally held judgments, I would say it is also quite safe to suggest that he made use of the method of reflective equilibrium.

References

[1] BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA/London: Harvard University Press.

[2] Brink, David O. 1994. "Common Sense and First Principles in Sidgwick's Methods." *Social Philosophy and Policy* 11 (1): 179-201.

[3] Crisp, Roger. 2002. "Sidgwick and the Boundaries of Intuitionism." In *Ethical Intuitionism: Re-evaluations,* edited by Philip Stratton-Lake. Oxford: Oxford University Press, 56-75.

[4] de Maagt, Sem. 2017. "Reflective equilibrium and moral objectivity." *Inquiry: An Interdisciplinary Journal of Philosophy* 60 (5): 443-65.

[5] Daniels, Norman. 1996. *Justice and Justification*, Cambridge: Cambridge University Press.

[6] Elgin, Catherine. 1996. *Considered Judgment*, Princeton: Princeton University Press.

[7] Goodman, Nelson. 1955. *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard University Press.

[8] Langkau, Julia. 2013. "The Method of Reflective Equilibrium and Intuitions" In *Was dürfen wir glauben? Was sollen wir tun? Sektionsbeiträge des achten internationalen Kongresses der Gesellschaft für Analytische Philosophie*, edited by Miguel Hoeltje, et al. University of Duisburg/Essen, 352-64. urn:nbn:de:hbz:464-20130612-081113-3

[9] Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

[11] It would be an interpretative inquiry of its own to determine if this general claim was Rawls's claim or if he indeed interpreted Sidgwick's epistemology for normative ethics as a form of coherentism and pointed therefore to his use of the method of reflective equilibrium.

[10] _____. 1974. “The Independence of Moral Theory.” *Proceedings and Addresses of the American Philosophical Association* 48: 5-22.

[11] _____. 1993. *Political Liberalism*. New York: Columbia University Press.

[12] Sidgwick, Henry. 1907. *The Methods of Ethics*, 7th edition. London: Macmillan and Co.

[13] Singer, Peter. 1974. “Sidgwick and Reflective Equilibrium” *The Monist* 58 (3): 490-517.

[14] _____. 2005. “Ethics and Intuitionism” *The Journal of Ethics* 9 (3-4): 331–52.

[15] _____, and Katarzyna de Lazari-Radek. 2014. *The Point of View of the Universe. Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press.

[16] Skelton, Anthony. 2010. “Henry Sidgwick’s Moral Epistemology” *Journal of the History of Philosophy* 48 (4): 491-519.

[17] Schneewind, J. B. 1963. “First Principles and Common Sense Morality in Sidgwick’s Ethics” *Archiv für Geschichte der Philosophie* 45 (2): 137-56.

[18] Sverdlik, Steven. 1985. “Sidgwick’s Methodology” *Journal of the History of Philosophy* 23 (4): 537-53.

Ist der Personenbegriff wirklich überflüssig für die biomedizinische Ethik?

Shingo Segawa, University of Münster, Germany

Abstract

Der Personenbegriff spielt eine entscheidende Rolle in der biomedizinischen Ethik insbesondere im Kontext moralischer Schutzwürdigkeit am Lebensanfang und Lebensende des Menschen (Abtreibung und Sterbehilfe). Dennoch schlägt Dieter Birnbacher vor, dass wir den Personenbegriff vollkommen aus diesem Themenbereich ausschließen sollten. In diesem Beitrag weise ich darauf hin, dass sein Vorschlag *nur zum Teil* richtig ist. Hier möchte ich die These zum Ausdruck bringen, dass der Personenbegriff für die biomedizinische Ethik *zum Teil* hilfreich ist.

Der Personenbegriff ist immer wieder umstritten, weil er auf verschiedene Weisen interpretiert wird, die wesentlich auf Kant bzw. Locke beruhen. Außerdem führt seine Einführung in die Debatte um Leben und Tod zur moralisch nicht leicht akzeptablen Konsequenz wie z. B. der moralischen Zulässigkeit der Kindestötung. Das Personsein ist untrennbar mit moralischer Schutzwürdigkeit verbunden, so dass es ausschließlich darum geht, ob ein menschliches Wesen eine Person ist. Daraus folgt, dass keine moralische Schutzwürdigkeit dem menschlichen Wesen zugeschrieben werden kann, die nicht als Person anzusehen ist. Aber diese Konsequenz erscheint mit gesellschaftlich weitgehend geteilten moralischen Wertvorstellungen unvereinbar. Ein großes Problem liegt dabei meines Erachtens darin, dass der Personenbegriff konzeptionell zur Gradualisierung moralischer Schutzwürdigkeit nicht beitragen kann.

Vor diesem Hintergrund macht Birnbacher diesen Vorschlag. Birnbacher muss sich deswegen mit der Frage befassen, wie wir ohne Berufung auf den Personenbegriff die moralische Schutzwürdigkeit des Menschen begründen können. Einer der wichtigsten Gründe dafür finde sich in der Empfindungsfähigkeit. Es ist unter normalen Umständen moralisch unzulässig, Schmerzen menschlichen Wesen mit Empfindungsfähigkeit zuzufügen. Wenn man in dieser Weise die moralische Schutzwürdigkeit des Menschen begründen kann, ist der Personenbegriff unbrauchbar. Das soll heißen, dass Birnbachers Vorschlag *völlig* richtig ist. Aber im Hinblick darauf, dass dieser Begriff sehr hilfreich für die Debatte um das Lebensende (Sterbehilfe) ist, sollten wir Birnbachers Vorschlag nicht die *völlige* Zustimmung geben. Sonst verlieren wir eines der geeigneten Argumente in dieser Debatte.

Einleitung

Die Leitfrage meines Beitrags lautet, ob der Begriff der Person wirklich hilfreich für die biomedizinische Ethik ist. Dieser Begriff wird häufig in Debatten um Leben und Tod, z. B. im Fall von Abtreibung und Sterbehilfe, als Begründung verwendet. In der Abtreibungsdebatte spielen der kantische und Locke'sche Personenbegriff eine entscheidende Rolle. Die beiden Begriffe der Person in diesem Bereich scheinen mir jedoch nicht hilfreich zu sein. Denn Autoren wie Otfried Höffe (2002) und Michael Tooley (1983), die den kantischen und Locke'schen Personenbegriff als Begründung ihrer eigenen These verwenden, setzen das Personsein mit dem moralisch zu berücksichtigenden Wesen gleich, sodass die Frage nach

dem Umgang mit den Menschen, die noch nicht oder niemals als Personen betrachtet werden können, gar nicht zur Diskussion steht. In dieser Verwendung lässt sowohl der kantische als auch der Locke'sche Personenbegriff von seiner logischen Struktur her nur zwei Antworten zu: Ein menschliches Wesen ist entweder eine Person mit ausgezeichneter Schutzwürdigkeit oder nichts. Dieser Ansatz erscheint nicht hilfreich für die Untersuchung des Umgangs mit den menschlichen Wesen, die Personen weder im kantischen noch im Locke'schen Sinne wären, weil irgendeine moralische Schutzwürdigkeit auch ihnen zukommen sollte bzw. könnte.

Damit lässt sich die Leitfrage meines Beitrags entwerfen, ob der Personenbegriff überhaupt hilfreich ist für die biomedizinische Ethik. Um die Frage einer Antwort zuzuführen, werde ich mich in meinem Beitrag mit den Beiträgen von Ludwig Siep und Dieter Birnbacher auseinandersetzen. Beide Autoren entwickeln diametral zueinanderstehende Ansätze dazu, sodass sie als ein pointierter Ausgangspunkt für die Beschäftigung mit meiner Leitfrage gelten können. Siep und Birnbacher ist gemeinsam, dass sowohl der kantische als auch der Locke'sche Personenbegriff sich eher lähmend als fördernd auf den Diskurs auswirken. Der entscheidende Unterschied zwischen beiden besteht darin, dass Siep, auf der einen Seite, eine affirmative Antwort auf die Leitfrage gibt. Dazu sucht Siep durch die Entwicklung eines neuen Personenbegriffs, der sich insbesondere nicht auf den Locke'schen Personenbegriff beruft, diesen Begriff in der biomedizinischen Ethik zu retten. Auf der anderen Seite beantwortet Birnbacher die Leitfrage negativ und schlägt daher vor, dass der Personenbegriff aus allen biomedizinischen Debatten auszuschließen ist. Daraus ergeben sich die folgenden drei Antworten auf die Leitfrage:

1: Ja, der von Siep entwickelte Personenbegriff ist für die biomedizinische Ethik geeignet. (Siep)

2: Nein, der Personenbegriff ist für die biomedizinische Ethik überhaupt ungeeignet. (Birnbacher)

3: Der Personenbegriff ist für die biomedizinische Ethik teilweise geeignet. (Segawa)

Ich werde in diesem Beitrag die dritte Antwort vertreten und den Grund dafür erklären. Hierbei ist zu beachten, dass ich Birnbachers Diagnose nicht für *völlig* falsch, sondern nur für *teilweise* falsch halte. So lautet meine These: Der Personenbegriff ist für die biomedizinische Ethik hilfreich, insoweit er in den angemessenen Bereichen verwendet wird. Unter der Voraussetzung, dass allein Personen eine moralische Schutzwürdigkeit haben und allein Wesen, die im hinreichenden Maße über irgendwelche für das Personsein geforderten Fähigkeiten bzw. Eigenschaften verfügen, als Personen anzuerkennen sind, scheint mir

dieser Begriff für die Untersuchung eines moralischen Status von den menschlichen Wesen nicht hilfreich zu sein. Diese Verwendung sieht man typischer weise in der Debatte um Abtreibung. Aber der Personenbegriff kann meines Erachtens dennoch für die biomedizinischen Diskussionen geeignet sein, insoweit es vor allem um den Respekt vor Personen geht, wie etwa in der Debatte um die moralische Zulässigkeit aktiver Sterbehilfe.

I Der Rettungsversuch des Personenbegriffs

Besonders relevant ist für Siep die Frage, wie man mit menschlichen Wesen umgehen sollte, die weder im kantischen noch im Locke'schen Sinne als Personen anzusehen sind. Siep hält die Konsequenz, die sich aus der Einführung der beiden Personenbegriffe in die Diskussion über moralische Schutzwürdigkeit der menschlichen Wesen ergibt, für moralisch unzulässig und sie kann mit dem Begriff der Person gelöst werden. Deshalb wird ihm die Aufgabe auferlegt, einen anderen Begriff der Person als bei Kant und Locke zu entfalten. Ob dieser Personenbegriff zur Auseinandersetzung mit dem moralischen Umgang mit Embryonen dienen könnte, werde ich im nächsten Schritt im Verbund mit dem Beitrag von Birnbacher überprüfen (II).

Aus der Sicht von Siep besteht das wesentliche Problem der Anwendung des kantischen und Locke'schen Personenbegriff auf die Debatte um den Umgang mit den menschlichen Wesen zu Lebensbeginn darin, dass zum einen Personalität (die Bedingungen für das Personsein) lediglich auf die kognitiven Fähigkeiten der Menschen wie Selbstbewusstsein oder Rationalität aufmerksam macht und dass sie zum anderen Gradualisierung zulasse (1993, 44; 2001, 454). So macht Siep den folgenden Vorschlag: Die Personalität kann in dem Ausmaß abgestuft werden, in dem ein menschliches Wesen die Bedingungen für das Personsein erfüllt (2001, 453-55).

Ein menschliches Wesen, so Siep, hat moralische Schutzwürdigkeit deshalb, weil es ein moralisches Recht besitzt. Außerdem wird die Gradualisierung der Personalität bei Siep gestattet. Dies erlaubt es, moralische Rechte einem menschlichen Wesen in dem Ausmaß graduell zuzuschreiben, in dem es die Personalität erfüllt. Hierbei lässt sich die Aussage nicht mehr folgen, dass lediglich Personen, die als vernünftig bzw. selbstbewusst gelten, über moralische Schutzwürdigkeit verfügen, weil auch die menschlichen Wesen, die weder vernünftig noch selbstbewusst sind, als Personen mit irgendwelchem moralischen Recht anzusehen sind. Es geht nun darum, worauf sich die Abstufung der Personalität gründet.

Siep versteht den Begriff der Person als Ganzheit der "leiblichen, emotionalen, intellektuellen und sozialen Leistungen" (2001, 457). Ausgehend davon ist es unumstritten, dass diese alle graduell sind und unter normalen Umständen sich im Verlauf der Jahre entwickeln. Siep erfasst diesen Entwicklungsprozess als ein moralisch zu berücksichtigendes Merkmal. Daraus folgt nicht nur, dass die Personalität nach dem Ausmaß abgestuft werden kann, in dem ein menschliches Wesen den Weg von noch-nicht-Personen zu Personen geht, sondern auch gleichzeitig, dass Ansprüche auf eine moralische Schutzwürdigkeit nicht als ja-oder-nein, sondern als stufenförmig aufzufassen sind. Mit dem Verweis darauf, dass sich das Personsein als graduell verstehen lässt und der Entwicklungsprozess der Menschen als moralisch wichtig gilt, versucht Siep die Geltung der Einführung des Personenbegriffs in die biomedizinische Ethik aufzuzeigen.

II Der Ausschluss des Personenbegriffs aus der biomedizinischen Ethik: Birnbachers Vorschlag

Gegen Sieps Rettungsversuch des Personenbegriffs wirft Birnbacher die Frage auf, ob der Personenbegriff für die biomedizinische Ethik wirklich hilfreich ist. Aus Birnbachers Sicht verbaut Sieps Strategie des Festhaltens an diesem Begriff nicht nur andere Formen der Zuschreibung moralischer Rechte, sondern bringt auch weitere Problematiken mit sich. Daher lautet Birnbachers These: Der Begriff der Person ist überflüssig für die biomedizinische Ethik, sodass wir auf diesen verzichten sollten.

> Auf dem Hintergrund der beschriebenen Dilemmata spricht viel dafür, bioethische Diskussionen ohne den Rückgriff auf den Personenbegriff zu führen oder ihm zumindest eine weniger zentrale Funktion zuzuweisen, als ihm gegenwärtig zugewiesen wird. (Birnbacher 2005, 73)

Birnbacher stimmt mit Siep überein, dass auf der einen Seite der Versuch, moralische Rechte durch die Anwendung sowohl des kantischen als auch Locke'schen Personenbegriffs zu begründen, inakzeptable Konsequenzen nach sich zieht und dass auf der anderen Seite diese Konsequenzen in irgendeiner Weise umgangen werden sollten. Aus Birnbachers Sicht scheint jedoch Sieps Rettungsversuch durch die Abstufung der Personalität mit zwei schwierigen Problemen belastet zu sein. Einerseits muss die Grenzziehung zwischen Personen im vollkommenen Sinne und Personen im unvollkommenen Sinne und Nichtpersonen willkürlich gewählt werden und folglich ein tiefgreifender Dissens darüber herrschen, wo bzw. auf welcher Basis die Grenze zu ziehen ist (Birnbacher 2005, 73-75). Andererseits entfernt dieses Personenverständnis sich sehr weit vom Alltagssprachgebrauch, nach dem die

Aussage "eine Person ist zu 20 Prozent-Person und eine andere Person ist zu 80 Prozent-Person" nicht folgen kann. Aus diesen Gründen muss eine Person Birnbacher zufolge alles oder nichts sein. Die Problematik der Anwendung dieses Personenverständnisses auf die Debatte um moralische Schutzwürdigkeit eines menschlichen Wesens, wie etwa Embryonen, ist aber bereits von Siep hingewiesen worden. Daher besteht das Ziel Birnbachers selbstverständlich darin, ohne den Personenbegriff moralische Rechte solcher menschlichen Wesen *graduell* zu begründen. Birnbacher schreibt wie folgt:

> … man bestimmte moralische Rechte haben kann, ohne jedes mögliche moralische Recht zu haben. […, S.S.] Man braucht Wesen, die Personen sind, nicht alle möglichen Rechte zuzusprechen, und man braucht Wesen, die keine Personen sind, nicht bestimmte oder alle moralischen Rechte abzusprechen. (2005, 74)

Es geht bei Birnbacher deshalb darum, wie moralische Rechte abgestuft werden können. Als Beispiel hierfür gibt es ein moralisches Recht auf Leidensvermeidung. Birnbacher nach lässt sich dieses Recht unter Rückgriff auf die Empfindungsfähigkeit auch Embryonen zusprechen, die zu einem bestimmten Stadium entwickelt sind. Wenn die Embryonen mit der Empfindungsfähigkeit ausgestattet sind, sollten sie als moralisch zu berücksichtigende Wesen betrachtet werden, wobei es nicht darum geht, ob sie als Personen anzusehen sind. Birnbacher erachtet tatsächlich die Empfindungsfähigkeit als Begründung eines der grundlegendsten moralischen Rechte für hinreichend. Können wir uns ohne Berufung auf den Begriff der Person mit der Diskussion über moralische Schutzwürdigkeit der menschlichen Wesen beschäftigen, spielt der Personenbegriff keine Rolle dabei. Wie Birnbacher herausstellt, öffnet die Aufgabe des Begriffs der Person einen anderen Weg dafür, eine moralische Schutzwürdigkeit der menschlichen Wesen zu Lebensbeginn auf die verschiedenen Weisen, wie z. B. Gattungsethik von Siep (2002) oder Pietät von Birnbacher,[1] veranschlagen zu können. Auch wenn diese Arten und Weisen noch nicht als Begründung in dieser Debatte garantiert sind, ergeben sich tatsächlich aus dem Verzicht auf den Personenbegriff die anderen Möglichkeiten nicht in der Weise eines Alles-oder-Nichts. Ich stimme daher Birnbacher nur darin zu, dass der Begriff der Person nicht hilfreich für die Debatte um moralische Schutzwürdigkeit der menschlichen Wesen *zum Lebensbeginn* ist. Aber anders als

[1] Vgl. Birnbacher 2005, 300: „Ähnlich wie die Pietätspflichten gegenüber menschlichen Leichnamen werden diese Pflichten deshalb stets nur schwache oder Prima-facie-Pflichten sein können, d. h., es werden Pflichten sein, die eine Abwägung mit konkurrierenden Pflichten und Rechten erlauben und gegebenenfalls hinter diesen zurückstehen müssen - so wie im Fall einer gerichtlich angeordneten Obduktion die Pietätspflichten gegenüber der konkurrierenden Pflicht zur Sicherung des Rechtsfriedens zurückstehen müssen. Darüber hinaus sind diese Pflichten zeit- und kulturrelativ. Falls die Vermutung berechtigt ist, dass die Einstellungen zur Embryonenforschung in Deutschland deutlich ablehnender sind als etwa in den USA, können wir nicht davon ausgehen, dass die Forschungsbeschränkungen, die hier zu Recht bestehen, auch dort moralisch verpflichtend sind."

bei Birnbacher folgt bei mir nicht, dass dieser Begriff aus der biomedizinischen Ethik ausgeschlossen werden sollte, weil meines Erachtens es bestimmte biomedizinischen Problematiken gibt, zu denen der Personenbegriff doch beitragen kann. Aus diesem Grund halte ich Birnbachers These nur *teilweise* für richtig.

III Ist der Personenbegriff wirklich überflüssig für die biomedizinische Ethik?

Ob mein Versuch dieses Beitrags gelingt, hängt von den folgenden Überlegungen dieses Abschnittes ab. Um meine These plausibel zu machen, dass der Personenbegriff für die biomedizinische Ethik hilfreich ist, behandele ich das Prinzip des Respekts vor Autonomie und weise auf die Rolle des Personenbegriffs dabei hin. Das Prinzip des Respekts vor Autonomie zählt zu den relevantesten Prinzipien in Diskussionen über die biomedizinische Ethik (Beauchamp and Childress 2013). Dieses Prinzip steht typischerweise im Vordergrund der Diskussion über die moralische Zulässigkeit der Sterbehilfe, die in vier Kategorien einzuteilen ist: Passive, indirekte, aktive Sterbehilfe und ärztlich assistierter Suizid. Da es bei meinem Beitrag nicht um die moralische Zulässigkeit der Sterbehilfe an sich geht, fokussiere ich mich in diesem Abschnitt nur auf die Rolle des Personenbegriffs in der Debatte um die moralische Zulässigkeit aktiver Sterbehilfe. Darüber hinaus beschränken sich die folgenden Überlegungen auf einen Fall: Einerseits leidet der Patient an einer unheilbaren Krankheit und an unerträglichen Schmerzen (z. B. Krebs in späten Stadien) oder der Patient ist nicht mehr in der Lage, ein verschriebenes tödliches Medikament selbst einzunehmen (z.B. vollständige Lähmung). Andererseits befindet sich der Handelnde, d. h. in diesem Kontext der Arzt, in der Lage, mit Sicherheit zu erwarten, dass sein Handeln in jedem Fall zum Tod des Patienten führt.

Passive Sterbehilfe zeichnet sich dadurch aus, dass eine lebensverlängernde medizinische Behandlung nach Einsetzen des Sterbeprozesses, etwa die Wiederbelebung, unterlassen wird und dadurch der Eintritt des Todes beschleunigt wird.[2] Dabei liegt die Annahme

[2] Der Nationale Ethikrat schlägt vor, dass statt von passiver Sterbehilfe der Terminus von Sterbenlassen in diesem Kontext gebraucht werden sollte. Vgl. Nationaler Ethikrat 2006, 53.

zugrunde, dass der Sterbeprozess bereits eingesetzt hat und folglich das absichtliche Unterlassen nicht mit der absichtlichen Herbeiführung des Todes verbunden ist.[3] Diese zwei Bedingungen machen das Merkmal für ihre moralische Zulässigkeit aus.

Indirekte Sterbehilfe ist dadurch charakterisiert, dass der Eintritt des Todes von dem an z. B. unerträglichen Schmerzen leidenden Patienten als Nebenfolge der Überdosis beschleunigt wird. Es ist besonders relevant, dass ein solches Handeln weder direkt noch indirekt vorsätzlich auf den Tod des Patienten zielt, sondern sein Tod nur in Kauf genommen wird.[4] Dieses Merkmal liefert das Argument für die moralische Zulässigkeit der indirekten Sterbehilfe.

Im Gegensatz zu den ersten beiden Formen der Sterbehilfe geht aktive Sterbehilfe mit der gezielt ausgeführten absichtlichen Herbeiführung des Todes vom Patienten einher. Die aktive Sterbehilfe, die sich durch absichtliche Handlung der Herbeiführung des Todes vom Patienten und durch ein fehlendes Einsetzen des Sterbeprozesses auszeichnet, steht dann den zwei Bedingungen für die moralische Erlaubtheit von Sterbehilfe gegenüber. Die Frage hinsichtlich der moralischen Zulässigkeit von Sterbehilfe fokussiert sich darauf, ob der ethisch radikale Unterschied zwischen den drei Formen aus der Differenzierung vom absichtlichen Handeln, das zum Tod führt (aktive Sterbehilfe), der Inkaufnahme des Todes (indirekte Sterbehilfe) und des absichtlichen Unterlassen der Behandlung (passive Sterbehilfe) abgeleitet werden kann. Hiergegen wird der detaillierte sowie umfassende Einwand durch die Analyse der Handlungstheorie vorgebracht (Birnbacher 1995). Unter der oben genannten Voraussetzung dieses Beitrags kann die Trennung zwischen dem absichtlichen Tun, dem in Kauf genommen Tun und dem als ein Handeln verstandenes Unterlassen aus Sicht des kausal-relationalen Verhältnisses nicht mehr gesichert angenommen werden.[5] Daraus folgt, dass der ethisch radikale Unterschied zwischen den drei Formen der Sterbehilfe nicht mehr haltbar ist. Der wesentliche Grund für die moralische Zulässigkeit der Sterbehilfe, unabhängig von ihren drei verschiedenen Formen, liegt darin begründet, dass der Patient sich

[3] Das absichtliche Unterlassen entspricht dem von Birnbacher ausgelegten „Geschehenlassen". Birnbacher erklärt den Unterschied zwischen Unterlassen und Geschehenlassen folgendermaßen: „Während ein Unterlassen (wie auch Handeln) auch unwissentlich erfolgen kann […, S.S.], kann ein Geschehenlassen immer nur *wissentlich* [im Original] sein" (1995, 104 und 2016, 90).

[4] Aus dieser Sicht soll der Terminus von indirekter Sterbehilfe aufgegeben werden: „Auf den bisher in diesem Zusammenhang verwendeten Begriff der ‚indirekten Sterbehilfe' sollte verzichtet werden, weil der Tod des Patienten weder direkt noch indirekt das Ziel des Handelns ist" (Nationale Ethikrat 2006, 54).

[5] Siep und Quante weisen darauf hin, dass nicht nur Tun, sondern auch „Unterlassung [an sich bereits unabhängig von willentlich oder unwillentlich, S.S.] ein raum-zeitlich reales Ereignis ist" (1998, 46).

selbst für Sterbehilfe entscheidet. Der Schlüssel für die moralische Erlaubtheit der aktiven Sterbehilfe ist Selbstbestimmung.

Es scheint mir jedoch selbstverständlich zu sein, dass jede Selbstbestimmung des Patienten, der nicht an einer unheilbaren Krankheit und an unerträglichen Schmerzen leidet, vor allem im Kontext der aktiven Sterbehilfe respektiert werden sollte. An dieser Stelle geht es darum, welche Selbstbestimmung des Patienten respektiert werden sollte. Meine Antwort darauf ist autonome Selbstbestimmung. Es stellt sich an dieser Stelle die Frage, wodurch sich autonome Selbstbestimmung auszeichnet. Mit Verweis darauf, dass das Konzept personaler Autonomie dazu beitragen kann, argumentiere ich für meine These. Denn für das Konzept personaler Autonomie ist der Begriff der Person notwendig.

Der Begriff der Person setzt sich nicht nur aus der Personalität (die Bedingungen für das Personsein), sondern aus der Persönlichkeit zusammen. Die Persönlichkeit charakterisiert in unserem alltäglichen Leben ein individuelles Merkmal der Person, sodass sie sich implizit oder explizit in ihren bestimmten ethischen Ansprüchen widerspiegelt. Damit wird verständlich, weshalb sich ein gewisser ethischer Anspruch für eine Person als ziemlich relevant darstellt, aber für eine andere Person nicht in der gleichen Weise. Die Persönlichkeit ist relevant für die Untersuchung personaler Autonomie. Denn dem Konzept personaler Autonomie zufolge ist eine Person in konkreten Selbstbestimmungen insofern autonom, als diese Selbstbestimmung ihrer Persönlichkeit entspricht. Wenn die Selbstbestimmung der Person nicht ihrer Persönlichkeit entspricht, ist sie nicht als autonom, d. h. nicht als moralisch zulässig, anzusehen. Infolgedessen kann das soeben Problem durch die Berufung auf das Konzept der personalen Autonomie vermeidet werden. Im Hinblick darauf, dass das Konzept der personalen Autonomie auf dem Begriff der Person basiert, sollte der Personenbegriff nicht aus der biomedizinischen Ethik ausgeschlossen werden. Somit komme ich zu dem Schluss, dass der Begriff der Person teilweise geeignet für die biomedizinische Ethik ist.

Literatur

[1] Beauchamp, L. T., and F. J. Childress. 2013. *Principles of Biomedical Ethics*. Oxford: Oxford University Press.

[2] Birnbacher, Dieter. 1995. *Tun und Unterlassen*. Stuttgart: Reclam.

[3] _____. 2005. *Bioethik zwischen Natur und Interesse*. Frankfurt a. M.: Suhrkamp.

[4] _____. 2016. “Unterlassungen.” In *Handbuch Handlungstheorie. Grundlagen, Kontexte, Perspektiven*, edited by Michael Kühler, and Markus Rüther. Stuttgart: Springer, 90-8.

[5] Höffe, Otfried. 2002. *Medizin ohne Ethik?* Frankfurt a. M.: Suhrkamp.

[6] Nationaler Ethikrat. 2006. *Selbstbestimmung und Fürsorge am Lebensende*. *Stellungnahme.* Berlin. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/Archiv/Stellungnahme_Selbstbestimmung_und_Fuersorge_am_Lebensende.pdf

[7] Siep, Ludwig. 1993. "Personbegriff und angewandte Ethik." In *Person und Sinnerfahrung. Philosophische Grundlagen und interdisziplinäre Perspektiven,* edited by Carl Friedrich Gethmann und Peter L. Osterreich. Darmstadt: Wissenschaftliche Buchgesellschaft.

[8] _____, and Michael Quante. 1998. "Ist die aktive Herbeiführung des Todes philosophisch zu rechtfertigen?" In *Das medizinisch assistierte Sterben. Zur Sterbehilfe aus medizinischer, ethischer, juristischer und theologischer Sicht*, Edited by Adrian Holderegger. Freiburg: Herder.

[9] _____. 2001. "Der Begriff der Person als Grundlage der biomedizinischen Ethik: Zwei Traditionslinien." In *Person. Philosophiegeschichte. Theoretische Philosophie – Praktische Philosophie*, edited by Dieter Sturma. Münster: mentis.

[10] _____. 2002. "Moral und Gattungsethik." *Deutsche Zeitschrift für Philosophie* 50 (1): 111-20:

[11] Tooley, Michael. 1983. *Abortion and Infanticide*. Oxford: Oxford University Press.

Is Hedonism a Version of Axiological Monism?

Adam Shriver, University of Oxford, UK

Abstract

Axiological monism refers to accounts of value that hold that there is just one type of goodness or value in the world. Hedonism is frequently taken to be one of the canonical examples of axiological monism since, according to hedonism, there is just one type of good: namely positive experience. However, hedonism is committed not just to the claim that positive experiences are the sole good, but also to the claim that negative experiences are the sole bad. And given that the goodness and badness of experience must be weighed against one another in order to reach an overall assessment of the welfare of an individual according to hedonism, we can ask whether hedonism truly retains the purported theoretical advantages that are thought to apply to monism.

I argue that pleasures and pains are sufficiently different such that hedonism cannot retain the advantages typically assigned to axiological monism. To make this case, I critically evaluate recent discoveries in the scientific study of pleasure and pain. The upshot of these two sets of evidence, I argue, is that the goodness of pleasure cannot be explained in the same manner as the badness of pain, and that there is no adequate way of trading the value of pleasure against the disvalue of pain interpersonally. As such, I argue that hedonism does not retain the advantages of axiological monism.

Introduction

Axiological monism refers to accounts of value that hold that there is just one type of value. Such views are typically contrasted with versions of axiological pluralism, which hold that there are multiple kinds of value. Hedonism about value has often been portrayed as a canonical example of a monistic account of value. However, Elinor Mason (2018), citing Shelly Kagan's recent (2014) work on ill-being, has suggested that incorporating the disvalue of pain into an account of value may problematize hedonism's status as a monistic account of value. This presentation expands on Mason's brief suggestion and argues that that the disvalue of pain complicates the view that hedonism possesses the theoretical advantages over pluralism that have traditionally been associated with axiological monism.

Monism is frequently expressed in two different ways. One common way of expressing it is exemplified in the SEP article on Value Pluralism (Mason 2018), which suggests that monism is the view that there is only one fundamental value. But Chris Heathwood's article "Monism and Pluralism" provides an example of a different way of describing monism. Heathwood writes that the oldest and simplest version of welfare monism is the view that "pleasure is the one thing of ultimate benefit to us *and* pain the one thing of ultimate harm" (2015, my

emphasis). So monism is sometimes expressed as "there is only one type of value," and sometimes expressed as "there is only one type of good thing and only one type of bad thing." However, the claim that there is only one kind of value is not equivalent to the claim that there is only one type of good thing *and* there is only one type of bad thing in the absence of further exegesis.

Given these two ways of describing monism, answering the question "Is hedonism a version of axiological monism?" may seem to depend simply on which definitions are used. However, I will focus on a different type of question: does hedonism retain the theoretical advantages that have been traditional associated with axiological monism? That is, can the advantages ascribed to monistic accounts of value as used in reasoning about well-being and impersonal value be realized by hedonistic accounts of pleasure and pain? I will argue that they cannot, or at least that more theoretical work is required in order to show that they can.

I Clarifying the Concept

I have the following in mind when talking about value in relation to axiological monism. First, axiological monism of course refers to intrinsic value rather than instrumental value. Second, the type of value under discussion is about the goodness or badness of states of affairs of the type consequentialists would care about, rather than the "moral value" of rightness or wrongness of actions or virtuous and vicious characters. And finally, we can distinguish between value *for* a given individual, or welfare value, and value simpliciter (or impersonal value). My focus will be on value simpliciter, although as will be seen I very much think this type of value depends upon the welfare on individuals.

II Advantages of Axiological Monism

Axiological monism is thought to have certain theoretical advantages relative to axiological pluralism (Heathwood 2015). One purported advantage of axiological monism is that it possesses a type of explanatory adequacy lacking in pluralistic accounts where different sources of value are not unified by a common feature. If a pluralistic account claims that a certain number of properties have value, we can ask what makes it the case that those properties and only those properties have value. If there is a common explanation for why those properties have value, then it seems likely that the theory would collapse into value

monism. If there is not a shared explanation, then the theory fails to explain something important.

Another purported advantage of monism is that it can account for the comparability of different values whereas forms of pluralism, arguably, imply that certain types of values are fundamentally incomparable. Consider the claim that pleasure is the only good. On this view, since all goods are pleasures, and since pleasures can be quantified by intensity and duration, it follows that all goods can in principle be compared to one another. In contrast, consider the pluralist view that both pleasure and achievement are goods. How does one measure a given amount of achievement against a given level of pleasure? It is not clear how such heterogeneous values can be comparable. Of course, numerous attempts have been made to explain how such comparisons can proceed, but there is at least a prima facie reason for thinking that a monistic account can more easily explain the comparability of value.

III Disvalue Needs to Be Included in a Monistic Account of Value

Were there no badness or disvalue in the world, a monistic account of goodness would be equivalent to a monistic account of value. Higher levels of the goodness would be better, and the world could be made "worse" only by preventing the more goodness from being realized, but things would never be worse than a state of neutrality. Unfortunately, however, it is the case that there are states that are worse than neutral; an existence of nothing but intense suffering is worse than having not existed at all. For this reason, we need more than just an account of goodness to provide a comprehensive account of value.

Moreover, a monistic account of positive value that is silent on disvalue is practically useless. Consider a case where you need to compare the value of Situation A vs Situation B. If you only know how much positive value is present in each situation, but know nothing about how much disvalue is present, then you cannot evaluate which situation is preferable. This is true for both decisions about prudential well-being and for decisions about impersonal value. As such, the purported comparability advantage of monism is practically inert if it only applies to accounts of goodness or positive value and not to badness or disvalue.

IV Disvalue and Value Require a Shared Explanation

The explanatory advantage of axiological monism is also lost if the monistic account does not include disvalue. Consider a discussion by Heathwood of a suggestion that we consider the following an account of monism: "X is good iff X is either a state of pleasure *or* a state of knowledge." The reason such an account should actually be regarded as pluralistic rather than monistic is that the explanation for why pleasures are valuable is different from the explanation of why knowledge is valuable. There is no shared explanation for pleasure and knowledge and thus no monism.

Now consider attempts to describe monism as follows: X is the only good, and Y is the only bad. Such accounts are monistic about goodness, and monistic about badness, but are they also monistic about value as a whole? Not without further argument. Here is an account of value inspired by my nephew: "Achievement is the only good, and suffering is the only bad." Putting aside whether the view is plausible, would it be a monistic account of value? I think most would agree that it clearly is not. The reason it is not, I believe, is the same reason why pleasure and knowledge cannot be combined into a monistic account of goodness; achievement and suffering are not related to one another in the right way. In particular, there is no shared explanation for why achievement would be the good and suffering the bad.

As such, theories that suggest that X is the only good and Y is the only bad are only monistic accounts of value if there is a shared explanation for why X is good and why Y is bad. Is hedonism such a theory? Of course, pleasures and pains are both experiences. But saying X has value iff X is an experience is inadequate in at least two ways. First, we can presumably have neutral experiences, and so the fact that something is an experience doesn't automatically entail that it has value. And second, it does not follow from the claim that "experiences can be good" that "experience can be bad," so there's still additional explanation required about how the claims about pleasure relate to the claims about pain. In what follows, I will consider whether a shared explanation can be provided for why pleasure is good and pain is bad.

V Symmetry's Discontents

The intuition that pleasure and pain are related to value in similar (or opposite) ways is deeply entrenched. However, there is also a tradition of questioning this relationship. Numerous philosophers, including Popper (1950), Hurka (2011), Mayerfeld (1999), Benatar

(2008) and Shriver (2014a, 2014b) have questioned, in various ways, whether pleasure and pain are symmetrically related to well-being or goodness. Many of these accounts have used intuitions about thought experiments to undermine the view that pleasures and pains can be treated similarly. In the sections that follow, I proceed with a different methodology; examining what we know about the science of pleasure and pain and attempting to build a theory about the value of pleasure and pain from the ground up rather than working backwards from intuitions.

VI Is There a "Common Currency" of Value Shared by Pleasure and Pain?

It has been suggested that the types of things we call "pleasure" are so dissimilar from one another that it is implausible to suggest that they are the same type of thing. In particular, Mill famously suggested that no amount of lower pleasures could, by themselves, ever equal the value of a higher pleasure. On such a view, hedonism's status as a monistic theory is jeopardized by consideration even of the differences between, say, the pleasure of eating strawberries and that of reading philosophy. But in a recent article, Roger Crisp and Morten Kringelbach appeal to evidence from the neurosciences to resist such a conclusion. They write the following:

> Consider… the bolt-on view, according to which experiences become pleasurable through the activation of a certain pleasure circuit, or certain circuits, common to all pleasurable experiences … Much of the neuroscientific evidence is of course not yet in, but at present the tenor of research suggests that the neural substrate of pleasure in quite different kinds of activity is quite similar … In other words, current neuroscience seems to favour the bolt-on thesis, and hence the denial, for the present at least, of the higher/lower thesis. (Crisp and Kringelbach 2018, 213-14)

The gist of their argument is that there is a well-defined set of brain regions that are active during the experience of pleasure, and that these same regions are active across different types of experiences that have been described as heterogeneous by critics of hedonism. As such, the argument goes, the idea that there is a qualitative difference between higher and lower pleasures seems at odds with what we currently know about the brain.

Can a similar argument be made in response to the suggestion that pleasure and pain do not have a common currency? As I've argued elsewhere in more detail (Shriver 2014b), the core brain regions involved in pleasure and the unpleasantness of pain appear to be distinct,

even if there are some brain regions active during both. One cannot point to activity in particular brain regions as evidence of a common currency of positive and negative experiences.

Moreover, there is no consistent relationship that has been found in psychology research between the value people assign pleasure and the value they assign pain. Of course, individuals at particular times trade off pleasures and pains in their decisions, but there is no clear ratio that holds across individuals or even within individuals across time. We have no evidence of a consistent internal valuation of the relationship between pain and pleasure. As such, there is no more evidence of a common currency existing between pleasure and pain than there is of that existing between pleasure and achievement; in both cases, the only thing we can say is that, when forced to choose between two options, we make a choice. As I will argue below, I think the problem lies in our inconsistent evaluation of pleasure.

The above evidence only provides a response to a potential argument for a common currency of pleasure and pain. Are there any positive arguments suggesting that pleasure and pain are in fact related to value in fundamentally different ways? I now turn to an argument along these lines.

VII Pain and Pleasure Are Structurally Different Types of Experiences

The standard story for experiencing pleasure is as follows: one has a desire and acts in order to fulfill that desire. If the desire is successfully fulfilled, pleasure ensues. In contrast, consider the case of pain: one has the experience of pain *and* the desire to escape the pain simultaneously. When the aversive desire is fulfilled, the negative hedonic experience goes away. These are oversimplified descriptions, of course, but I believe represent the canonical instances of these cases. And I will argue that these differences in fact are indicative of important features of pleasure and pain.

One complication with the above story about pleasure is now familiar to many philosophers. Kent Berridge, a psychologist at Michigan, has demonstrated that 'wanting' (or appetitive motivation) can be dissociated from 'liking' (or pleasure). Through direct interventions on the brain, Berridge and others were able to induce instances of liking a particular taste sensation without wanting it, and instances of wanting without liking. Cases of addiction where addicts no longer get much pleasure from the drugs are also thought to be clear examples

where the 'liking' component has come apart from the motivation to pursue particular rewards.

Can a similar story be told about pain? People given low doses of morphine, cancer patients who have undergone a procedure known as cingulotomy where the cingulate cortex is lesioned, and people with a rare condition known as pain asymbolia all report instances of feeling pains but not finding the pain unpleasant (Aydede 2005; Shriver 2006). Thus, there seems to be a dissociation between "having a pain" and "having an unpleasant sensation" that may initially seem to mirror the case of pleasure.

However, this appearance of similarity is misleading. In typical cases of pleasure and pain, there are not just two relevant components of the phenomenon (the experience itself and the motivational signal), but actually three: the sensory components (SC) of the experience, the hedonic components (HC) of experience, and the motivational component (MC). In the case of getting a pleasurable shoulder massage, the sensory component would consist of details about the representation of touch in a particular location, the hedonic component would consist of the accompanying pleasure, and the motivational component would consist of one's desire for the experience to continue. In the case of a painful experience of stepping on a tack, the sensory component would consist of a representation of a puncture in a particular region of the body with a certain intensity of pain, the hedonic component would consist of the unpleasantness of that experience, and the motivational would consist of the urge to pull one's foot away from the tack.

With pleasure, evidence from the sciences has shown that all three components can be dissociated from one another. One can have a representation of rubbing on one's shoulder (SC) without any pleasure (HC) and without any motivation for it to continue (MC). And, as was mentioned above, research has shown that the hedonic and motivational components can also be pulled apart from one another. In both humans and other mammals, certain brain states can result in motivation signals even without hedonic tone ('wanting' without liking'). Conversely, other states produce changes in hedonic tone without influencing motivational signals (thus influencing 'liking' but not 'wanting'). As such, the sensory, motivational, and hedonic components of pleasure *all* appear to be separate from one another.

In the case of pain, experiments have shown is that the sensory components of pain can be dissociated from its hedonic tone and from the desire to escape, but it has not been shown that unpleasantness of pain can be separated from the motivation to avoid the experience. In fact, there is good evidence that precisely the same processes involved in generating the unpleasantness are also crucial for the motivational component of pain. Administering morphine to rats or lesioning their anterior cingulates in conditioned place preference tasks

produces similar dissociations as in humans: they still show pain behavior but no longer are motivated to escape the situation nor to avoid it in the future (LaGraize et al. 2004, 2006). Furthermore, glutamatergic activation of the anterior cingulate in rats can induce place avoidance even in the absence of noxious stimulation, while inhibiting activation in the anterior cingulate can block place avoidance even in the presence of noxious stimulation (Johansen and Fields 2004). Likewise, direct stimulation of the insula has produced pain sensations (Ostrowsky et al. 2002), while patients with certain insula lesions are characterized by an indifference to potentially harmful stimuli (Grahek 2001). Thus, in the case of pain, unlike in that of pleasure, there is evidence that the key neuronal substrates of unpleasantness are the same substrates that produce motivational signals. This evidence suggests that the feeling of unpleasantness cannot be separated from a motivating signal to avoid the experience (though this signal, of course, can be overridden by other, competing motivations)

There's also a good supporting evolutionary story that can be told about why pain and pleasure have different relationships to motivation. In general, an organism failing to immediately respond to painful events in the environment can have catastrophic consequences including death. As such, it makes sense for the relationship between pain and motivation to have evolved to be direct and urgent. In contrast, failing to pursue pleasures can lead to disadvantages related to evolutionary fitness, but in general are not catastrophic. Thus, a contingent and weaker relationship between pleasure and motivation would be consistent with evolutionary fitness.

So, to summarize, pleasure and pain are related to our motivation systems in importantly different ways. In particular, aversive motivation is intrinsically linked to the unpleasantness of pain, whereas appetitive motivation is only contingently linked to the pleasantness of pleasure, and can in some cases be separated from it altogether. This is consistent our everyday experiences where appetitive desire and pleasure often occur at separate times whereas the unpleasantness of pain is always accompanied by aversive desire. In the final sections, I explain why these differences undermine hedonism's claims to the advantages of axiological monism.

VIII Are Pleasure and Pain Comparable?

Recall that one of the purported advantages of axiological monism over pluralism is that monism can account for the comparability of different values whereas pluralism may suggest that certain values are incomparable. From the previous section, then, we may ask

whether pleasure and pain are comparable on a single scale of value. For example, we can ask whether avoiding a certain quantity of pain is equally valuable to experiencing some particular quantity of pleasure.

I have already provided some hints as to why there are reasons to doubt that pleasure and pain are comparable in the right way. It is useful to consider the comparison of pain against pleasure with that of pleasures against other pleasures and pains against other pains. Recall the evidence for neural correlates of pleasure in the discussion of Crisp and Kringelbach. Not only does activation in particular brain regions track whether pleasure is present or not, it is correlated with the intensity of the pleasures involved. When comparing the pleasure of eating strawberries with that of hiking up a mountain, for example, we can in principle track the intensity of these different types of pleasure by looking at activation in these regions and can compare the pleasures accordingly.

Similarly in the case of pain, significant progress has been made in recent years in developing what has been called a "neural signature of pain." At Colorado, Tor Wager's group used a machine learning algorithm to develop a system that could use brain imaging data to predict, with 95% accuracy, not only whether a person was in pain but also how intense the pain is (2010). In other words, pain intensity, as well as pain unpleasantness, can be quantified quite straightforwardly and can be used to compare the disvalue of different pains.

However, when it comes to comparing the value of pleasure against the disvalue of pain, there is no similar measure we can use. As I noted previously, there's no consistent tradeoff between pleasure and pain across individuals, or even within individuals across different times. Moreover, I believe the discussion of the previous section provides us with reasons for being skeptical that such a tradeoff function could ever be produced. The unpleasantness of pain is directly related to motivational urges. However, the relationship between pleasure and motivation is only contingent, and turns out to be quite variable. As Mayerfeld suggested, and as was apparent in some of the other examples, it is possible to be indifferent to future pleasures. The full story about the relationship between reward, desire, and pleasure is more complicated than I have described here,[1] but the upshot of the previous discussion is that different people can and do assign different value to pleasure. As such, there is reason to be skeptical that it is possible to trade pleasures for pains across individuals in a principled manner since different people will care about pleasure to different degrees.

[1] See Schroeder 2004 for a helpful overview.

Without the ability to "trade off" pleasure and pain, the explanatory advantages of hedonism are lost. We might believe that pleasure is the good, pain is the bad, and know all of the facts of how much pleasure and pain will be produced by a given decision between choice X and choice Y, but if we don't know how pain and pleasure can be traded against one another we will be unable to know which outcome is preferable. As such, the theoretical advantages of comparability seem lost.

IX Is There a Shared Explanation for the Goodness of Pleasure and the Badness of Pain?

Given that pleasure and pain evolved to solve separate problems, have dramatically different influences on our behavior, and are structurally related to our brain's capacity to value and our perceived valuations in fundamentally different ways, I answer "no" to this section's title question.

The comparability advantage of monism and the explanatory advantage have been presented as separate from one another. However, as we saw in Heathwood's discussion of the hypothetical view that pleasure and achievement are good, these two advantages are in fact inextricably linked. The explanation for why a given state is good or bad is what allows us to sort the different types of value in the world. If there is only one explanation of what makes something valuable, then there is only one type of value. If, on the other hand, we require different explanations for why pleasure and achievement are valuable, and we think they are both valuable, then we are forced to adopt a pluralist position.

However, part of explaining which states provide value is involves giving us some idea of how to measure the given level of value. Saying that pleasure is the good implies that more pleasure adds more goodness. Saying that pain is bad implies that more pain adds more badness. But stating that pleasure is good and pain is bad, without providing some idea of how they can be traded off against one another, does not provide a full story about how experience contributes to the value of the world. As such, I suggest that there is no more reason to suppose that there is a shared explanation for how pleasure and pain are related to value and disvalue then there is reason to think that a common explanation can be provided for why pleasure and achievement are both, independently, goods.

X Conclusion

The upshot of my argument is as follows: first, the badness of pain and the goodness of pleasure do not share a similar explanation, since the unpleasantness of pain cannot be dissociated from a motivational urge to avoid the pain, whereas the pleasantness of pleasure can and does come apart from the desire for the relevant experiences. Second: there is no reliable way of trading pleasures for pains between individuals, as the value of pleasure assigned by individuals varies more dramatically than the disvalue assigned to pains. As such, I argue that hedonism does not retain the advantages of axiological monism. This does not mean that hedonism is an incorrect account of value, but it does suggest that additional work is required in order to explain how positive and negative experiences are related to one another.

References

[1] Aydede, M., ed. 2005. *Pain: New Essays on its Nature and the Methodology of its Study*. Cambridge, MA: MIT Press.

[2] Benatar, D. 2008. *Better Never to Have Been: The Harm of Coming into Existence*. Oxford: Oxford University Press.

[3] Crisp, R., and M. Kringelbach. 2018. “Higher and Lower Pleasures Revisited: Evidence from Neuroscience.” *Neuroethics* 11 (2): 211-5.

[4] Grahek, N. 2001. *Feeling Pain and Being in Pain*. Cambridge, MA: MIT Press.

[5] Heathwood, C. 2015. “Monism and Pluralism about Value.” In *Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson. Oxford: Oxford University Press, 136-57.

[6] Hurka T. 2011. *The Best Things In Life*. New York: Oxford University Press.

[7] Johansen, J.P., and H.L. Fields. 2004. “Glutamatergic Activation of Anterior Cingulate Cortex Produces an Aversive Teaching Signal.” *Nature Neuroscience* 7: 398–403.

[8] Kagan, S. 2014. “An Introduction to Ill-Being” *Oxford Studies in Normative Ethics*. *Volume 4*, edited by Mark Timmons. Oxford: Oxford University Press.

[9] LaGraize, S., C. Labuda, R. Rutledge, R. Jackson, and P. Fuchs. 2004. “Differential Effect of Anterior Cingulated Cortex Lesion on Mechanical Hypersensitivity and Escape/Avoidance Behavior in an Animal Model of Neuropathic Pain.” *Experimental Neurology* 188: 139–48.

[10] LaGraize, S., J. Borzan, Y.B. Peng, and P. Fuchs. 2006. “Selective Regulation of Pain Affect Following Activation of the Opiod Anterior Cingulate Cortex System.” *Experimental Neurology* 197: 22–30.

[11] Leknes, S., and I. Tracey. 2010. “Pain and Pleasure: Masters of Mankind.” In *Pleasures of the Brain*, edited by M. Kringelbach and K. Berridge. New York: Oxford University Press, 320-35.

[12] Mason, Elinor. 2018. "Value Pluralism" *The Stanford Encyclopedia of Philosophy* (Spring 2018 edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/value-pluralism/.

[13] Mayerfeld, J. 1999. *Suffering and Moral Responsibility*. New York: Oxford University Press.

[14] Mill, J.S. 1979. *Utilitarianism*. Indianopolis: Hackett Publishing.

[15] Miller, J.M., S.R. Vorel, A.J. Tranguch, E.T. Kenny, van Mazzoni, W.G. Gorp, and H.D. Kleber. 2006. “Anhedonia After a Selective Bilateral Lesion of the Globus Pallidus.” *American Journal of Psychiatry* 163: 786–8.

[16] Moore G. E. 1903. *Principia Ethica, Revised Edition*. Cambridge: Cambridge University Press.

[17] Ostrowsky, K., M. Magnin, Ryvlin, J. Isnard, M. Guenot, and F. Mauguière. 2002. “Representation of Pain and Somatic Sensation in the Human Insula: A Study of Responses to Direct Electrical Cortical Stimulation.” *Cerebral Cortex* 12 (4): 376–85.

[18] Popper K. 1950. *The Open Society and Its Enemies*, Princeton, NJ: Princeton University Press.

[19] Shriver, A. 2006. “Minding Mammals.” *Philosophical Psychology* 19 (4): 433-42.

[20] _____. 2014a. “The Asymmetrical Contributions of Pleasure and Pain to Animal Welfare.” *Cambridge Quarterly of Healthcare Ethics* 23 (2): 152-62.

[21] _____. 2014b. “The Asymmetrical Contributions of Pleasure and Pain to Subjective Well-Being.” *Review of Philosophy and Psychology* 5 (1): 135-53.

[22] Wager, T.D., L.Y. Atlas, M.A. Lindquist, M. Roy, C. Woo, and E. Kross. 2013. “An fMRI-Based Neurologic Signature of Physical Pain.” *New England Journal of Medicine* 368: 1388–97.

Neurofeedback-Based Moral Enhancement and Moral Reason

Koji Tachibana, Kumamoto University, Japan

Abstract

Some neuroethicists criticize the very possibility of moral bioenhancement techniques because a moral state acquired through bioenhancement techniques is not actually moral; such a state is neither reached through moral reasoning nor accompanied by moral reason. I will examine this criticism and argue that neurofeedback-based moral enhancements can overcome the criticism. Neurofeedback-based moral enhancements may not directly endow individuals with moral reasons, but it can do so indirectly. Furthermore, even if these enhancements cannot occur even indirectly, this device can be a tool for moral education because a person need not provide a moral reason to be/become legitimately moral. Therefore, neurofeedback-based moral enhancements can be acceptable tools for moral education, even if they do not enhance any moral reason or reasoning.

I A Critique of Moral Bioenhancement in General

The possibility of moral bioenhancements has been subject to criticism on the grounds that any proposed bioenhancement technique cannot be called *moral* because it cannot endow individuals with moral reasons. For example, John Shook and James Giordano (2016, 118–19) write that

> a subject can produce different moral judgments without anyone, including experimenters, understanding which components of moral cognition have been adjusted and why those adjustments caused differing moral judgments. Subjects would be unable to say why they think differently about moral matters, even in the ordinary terms of folk moral psychology [...]. There is no promise that the subject will introspectively grasp why.

Similarly, John Harris (2016, 270) rejects the possibility of moral bioenhancements, claiming that

> morality was basically a matter of choosing what is for the best all things considered, not simply being well motivated or pro-social; in short that to be good is not simply happening to do no evil but choosing for a reason, choosing on the basis of evidence and argument, not to do wrong.

This criticism is comprised of the following two-fold assumption. First, moral facts contrast with moral reasons—i.e., moral facts concern the fact that a particular emotion, judgment, or behavior is morally appropriate in a certain situation, whereas moral reasons concern the understanding why such a particular emotion, judgment, or behavior is morally appropriate in the situation. Second, morality must contain moral reasons as well as moral facts—i.e., for a person's emotion/judgment/behavior to be moral, it must be based on his/her own moral reasoning or moral reason(s). Variations on those assumptions can be observed widely in the history of moral philosophy. For example, when distinguishing between *Legalität* (legality) and *Moralität* (morality), Immanuel Kant (1788) states that a moral behavior must be something that not merely corresponds with what to do but also derives from the respect for moral law or from duty to the behavior—the Kantian notion of moral reason. Another example can be found in Aristotle (1998), who says that one can conduct a right behavior based on a wrong reason/reasoning; accordingly, virtue requires *ho orthos logos* (a right reason/reasoning). Even though traditional moral philosophers have not reached an agreement on what kind of reason is moral, they seem to agree that moral facts are different from moral reasons and that morality requires moral reasons. (However, in the later sections, I will refine the view on Aristotle's notion of moral reason.)

The internal link between morality and moral reason can imply the deservingness of moral education. For an educational practice to deserve to be called *moral* education, it must endow individuals with moral reasons or the skill of moral reasoning: even if a practice teaches them moral facts, it does not deserve to be called *moral education* if they cannot come to understand the reasons for the moral facts. Hence, inflicting corporal punishment on children to enforce moral behaviors is not a moral education because such behaviors are the result of fear rather than of moral reasons. Since this implication applies not only to traditional teaching methods and aids but also to novel techniques (e.g., moral bioenhancements), proposed moral bioenhancement techniques, such as pharmaceuticals or surgical operations, do not deserve to be called moral enhancement nor the tools for moral education if the techniques do not endow individuals with moral reasons.

Three options are available to rebut such a criticism. The first option is to claim that moral bioenhancement techniques can directly endow individuals with moral reasons, as some claim (see, e.g., Kabasenche 2016; Persson and Savulescu 2016). This claim can only be verified with evidence showing that a moral bioenhancement technique does endow individuals with moral reasons. This is purely an empirical issue and not supported at present because such a technique has yet to be realized. The second option is to claim that a proposed moral bioenhancement technique can be an authentic tool for moral education, arguing that such a technique can indirectly endow individuals with moral reasons. This option comprises several types of arguments, two of which I propose in Sections II and III of this paper.

The third option goes further and argues that a proposed moral bioenhancement technique can be a tool for moral education even if it cannot endow individuals with moral reasons even indirecty. I consider this option in Section IV of this paper.

II Activating Moral Reasons by Overcoming the Weakness of the Will

The first argument for authentic moral education through bioenhancements concerns the cases of the weakness of the will, in which a participant understands a relevant moral fact and even has a reason to act in accordance with that fact but cannot activate it due to the weakness of his/her will. Imagine a case in which a wife confronts her husband for his cruelty. He berates her whenever she does not complete the housework on her own, and he scolds their seven-year-old son whenever the latter does live up to his fatherly expectations. These conflicts make the husband/father worry that his family is becoming dysfunctional, and he knows that his wife and son do their best and that he should help them, but he never supports his son or helps his wife with the housework. Thus, he knows what to do and why. He wishes he were more sympathetic with his son and wife, wanting to help them more, but he cannot stop seeing them as weak when they fail to achieve his desired results.

In this case, the husband/father understands the moral fact as well as the moral reason, but he lacks the will to activate that reason. To change himself, he may visit a clinic where a doctor prescribes oxytocin. As oxytocin is expected to enhance a patient's tenderness, a constant absorption of this pharmaceutical could make him sympathize more with his wife and son (see Churchland 2011, ch. 3–4). He may also visit a "cosmetic and ethical neurosurgery" and receive a course on neurofeedback-based training to enhance his tenderness (see Moll et al. 2014) or to cognitively reappraise his cruelty (see Sarkheil et al. 2015). After a few weeks of training, he could successfully sympathize with others more than ever.

Whatever the method he adopts, pharmacological or neurofeedback-based bioenhancements, he would become more kind to his family, putting his supportive behavior into practice. Enhancing his sympathy for others, he uses such moral bioenhancement techniques to activate his moral reasons that leads to moral behaviors. He can also explain his helpfulness to his family by saying "because you are in a jam" in its true sense. This means that these techniques can endow individuals with a moral reason to act in the sense that a person can activate a reason that he/she already understands as morally right but could not previously practice due to the lack of related emotions that trigger the behavior the reason justifies.

III Endowing Moral Reasons through Existing Education Networks

The second argument accepts the criticism that a proposed moral bioenhancement technique might not, in itself, be able to endow individuals with moral reasons in any sense. Although it can endow them with moral facts, this argument claims that those who acquire moral facts can also acquire moral reasons with the aid of teaching practices. For example, Glannon (2015, 1261) alluded to this:

> it is doubtful that pharmacological modification of our cognitive and affective capacities alone would make us more responsive to moral reasons when acting. [...] Yet the right type of education could complement psychopharmacology in making us more responsive to the interests of those who exist now and those who will exist in the future. The social environment also influences brain function and provides cues that prompt us to act in different ways. Moral enhancement would thus require a full complement of education, environmental modification, and psychopharmacological intervention.

As Glannon expects, the traditional moral education network, including schools, churches, and local communities, are able to play the role of endowing individuals with moral reasons by teaching why a particular emotion or motivation is morally appropriate in a certain situation.

However, the pharmacological moral enhancements that Glannon (2015) had in mind do not seem to be a good option because they have at least three defects as tools for moral education. First, chemical intervention, to an extent, is invasive to the human mind and body. Such invasiveness would not fit with the ideal of current educational standards due to the norm that morality must be educated through noninvasive and voluntary changes. Second, pharmaceuticals are a sort of automatic chemical modification. Such a modification is not compatible with social norms, including the value of effort and authenticity; we assume that one's authentic moral performance must be acquired through his/her own efforts. Third, pharmacological moral enhancements do not contribute to moral diversity because they are not varied enough to meet our moral complexity. Rather, they tend to be limited to popular effects, such as sociability and cheerfulness. In short, pharmacological moral enhancements can threaten moral diversity and promote moral uniformity.

Neurofeedback-based moral enhancements can avoid these problems because they have unique features, including noninvasiveness, the requirement of a participant's effort, and flexibility in targeted moral faculties (Collura 2014). By virtue of these features, neurofeed-

back-based moral enhancements can save morality and, accordingly, harmonize with existing moral education networks, being part of such networks (Tachibana 2017, 2018). It can also lead its participants to use other teaching methods for endowing them with moral reasons because those who have learned moral facts through neurofeedback-based training will become more open to the opportunities to learn the reasons. Then, other teaching methods in the network, such as instructions from school-teachers or parents, can more easily lead subjects to understand moral reasons. Therefore, being part of the educational network, neurofeedback-based moral enhancements can, in itself, only indirectly and sequentially endow a subject with moral reasons. It is indirect because it needs to work in tandem with other teaching methods. It is also sequential because moral reasons are only learnt after the subject is endowed with moral facts.

This sequence would not deprive neurofeedback-based moral enhancements of the title of *moral education* because, as is typical in upbringing and religious education, it often happens that a learner who does not understand the reason for a moral fact at first comes to understand the reason later by following orders or a doctrine. Because upbringing and religious education require time and experience to endow the subject with an understanding of moral reasons, it does not deprive him/her of authentic education; nor do neurofeedback-based moral enhancements. The indirectness is also acceptable, as Aristotle (1998, I4, 1095b7) discerned when he compared "fact (*to hoti*)" with "reason (*to dioti*)". For Aristotelian virtue ethics, we should learn the fact first through habituation or repetition of morally right behaviors; we then gradually understand the reason for the fact through teaching—teaching here means a sort of lecture on ethics, such as "Aristotle's lectures" (Burnyeat 1980, 71–72).

In the case of neurofeedback training, a trained person comes to understand firstly moral facts—feeling/recognizing/behaving in a certain way—without fully understanding the reason why that particular feeling/cognition/behavior is morally desirable. However, he/she comes to understand the reason gradually through his/her everyday experiences and other teaching methods. Such experiences and methods fine-tune what was learned through the neurofeedback training to fit into actual social environments inasmuch as principles of upbringing or tenets of religion have been fine-tuned through actual social environments. Therefore, working as part of a traditional moral education network, neurofeedback-based moral enhancements can endow individuals with moral reasons in an indirect but legitimate way.

IV Morality without Moral Reasons

The third option doubts the very assumption that morality requires moral reasons. This option comprises at least three arguments. First, we can legitimately be moral even if we have moral facts without moral reasons for the facts. This argument can be seen in Hesiod's words, as quoted in Aristotle's *Nicomachean Ethics* (I4, 1095b10–13 [Aristotle 1998, 6]):

> Far best is he who knows all things himself;
> Good, he that hearkens when men counsel right;
> But he who neither knows, nor lays to heart
> Another's wisdom, is a useless wight.

A person who listens to and remembers a virtuous person's rational advice is said to be "good (*esthlos*)" though not "far best (*panaristos*)." Furthermore, just before this quote, Aristotle even declares that such a good person "will not need the reason as well" (*NE* I4, 1095b6–7 [Aristotle 1998, 4–5]). Therefore, as Burnyeat (1980, 71) puts it, "Aristotle quotes the Hesiodic verses in all seriousness." This quote does not deny that those who have moral reasons as well as moral facts are morally desirable. However, it also admits the latter are still legitimately moral. Consequently, an educational practice that can teach moral fact, but not moral reasons, is also a legitimately moral education. Following Aristotle's notion of morality (and presumably also some sort of consequentialism at least), having moral reasons is not a necessary condition for a person to be morally good.

Second, emotions, also, can play the role of moral reasons. A German notion, *Mitleid*—an approximate English translation is "compassion"—which Schopenhauer (1819) formulates as the only basis of human morality, is a typical case; a behavior based on Mitleid is said to be morally good in all things. Sympathy is another example. As is typical in cases of euthanasia or dying with dignity, the fact that a person suffering from a severe disease can arouse sympathy enough to give a moral reason to kill the person, whereas the fact that sunlight is blinding does not arouse any emotion which legitimately give a moral reason to kill a man (Camus 1942). In general, some emotions, such as compassion and sympathy, can thus be reasons for moral action.

Third, moral facts do not necessarily require knowledge of moral reasons if correctly used. Imagine two cases in which students acquire theoretical knowledge without understanding the reasons for that knowledge. In one case, a student correctly writes the atomic symbol "He" next to "H" on a short examination about the periodic table of elements because he remembers a mnemonic device such as a pun; in the other case, a student mechanically, routinely, and correctly applies a formula, such as Pythagorean Theorem, to a question on

a mathematics test. If they were asked to explain why their answers were correct, they could say no more than "because a pun song says *He* is next to *H*" or "because I applied Pythagorean Theorem." In these cases, the mnemonic device and the formula work as a black box that does not provide the reasons but conclude the correct answers in an acceptable way. This suggests that theoretical knowledge can, to some extent, be approved to be legitimate without giving a reason for that knowledge.

The same logic applies to the case of sociomoral knowledge. If you are invited to a house party that starts 6:00 p.m., your chosen time to arrive is culturally dependent. Supposedly, 6:05 p.m. would be a correct time of arrival in the U.S., whereas in Japan, 5:55 p.m. would be correct. Table manners are another example: slurping pasta is ill-mannered in one culture, whereas eating pasta quietly can be relatively ill-mannered in another culture (e.g., soba noodles in Japan). In both cases, we have difficulty in explaining the reason why a certain manner is right (or wrong) in a given culture because understanding such reasons requires a broad and complicated knowledge of cultures, religions, histories, and so on. However, the lack of understanding the reasons does not deny that we know what to do morally in each social situation. In this sense, sociomoral knowledge is exempt from providing reasons to some extent. It can lead to a radical form of the exemption—namely, to the point where we cannot give any better explanation than "that is what I/we do." As Wittgenstein (1958, §217) puts it, we reach the "bedrock".

These three arguments suggest that we should doubt the assumption that morality requires moral reasons. If giving a reason is not a necessary condition for being/becoming moral, neurofeedback-based moral enhancements can be a tool for moral education.

V Conclusion

The possibility of moral bioenhancements has been subject to various criticisms. However, the criticism on the grounds that any proposed bioenhancement technique cannot endow individuals with moral reasons will be rejected for two reasons. First, neurofeedback-based training will be able to endow individuals with moral reasons in indirect ways. Second, this training can be a tool for moral education even if it cannot endow them with such reasons in any sense, because bestowing moral reasons is not a necessary condition for a proposed moral education to be a legitimate teaching method. These two reasons that this paper has sketched entitles neurofeedback-based moral enhancement to be a legitimate tool for cultivating human's morality.

References

[1] Aristotle. 1998. *Nicomachean Ethics*. Translated by W. D. Ross. Revised by J. L. Ackrill and J. O. Urmson. Oxford: Oxford University Press.

[2] Burnyeat, M. 1980. “Aristotle on Learning to Be Good.” In *Essays on Aristotle’s Ethics*, edited by A. O. Rorty, 69–92, Berkeley: University of California Press.

[3] Camus, A. 1942. *L’Étranger*. Paris: Gallimard.

[4] Churchland, S. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton: Princeton University Press.

[5] Collura, T. F. 2014. *Technical Foundations of Neurofeedback*. New York: Routledge.

[6] Glannon, W. 2015. “Reflections on Neuroenhancement.” In *Handbook of Neuroethics*, edited by J. Clausen and N. Levy, 1251–65. New York: Springer.

[7] Harris, J. 2016. “Moral Blindness—The Gift of the God Machine.” *Neuroethics* 9: 269–73.

[8] Kabasenche, W. 2016. “Moral Formation and Moral Enhancement.” *AJOB Neuroscience* 7 (2): 130–31.

[9] Kant, I. 1788. *Kritik der praktischen Vernunft*. Riga: J. F. Hartknoch.

[10] Moll, J., J. H. Weingartner, Bado, R. Basilio, J. R. Sato, B. R. Melo, I. E. Bramati, R. de Oliveira-Souza, and R. Zahn. 2014. “Voluntary Enhancement of Neural Signatures of Affiliative Emotion Using fMRI Neurofeedback.” *PLOS ONE* 9 (5): e97343.

[11] Persson, I., and J. Savulescu. 2016. “Moral Bioenhancement, Freedom and Reason.” *Neuroethics* 9: 263–68.

[12] Sarkheil, R., A. Zilverstand, N. Kilian-Hütten, F. Schneider, R. Goebel, and K. Mathiak. 2015. “fMRI Feedback Enhances Emotion Regulations Evidenced by a Reduced Amygdala Response.” *Behavioral Brain Research* 281: 326–32.

[13] Schopenhauer, A. 1819. *Die Welt als Wille und Vorstellung*. Leipzig: F. A. Brockhaus.

[14] Shook, J., and J. Giordano. 2016. “Moral Enhancement? Acknowledging Limitations of Neurotechnology and Morality.” *AJOB Neuroscience* 7 (2): 118–20.

[15] Tachibana, K. 2017. “Neurofeedback-Based Moral Enhancement and the Notion of Morality.” *Annals of the University of Bucharest: Philosophy Series* 66 (2): 25–41.

[16] _____. 2018. “Neurofeedback-Based Moral Enhancement and Traditional Moral Education.” *Humana Mente: Journal of Philosophical Studies* 11 (33): 19–42.

[17] Wittgenstein, L. 1958. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Oxford: Blackwell.

Creating an Obligation: Bentham and the Normative Question

Piero Tarantino, Sciences Po Law School Paris, France

Abstract

An unrecognized and unexplored development in the history of thought is the linguistic account of the normative character of standards of behaviour, practical norms, moral values, legal rules and religious principles provided by Jeremy Bentham. A recent and substantial attempt to fill this historical and theoretical gap is my research monograph *Philosophy, Obligation and the Law: Bentham's Ontology of Normativity* (2018), which offers a comprehensive investigation into Bentham's theory of real and fictitious entities, and – in particular – examines its application to the fields of morality and law. After providing a short account of the debate devoted to the presentation and discussion of *Philosophy, Obligation and the Law: Bentham's Ontology of Normativity*, during the ISUS Conference 2018 in Karlsruhe, the present paper aims to give a general overview of the book. In the first part, I will try to throw light on the Bentham's contribution to the so-called *normative question*; I will then explain the methodology I adopted and, in connection with this, outline the structure and the content of the book; finally, I will focus on certain crucial aspects of my reconstruction and interpretation of Bentham's view of the ontology underlying the normative realm.

Introduction

An unrecognized and unexplored development in the history of thought is the linguistic account of the normative character of standards of behaviour, practical norms, moral values, legal rules and religious principles provided by Jeremy Bentham. A recent and substantial attempt to fill this historical and theoretical gap is my research monograph *Philosophy, Obligation and the Law: Bentham's Ontology of Normativity* (2018), which offers a comprehensive investigation into Bentham's theory of real and fictitious entities, and – in particular – examines its application to the fields of morality and law.[1]

The International Society for Utilitarian Studies (ISUS) Conference 2018 in Karlsruhe was the ideal place in which to give the first official presentation of *Philosophy, Obligation and the Law: Bentham's Ontology of Normativity*. The debate on this book served to draw attention to the normative aspects and implications of Bentham's theory of real and fictitious entities from an interdisciplinary point of view. It was an occasion that brought together leading Bentham scholars who are specialists in different fields, more specifically philosophy, law,

[1] For more information about the book, please visit https://www.routledge.com/Philosophy-Obligation-and-the-Law-Benthams-Ontology-of-Normativity/Tarantino/p/book /9781138496576.

history, political science and English studies: (in alphabetical order) Malik Bozzo-Rey, Emmanuelle De Champs, Gianfranco Pellegrino and Philip Schofield. This debate was of interest not only to those studying Bentham's thought, but also to those wishing to understand the historical roots of the contemporary *normative question* as an investigation into the action-guiding claim of the practical domain.

Each speaker dealt with one of the four interrelated aspects, examined in my book, of Bentham's theory of real and fictitious entities: ontology, epistemology, normativity and motivation: Schofield focused on the place of logic and language in the future direction of Bentham studies; De Champs reconstructed the historical sources of Bentham's theory of real and fictitious entities; Bozzo-Rey explored the function of norms and obligations in Bentham's theory of law; Pellegrino compared Bentham's view of fictitious entities with the metaethical framework of contemporary fictionalism.

After providing a short account of the debate devoted to the presentation and discussion of *Philosophy, Obligation and the Law*, during the ISUS Conference 2018 in Karlsruhe, the present paper aims to give a general overview of the book. In the first part, I will try to throw light on the Bentham's contribution to the so-called *normative question*; I will then outline the structure and the content of the book; finally, I will focus on certain crucial aspects of my reconstruction and interpretation of Bentham's view of the ontology underlying the normative realm. This paper makes continuous reference to *Philosophy, Obligation and the Law*, by summarizing its main topics, which are examined at length in the book.

I Bentham and the Normative Question

Bentham's thought is in the seventeenth- and eighteenth-century British moral tradition, namely the philosophical context from which the current normative question, in the sense of an investigation of the foundations of practical reality, arose. In morality, law and religion, norms, rules, values and virtues do not merely express a belief about an action, but they recommend its approval and adoption. In this sense, practical notions entail an obligation, inasmuch as they exercise the prerogative to make a claim on their subjects' conduct in order to influence it. Their normativity consists in this prerogative, namely in this guidance claim, having a binding force.

The notion of obligation is at the heart of the concept of normativity. The understanding of the claim that an obligation makes on us to obtain compliance is the objective of the normative question. It requires the enlightenment of the roots of practical disciplines such as

morality, law and religion. It is important to clarify why moral, legal and religious rules have an action-guiding authority, thereby uncovering what endows them with a normative feature. The ultimate purpose of the normative question, therefore, is to explain and justify the claim of morality, law and religion to direct the agent's behaviour.

Bentham was deeply influenced by the view of an empirical basis of morality. Yet he reconciled that sensory approach with the idea of the linguistic construction of ethical elements. Bentham understood the concept of obligation as a fictitious entity, i.e. a name, created by the human mind in relation to the feelings of pleasure and pain, which are regarded as real entities. The desire to enjoy pleasure and avert pain makes binding the action instrumental in fulfilling that desire. Obligation is thus the result of harmonious cooperation between sensibility and intellectual activity, including particularly the faculties of language and imagination.

The fictitious notion of obligation has meaning and truth only if it is referred to pleasure and pain: they are the pillars on which the human mind builds the practical domain. The binding and motivating force of an obligation, in which its character consists, springs from pleasure and pain, which are individually perceived. So, the investigation of the nature of an obligation is joined to an investigation of its normative foundations. The ontological and normative – as well as the epistemological and motivational – features of an obligation depend on pleasure and pain. From these empirical roots the human mind can create a fictitious ontology endowed with normativity. Ethical elements turn out to be human artefacts, anchored to physical reality: they can make a guidance claim on us by virtue of their constitutive relation with pleasure and pain, which have the form of reward or punishment.

Although fictitious, an obligation purports to direct human behaviour. Bentham's ontology of the practical world, which is based on the distinction between real and fictitious entities, needs to be scrutinized in its normative and motivational facets. Bentham's account of ethics is intertwined with his theory of real and fictitious entities; consequently, his linguistic approach to ontology provides the key to interpreting his view on the constitution of morality and law. Moral and legal elements are fictitious entities, that is, experience-based products of the human mind, whose operations have to be identified and analysed in order to uncover the intimate nature of morality and law and their function in regulating human action. When conceiving of an obligation as a fictitious entity, one needs to explain how it can influence an agent's conduct by obtaining his/her compliance. Dealing with such issues requires first a reconstruction of Bentham's theory of real and fictitious entities and then

the examination of its application to the ethical field, in order to clarify the ability of morality and law to provide people with reasons for action.[2]

II The Structure of the Book

By focusing on the concept of obligation, *Philosophy, Obligation and the Law* explores Bentham's ontological and linguistic view, and aims to identify the specific features of ethical fictitious entities. The book is divided into two parts. In the first part I examine the ontological and epistemological foundations of Bentham's distinction between real and fictitious entities, whose interrelations provide the framework of the natural world and the ethical world, as they are represented by human beings. In the second part I seek to throw light on both the normative and motivational aspects of moral and legal notions, including an obligation, according to Bentham.

In pursuit of these aims, I focus on logic, theory of language, physics, metaphysics, metaethics, axiology, the doctrine of virtue, the freedom of the will, the structure of practical reasoning and action with reference to the law. Despite its richness and complexity, Bentham's treatment of these topics is much neglected in philosophical literature. Only few scholarly works tackle it and most of these are not recent. Nevertheless it is important to show the centrality of these issues to Bentham's legal reform.

Understanding, from Bentham's perspective, what an obligation is and how human beings become aware of it brings us to an investigation into the general nature of a fictitious entity. Only by throwing light on the framework of the human representation of the world, basically articulated according to the connection between real and fictitious entities, can the ontological and the epistemological character of the notion of obligation be identified and, consequently, the structure of ethics disclosed. Obligation is, in fact, the constitutive ingredient of practical reality as a whole, characterizing with its normative force each single component. Along with obligation, on which our present focus lies, practical concepts such as virtues, values, standards of behaviour and norms have a guidance function. They make a normative claim on us, which needs to be explored in the light of the distinction between real and fictitious entities.

[2] Cf. Tarantino 2018, 1-9, in which the relation between Bentham's philosophy and the normative question is fully explained.

My work is committed to exploring Bentham's ontology and theory of language with the purpose of identifying the general features that ethical fictitious entities, including obligation, share with other kinds of fictitious entities, and then of bringing out the particular aspects which distinguish the ethical domain. In this way, it is possible to trace, in Bentham's philosophy, a path that guides us from the comprehensive definition of a fictitious entity, with reference to its relation to a real entity, to the characterization of its ethical specifications and, finally, to the grasping of the normative function of an obligation, which is the core element in the constitution of the practical sphere (Tarantino 2018, 11-12).

III The Content of the Book

Bentham maintained a lifelong commitment to assessing the nature of various practical domains such as morality, law and also religion. As their distinctive property, the elements making up practical domains are characterized by a linguistic ontology, which is endowed with normativity, that is, with the property to make claims on the agents' behaviour to require obedience. In other words, concepts such as duty, goodness, rightness and virtue have a directive or guidance function, because they prescribe or recommend the performance of the act to which they refer.

The normative claim essentially characterizes ethical standards; therefore, any investigation into their authoritative feature is an investigation into the foundations of ethics and the obligation entailed by them. The distinction between real and fictitious entities provides Bentham with the ontological and epistemological framework to outline the structure of normative domain, with special reference to its constitutive relation with human motivation, and then bring about the reform of morality and law.

We can single out certain crucial points in my reconstruction and interpretation of Bentham's ontology of normativity: (a) the constructive function of language, (b) the fictitious framework of human knowledge, (c) the naturalistic foundations of the normative domain, (d) the instrumental connection between normativity and motivation. Let us re-examine these points in short, thereby exploring how they are related to the controversy on the status of normativity.[3]

[3] Cf. Tarantino 2018, 219-24 for an overview of the main topics of the book, which are outlined in this section of my paper.

III.a The Constructive Function of Language

Terms such as motion, quality, relation, cause, virtue, goodness, obligation, right and power represent mere artefacts characterized by a linguistic form of existence, as they are names produced by the activity of the human mind carried out on perceptual elements. On the other hand, real entities are names denoting empirical objects, whose existence lies in human perception. Bentham stresses the linguistic nature of real and fictitious entities which are, precisely speaking, names of real entities and names of fictitious entities included in a sentence on which their meaning depends. Real and fictitious entities, along with their relations, are resolved in language which accounts for them as nouns.

In Bentham's philosophy then, language has an ontological and epistemological function, since it contributes to our representation and construction of the world. Our imagining of the external world is shaped by language which plays a performative role, by providing a form of existence, though fictitious, to the framework of the natural and the practical realms. This linguistic ontology, depending on the creative power of the human mind, is complementary to and interrelated with the empirical ontology captured by sense-perceptions and consisting in impressions and ideas. Bentham emphasizes the existence-conferring ability of language, which enables human beings to consistently structure their knowledge and coordinate their behaviour in society (Tarantino 2018, 220-21).

III.b The Fictitious Framework of Human Knowledge

Bentham questioned the ontological and epistemological nature of the elements making up the mathematical, the physical, the moral, the legal and the religious domains of knowledge. In particular, he challenged the effective correspondence of beliefs and statements to an alleged external reality, on which the idea of truth resides, thereby assuming that the human mind has through language autonomous constructive ability in conceiving of the world.

Fictitious objects are contrivances with a theoretical and a practical function. From a theoretical standpoint, they are instrumental in our image of the world, giving an order to perceptions and making sense of experience; consequently, our representation of it is the outcome of an organizational process of the human mind. From a practical standpoint, fictitious objects make reasoning and decision possible, by orienting and directing people's individual and collective behaviour. According to Bentham, the activity of the human mind consists in creating fictitious entities; without these entities knowledge and action, and more generally

social life, would be impossible. In other words, fictitious objects allow people to think, communicate and act.

The relation between real and fictitious entities lies at the basis of Bentham's account of the natural world and of the science of morality and law. It not only provides the basis for his political and legal reform, but also underpins the linguistic ontology characterizing the various fields of knowledge. Bentham's approach to fictitious entities, thus, affords a critical point of view to rethink the constitution and the structure of the mathematical, the physical, the moral, the legal and the religious realms (Tarantino 2018, 220-21).

III.c The Naturalistic Foundations of the Normative Domain

The relation that a fictitious entity maintains with its empirical source, identified as a real entity, is constitutive of its import and truth. Divorced from reality, a fictitious entity is nothing but falsehood and nonsense. Paraphrasis, as a linguistic method to disentangle the obscurities and ambiguities of fictitious entities, reveals the empirical foundations on which the construction of the practical realm resides. So the fictitious phrase *An obligation is incumbent on a man* should be explained as *Pain is incumbent on a man if he does not act in compliance with an obligation*. Pain then is causative of an obligation, that is, of the binding power in which it consists.

Bentham conceives of values, duties, virtues and standards of behaviour as fictitious entities, namely as linguistic elements, whose import and truth depend on the empirical perceptions of pain and pleasure, which are real entities. Ethical standards turn out to be human artefacts, related to perceptions; more precisely, their guidance claim on us lies in the constitutive relation ethical standards have with pain or pleasure. Indeed, pain and pleasure are the foundations of the ethical domain and of its peculiar normative claim.

Normativity, as the distinctive property of the ethical area, consists in the ability to provide an agent with reasons for action. From Bentham's perspective, a reason for action is normative inasmuch as it is based on pain or pleasure. Values, virtues, norms, commands and standards of behaviour purport to direct an agent if they are connected with pain or pleasure; their normative, that is, their action-guiding prerogative derives from these perceptions. In Bentham's view normativity has naturalistic foundations which, though external to the agent, rely on his/her psychological structure: his/her desire to enjoy or maximize pleasure and his/her desire to avoid or minimize pain direct and guide the agent to choose the

course of action which leads him/her to the achievement of that pleasure or to the avoidance of that pain.[4]

IV The Instrumental Connection between Normativity and Motivation

By putting forward his model of normativity, Bentham contributes to the shaping of the early modern idea of obligation as an internal requirement of a self-governing moral agent's thinking. The creation of an obligation, along with the notions involving it, such as rights, values and virtues, depends on the ability of its issuing authority to arouse in the agent his/her desire to avoid punishment, deriving from disregarding that obligation, or his/her desire to gain praise, deriving from complying with that obligation. Punishment and praise are sanctions, consisting in pain or pleasure, which drive the agent to choose a course of action.

The connection that an obligation has with motivation is crucial. In order to feel the pressure to conform his/her action to a standard of behaviour, a pressure in which the normative claim resides, an agent needs to perceive the pain and the pleasure related to that standard and, then, to be motivated to act from that perception. Ethical values provide the agent with guidance, by recommending him/her to behave in a certain way; their guidance function springs from the pain or pleasure that the conformity with these values is expected to bring forth.

The agent's sensibility to the pain and pleasure flowing from a sanction causes in the agent a motive for avoiding that pain and attaining that pleasure. This motive creates the obligation to adopt suitable means to that end. In Bentham's outline of practical judgement, which is a form of instrumental rationality, desire generates requirements to act or to forbear from acting. The normative force of a sanction appears to be dependent on its motivational power. Motivation has indeed a causative role in obligation and then action. According to Bentham's principle of the determination of action, the agent constrains himself/herself to endorse a form of conduct when that form of conduct is instrumental in the achievement of his/her end. Put differently, an agent imposes on himself/herself a duty, compliance with which is aimed at the fulfilment of his/her interest.

[4] See Tarantino 2018, 221-22 and, for a full examination of Bentham's method of paraphrasis applied to the notion of obligation, 98-104.

The relation between normativity and motivation, namely between duty and desire, is thus fundamental to understanding Bentham's idea of ethics and of its decision- or action-guiding claim. Something can bind an agent to adopt a certain behaviour inasmuch as this thing motivates the agent to behave accordingly. Motivation consists in the desire to achieve the pleasure or escape the pain stemming from the adoption of that behaviour. Therefore, one can say that desire makes an action binding. The agent's volition and, thus, his/her behaviour is moved by his/her desire to avoid pain and to gain pleasure. As they are the motivating factors of deliberation and action, pain and pleasure can be considered the sources of normativity; in fact, obligation springs from desire for pleasure and aversion to pain.

Bentham regards ethical entities and the normative phenomena characterizing them as resulting from the agent's motivation. Moral and legal obligation or, more generally, normativity is a linguistic creation based on the real entities of pain and pleasure, which excite individual motivation. The motivating aspect of pain and pleasure turns out to be the key to understanding Bentham's ontology of normativity.

An action is right and thus worthy to be performed inasmuch as it leads to pleasure or, at least, it entails the reduction of pain. The rightness of that action resides in its utility and this utility is measured by the agent himself/herself according to his/her receptiveness to a certain pleasure or a certain pain. Despite appearances, Bentham does not put forward a merely reductionist theory of normativity, according to which obligation depends on an external fact, i.e. a sanction. No doubt, it is true that sanction, as a source of pain or pleasure, is an objective event related, for example, to the compliance or non-compliance with a command prescribed by an authority. However, it is also true that the pain and the pleasure deriving from a sanction are subjective perceptions; in other words, pain and pleasure need to be felt by the agent so that they can make their normative claim. Bentham traces back normativity to the agent's conative states, such as desires, interests, dispositions and wants.[5]

V Conclusion

Bentham rethinks this connection between normativity and motivation within a general fictitious context based on language. Bentham's idea of obligation, as a fictitious construction

[5] See Tarantino 2018, 222-23 and, for an examination of the relation between obligation and sanction in terms of the relation between normativity and motivation, 199-210.

of the human mind depending on naturalistic foundations, is an original position in the debates on normativity. By virtue of this, Bentham's approach enables us to re-assess the premises of the normative question and to explore it from a new perspective.

Bentham's idea of the ontology of ethics as a linguistic creation of the human mind underpins his theory of normativity. The ethical domain, however, is not an arbitrary construction; this is the reason why Bentham levels his criticism against the deceptive use of fictitious entities which is aimed at the protection of the interests of the ruling few. On the contrary, ethics has a firm foundation in empirical reality or, more precisely, in the physiological and psychological constitution of human nature, which is naturally oriented to pursue happiness, namely to seek pleasure and avoid pain. The empirical sources of the practical realm provide the guidelines for the legislator's and judge's decisions, so that they can achieve the greatest happiness of the greatest number. (Tarantino 2018, 223)

References

Bentham's Works

[1] *The Collected Works of Jeremy Bentham*, general editors Burns J. (1961-79), Dinwiddy J. (1977-83), Rosen F. (1983-95), Rosen F. and Schofield P. (1995-2003), Schofield P. (2003-), London: The Athlone Press, 1968-81; Oxford: Clarendon Press, 1983- in progress.

[2] *De l'ontologie et autres textes sur les fictions*, texte anglais établi par Schofield P., traduction et commentaires par Cléro J.-P. et Laval C., Paris: Éditions du Seuil, 1997.

[3] Bowring (abbr.): *The Works of Jeremy Bentham, Published under the Superintendence of his Executor, John Bowring*, 11 vols., Edinburgh: Tait, 1838-43.

Studies

[4] Bouveresse, J. 1993. "*La théorie des fictions chez Bentham*." In *Regards sur Bentham et l'utilitarisme,* edited by K. Mulligan, and R. Roth, 87-98. Genève: Librairie Droz.

[5] Bozzo-Rey, M. 2009. "Loi, fiction et logique dans la pensée juridique de Jeremy Bentham" *Annales de Droit* 3: 27-50.

[6] _____. 2014. "Reducing the Limits of the Realm of Possibilities: Law, Action and Will in Jeremy Bentham's Thought" in *Tusseau*: 338-57.

[7] Cléro, J.-P. 1993. “La théorie des fictions chez Jeremy Bentham” *Nouvelles de la République des Lettres* 2: 47-71.

[8] _____. 2000. “La valeur d’une théorie des fictions” *Laval théologique et philosophique* 56: 439-61.

[9] _____. 2014. *Essai sur les fictions*. Paris: Hermann.

[10] Darwall, S. 1995. *The British Moralists and the Internal “Ought”: 1640-1740*. Cambridge: Cambridge University Press.

[11] De Champs, E. 1999. “The Place of Jeremy Bentham’s Theory of Fictions in Eighteenth-century Linguistic Thought” *Journal of Bentham Studies* 2: 1-28.

[12] _____. 2015. *Enlightenment and Utility: Bentham in French, Bentham in France*. Cambridge: Cambridge University Press.

[13] Haakonssen, K. 1996. *Natural Law and Moral Philosophy: From Grotius to the Scottish Enlightenment*. Cambridge: Cambridge University Press.

[14] Harrison, R. 1983. *Bentham*. London: Routledge & Kegan Paul.

[15] Hart, H. 1982. *Essays on Bentham*. Oxford: Clarendon Press.

[16] Hume, L. 1981. *Bentham and Bureaucracy*. Cambridge: Cambridge University Press.

[17] Jackson, B. 1998. “Bentham, truth and the semiotics of law” *Current Legal Problems* 51: 493-531.

[18] Kelly, 1990. *Utilitarianism and Distributive Justice*: *Jeremy Bentham and the Civil Law*. Oxford: Clarendon Press.

[19] Korošec, G. 1994. “The Role of Fictions in Law: Hume, Adam Smith and Bentham” *Filozofski Vestnik / Acta Philosophica* 15: 151-68.

[20] Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

[21] Laval, C. 1994. *Jeremy Bentham. Le pouvoir des fictions*. Paris: Presses Universitaires de France.

[22] Ogden, C. 1932. *Bentham’s Theory of Fictions*. London: Kegan Paul.

[23] Pellegrino, G. 2010. *La fabbrica della felicità: Liberalismo, etica e psicologia in Jeremy Bentham*. Napoli: Liguori.

[24] Postema, G. 1983. “Facts, Fictions, and Law: Bentham on the Foundations of Evidence” *Archiv für Rechts- und Sozialphilosophie* 16: 37-64.

[25] _____. 1986. *Bentham and the Common Law Tradition*. Oxford: Clarendon Press.

[26] Raz, J. 1975/1990. *Practical Reason and Norms*. Oxford: Oxford University Press.

[27] Quinn, M. 2013. "Fuller on Legal Fiction: A Benthamic Perspective" *International Journal of Law in Context* 9: 466-84.

[28] Rosen, F. 2003. *Classical Utilitarianism from Hume to Mill*. London: Routledge.

[29] Schofield, 2006. *Utility and Democracy: The Political Thought of Jeremy Bentham*. Oxford: Oxford University Press.

[30] _____. 2009. *Bentham: A Guide for the Perplexed*. London: Continuum.

[31] Tarantino. 2018. *Philosophy, Obligation and the Law: Bentham's Ontology of Normativity*. Abingdon/ Oxon/ New York: Routledge.

[32] Tusseau, G. 2011. "Jeremy Bentham et les fictions du droit." In *L'imaginaire en droit*, edited by G. Darcy, and M. Doat, 383-433. Bruxelles: Bruylant.

[33] _____. 2011a. *Jeremy Bentham: la guerre des mots*. Paris: Dalloz.

[34] _____, ed. 2014. *The Legal Philosophy and Influence of Jeremy Bentham: Essays on "Of the Limits of the Penal Branch of Jurisprudence"*. London: Routledge.

Does Adam Smith's Moral Theory Truly Diverge from Humean Utilitarianism?

Hiroki Ueno, Hitotsubashi University, Japan

Abstract

In this paper, I argue that Adam Smith's philosophical system is essentially dependent upon the principle of utility, and that the principle of justice should be regarded as being founded considerably on a utilitarian basis, in the same manner as ascribed to David Hume. This theoretical structure should, first of all, be detected in Smith's moral theory, or ethics as the argument of human social/sociable nature. This view runs contrary to the majority of the literature on *The Theory of Moral Sentiments*, which tends to emphasise that his moral argument should be interpreted as being in contrast with Hume's utilitarian foundation of morality. This insertion and exaggeration of differences between Hume and Smith is partly derived from the deontological interpretation of Smith's moral theory, which usually confuses his explication of the moral motivation of the common people to do the right thing with the criterion for judging the social institution as a whole. The second purpose of this paper is to demonstrate that the relationship between public utility and justice that is latent within his ethics functions as the basis of his practical science of good policing, or political economy. Smith scrutinises the "invisible" mechanisms that enable each social action — referring solely to the natural sense of justice — to result in the most socially beneficial outcomes. And while he does emphasise that individual moral agents seldom regard the consequence of their specific actions upon wider society as a primary interest in their everyday life, this fact itself does not mean that there is no significant part played by the principle of utility within his theory; on the contrary, it has so essential a role that Smith is at pains to acknowledge the exceptional cases in which consideration of the public utility ought to take precedence over the common moral sense of justice.

Introduction: The Exaggeration of Differences between Hume and Smith

The first assertion of this paper is that Adam Smith substantially shares in what might be dubbed the 'liberal utilitarian scheme', among whose leading systematic formulators was David Hume. According to this theoretical framework, the liberal principle of justice is to be respected as much as possible, but could be regarded as subordinate to the principle of utility under certain special or exceptional circumstances. As such, particular attention will be paid to the implicit yet significant role of the "public utility" in Smith's moral theory in order to illustrate what is fundamentally shared by these two Scottish Enlightenment figures. It must be acknowledged here that — though further investigation of the point lies beyond the scope of this paper — this notion of liberal utilitarianism nevertheless remains distinct from "Benthamite" utilitarianism. Hume's scepticism regarding constructivism should be noted in this respect, despite often being attributed rather to Smith. This alone

signifies that Hume, in a similar manner to Smith, must be essentially differentiated from the particular type of utilitarian usually exemplified by Jeremy Bentham. While it is thus distinct from constructivist utilitarianism, however, Smith's liberal moral theory, as well as Hume's, should still be situated within another, separate utilitarian tradition of its own.

I Adam Smith on the Principle of Utility[1]

This section intends to demonstrate that even in Smith's moral theory, let alone Hume's, the principle of public utility actually plays a much more significant role than *The Theory of Moral Sentiments* might present to the casual reader; and that not only his political economy (which will be dealt with in the next section), but also his moral philosophy as a whole should be interpreted as being more intrinsically dependent on the principle of utility.[2]

Smith is explicit in recognising that moral judgements from a utilitarian perspective should have priority over the natural sense of justice in some cases. This occurs almost immediately after the passage in which he asserts that the reasons for a "just/unjust" determination on the part of the common people are not founded on their rational consideration of the public good or interest. Shortly after maintaining that "it is not a regard to the preservation of society, which originally interests us in the punishment of crimes committed against individuals" (TMS II.ii.3.10, 89), Smith somewhat unexpectedly acknowledges the limits of his argument as follows:

> Upon some occasions, indeed, we both punish and approve of punishment, merely from a view to the general interest of society, which, we imagine, cannot otherwise be secured. Of this kind are all the punishments inflicted for breaches of what is called either civil police, or military discipline. (TMS II.ii.3.11, 90)[3]

A "sentinel" (sentry) "who falls asleep upon his watch" is Smith's primary example, offered to demonstrate that consideration of the common or public interest is sometimes necessary in order to justify certain instances of severe laws or administrations of justice.[4] It is deemed

[1] In this paper, all references to Adam Smith are to *The Glasgow Edition of the Works and Correspondences of Adam Smith* (Oxford: Clarendon Press and Indianapolis: The Liberty Fund). The references use the standard abbreviations as listed.

[2] Regarding how Smith himself explains the system of justice (as the fundamental component of the "commercial politics") historically emerging from pre-commercial societies, based primarily on the principle of authority but gradually on the principle of utility, see Berry 2013, 94-96, 101-3.

[3] Cf. TMS II.ii.3.8, 89.

[4] The same example can also be found in LJA ii.92; LJB, 182.

entirely reasonable by Smith himself that the sentinel "suffers death by the laws of war, because such carelessness might endanger the whole army"[5]; yet, according to any natural sense of justice shared by the common people, this level of punishment is too severe to "go along with" in spite of Smith's determination that it is not only necessary but just and proper in itself. He emphasises this gulf between the natural common judgement and what is actually required thus:

> The natural atrocity of the crime seems to be so little, and the punishment so great, that it is with great difficulty that our heart can reconcile itself to it. Though such carelessness appears very blamable, yet the thought of this crime does not naturally excite any such resentment, as would prompt us to take such dreadful revenge. (TMS II.ii.3.11, 90)

If this is so, what renders this seemingly excessive punishment just and proper? Looking elsewhere, Smith elaborates that this level of severity, which a natural common sense cannot abide, can only be justified in the "consideration of the general interest of society" (TMS II.ii.3.7, 88) or the "welfare of society" (TMS II.ii.3.9, 89).[6] This is because crimes of this kind "do not immediately or directly hurt any particular person; but their remote consequences, it is supposed, do produce, or might produce, either a considerable inconveniency, or a great disorder in the society" (TMS II.ii.3.11, 90).

What is of crucial significance in the discussion here is a recognition that the system of justice, including its apparently excessive but actually just and proper penalties, is, in these cases, supported and maintained only by identifying with the interests of the public as a whole. The special cases relating to civil police or military discipline cannot be subjected to nor approved by a natural sense of justice. In substance, this amounts to an admission by Smith that considering the principle of public utility is indispensable in maintaining justice, and that a sense of propriety is not in itself sufficient for this purpose. It is obvious, then, that the consideration of the public utility supersedes any natural basis for judging what is just or right in these cases. For this reason, the perceived distance between Smith and Hume should be regarded as being much less than is generally assumed by many scholars of the former.

If Smith posits that what is just is primarily determined by a natural sense of propriety, without reflecting the interests of society as a whole, but should sometimes be judged according to the public utility nevertheless, then the importance of the principle of utility in Smith is

[5] "This severity may, on many occasions, appear necessary, and, for that reason, just and proper" (TMS II.ii.3.11, 90).

[6] In one of his *Lectures on Jurisprudence*, he says "this [severe punishment] is intirely founded on the consideration of the publick good," that is, "the safety of a multitude." (LJA ii.92).

not so different from that of Hume.[7] Hume does not always relate justice to the conscious consideration of the public interest either. A further point to be made is that, particularly in the cases of maintaining justice within larger societies, it is necessary for people to have a wider appreciation of the public welfare, which is particularly emphasised in Hume's *An Enquiry concerning the Principle of Morals*.[8] Taking this aspect in Smith into account, it is thus almost disingenuous to suggest that Smith is entirely free of and independent from the Humean utilitarian perspective.[9]

II Political Economy: Justifying the System of Justice in Terms of "Good Police"

Smith uses the example of the sentinel to point out the difficulty with which a natural moral judgement can be made compatible with the severity of institutional justice, suggesting that a "man of humanity must recollect himself, must make an effort, and exert his whole firmness and resolution, before he can bring himself to inflict it [the death penalty], to go along with it when it is inflicted by others." Immediately following this, he adds a counter-example: "It is not, however, in this manner, that he looks upon the just punishment of an ungrateful murderer or parricide. His heart, in this case, applauds with ardour, and even with transport, the just retaliation which seems due to such detestable crimes" (TMS II.ii.3.11, 90-91). Smith's intention in juxtaposing these two cases is easy to appreciate, namely that he would like to show the plurality or variety of the principles on which different judgements of justice are founded, and that he suggests it is rather rare cases that the actions should be judged in terms of their consequential impact on the public interest, instead of in

[7] Hume's argument for the utilitarian basis of justice in his *Second Enquiry* is actually correspondent to what has been investigated here in regard to Smith. In the 3rd section titled "of Justice" he makes good use of several philosophical or hypothetical scenarios in order to demonstrate that few people would observe the laws of justice and equity under these circumstances, which totally deprive justice of its usefulness (this does not, however, necessarily denote that each member of the society is always aware of its utility to the public whenever they act in accordance with justice). See Hume 1998, 13-19. In addition, the notion that the state of emergency or necessity justifies the suspension of the laws of justice in the name of public welfare or safety (*ibid.*, 15-6) is also shared by Smith. Cf. Hont 2005, ch. 5.

[8] Hume 1998, 23-25, 45-46. However, even in this *Second Enquiry*, significantly, this does not necessarily mean that Hume considers each person's act of justice as being morally motivated by the conscious consideration of the public utility. See, for example, *ibid.*, 26-27.

[9] What seems to be misunderstood in Hume by scholars of Smith can be said similarly with regard to Francis Hutcheson as well. The selfish passions and ego-centric human disposition (or "constitution") are paid equally special attention to in Hutcheson, as well as benevolent or beneficent human inclination. See Hutcheson 2005, 6-12; Hutcheson and Turco 2007, I.v.3, 82-84.

relation to their causes. This implies that the major reference for determining justice is still an appropriate balance between actions and their causes or circumstances; and any results of the actions are just a secondary factor to judge their morality.

Smith's portrayal of this discussion has the potential to be considerably misleading, however. What is demonstrated in the previous section is that Smith recognises some situations wherein consideration given to the public utility determines what is right even though it is contradictory to the natural sensibilities of the common people. It does not necessarily follow, however, that on other occasions the principle of utility therefore gives way to the principle of natural justice. This section intends to demonstrate that the principle of utility is equally satisfied during these ordinary situations. Smith virtually admits that almost all moral actions and institutions suitable for the laws of justice — or judged reasonable from the sense of natural reason — are also appropriate when judged from the perspective of public utility. This denotes that, in ordinary situations, not only the natural principle of justice but also the principle of utility are satisfied even though each individual agent is not necessarily conscious of the fact. What is often ignored, from a deontological perspective, but is significant in itself, is the idea, characteristic of the Scottish Enlightenment, that moral actions whose main intention is to fulfil the natural principle of justice *consequently* maximise the public utility at the same time. Therefore, it is not simply a matter of the former principle overriding the latter principle here.

As Knud Haakonssen described in systematic fashion, Smith's moral philosophy includes the so-called "historical jurisprudence", or historical sociology in the modern sense, wherein the development of a legal system is analysed in relation to different stages of manners and modes of subsistence shared in common by each society. According to the "four stages" theory, while savage and barbarous nations dedicate themselves to hunting and herding, the majority of feudal societies are agrarian, with the European civilisation of Smith's era seeing the genuine emergence of commercial societies during the Age of Enlightenment. In accordance with these social and economic stages, their laws and justice systems also change (albeit while the causal relation between the ways of life and the system of justice is not so much unilateral as reciprocal).[10] Following the publication of Haakonssen's landmark work, Smith's historiography of Europe has been generally understood as a history in which the natural laws of justice are presented as having gradually developed based on human social and sympathetic nature, resulting in "the impartial spectator,"[11] while the real

[10] For a simple outline of Smith's historical jurisprudence and four stages theory, see, for example, Lieberman 2006, 227-31.

[11] Haakonssen 1981. Cf. Haakonssen 2003.

positive laws, such as mercantile regulation policies,[12] are criticised according to a developing natural sense and laws of justice. "The laws of police" are, from a perspective of this kind, looked upon as a typical example of the objects of Smith's "legal criticism".

This interpretative framework, pioneered by Haakonssen (and partly by Donald Winch, 1978) is problematic to some degree, however. There are at least two reasons for this. One of the reasons is that Smith does not ignore their public utility at all when deeming more than a few laws of police to be unjustifiable. It is certain that a large number of the irrational laws installed by the feudal governments were, according to him, the laws of police under the pretext of benefitting the public utility or welfare, and that they are criticised by Smith as being contradictory to natural justice shared by the people as impartial spectators (WN IV.viii.17; V.ii.k.64, 75). It should also be noted, however, that these unjustifiable laws are judged to be in opposition to the common cause of improving the public utility as well; they are criticised not just because they appear to the impartial spectator as violating the laws of natural justice, but because they are not sound laws of police either. With respect to the "institution of entails" and the "right of primogeniture" as a symbol of "oppressive" government, as David Lieberman rightly posits: "whereas in the first part of the lectures (on justice), these institutions were condemned as the *unjust* remnants of an earlier and oppressive political order, now [i.e. in Smith's analysis of the causes of the slow progression toward opulence in the next part of the lectures on police] they were condemned as "extremely prejudicial to the *public interest*" on account of their "great hindrance to the progress of agriculture" (LJB, 289-95). In this example, as elsewhere, the "principle of law and government" as applied to "justice" and to "police" offered two complimentary frameworks for the assessment on the same body of positive law" (Lieberman 2006, 238. Emphases added). The reasoning in Smith's argument is thus completely distinct from cases where it is reasoned that the laws of police that are regarded as being useful for public utility can be disapproved of only by the criterion of natural justice.

The second reason why Haakonssen's interpretation is not acceptable without some reservation is that Smith both possesses and emphasises a utilitarian point of view when evaluating laws that are in harmony with natural justice. The dual perspectives characteristic of Smith's "science of a legislator," as shown above, should not be ignored here either, which make it possible to evaluate the same laws from two different angles: the principles of both natural justice and public utility. As seen in Lieberman's argument, the distinction between

[12] For example, Smith describes the prohibition of the British woollen trade as a law of police introduced in the name of necessity or utility (that is, "the wealth and strength of the nation") and that the death penalty for those who only exported wool cannot be sympathised with according to the natural sense of an impartial spectator (or "in naturall equity"), meaning that this law did not function well in practice. Cf. LJA ii.91-2; LJB 182; WN IV.viii.17.

justice and police should not be understood as representing the objective or substantial difference between the *laws* of justice and those of police. Ultimately this distinction should be attributed to the analytical observers who are attempting to morally evaluate legal institutions.[13] As such, the same laws can be assessed according to the two different principles without any contradiction. And in reality, Smith discusses every law that has already been judged as just and appropriate by natural justice is simultaneously beneficial to the public utility in the second part of his jurisprudence (that is, theory of police). Naturally just and proper laws are therefore regarded as being also the best laws of police.

To sum up, it is not strictly accurate to suggest that the civilising process evident in Smith's historical jurisprudence is one wherein the laws of police have been gradually superseded by the laws of justice. More precisely speaking, a great many laws had been altered during this process to become ones that are endorsed by the natural sense of justice, and which maximise public utility at the same time; any truly civilised laws are good and just in terms of policing *as well as* in terms of justice. What should be added to this is a point examined earlier in this paper, namely that in some crucial instances the common people are required to consciously prioritise the public utility over the natural principle of judging justice, in order to uphold the justice system. When we consider that the matter of public utility is a requirement that must be met in both cases, it is scarcely possible to argue that civil police or military discipline are the only cases in which the principle of utility comes to matter. The principle of utility must be satisfied — whether consciously or unconsciously — not only in cases of civil police, but in more ordinary circumstances as well (although this is often ignored on the basis of a Kantian deontological standpoint).[14] In contrast, the principle of natural justice is usually respected when determining morality on a daily basis, while in relatively rare but critical instances this principle is suspended and must yield to the principle of utility.

III Conclusion

For the reasons discussed, Adam Smith's moral theory should be, first of all, interpreted as a successor to "Humean" utilitarianism, which can be seen as somewhat distinct from so-

[13] Cf. "The distinction between "justice" and "expediency" served to distinguish two distinct moral perspectives on law and government. However, it emphatically did not carve out two separate and autonomous regions of social life, each exclusively shaped by a single and different moral virtue" Lieberman 2006, 237.

[14] In reality, Kant himself seems to be fully aware of the aspect emphasised in this paper, when considering not only his ethics as normative theory but his political philosophy and anthropology as well.

called "classical utilitarianism" in the sense that there is more room for situating "spontaneous order" or "unintended consequences". In Hume and Smith's version of utilitarianism, individual actions will unconsciously result in satisfying the principle of public utility in most instances, although these natural judgements should sometimes be modified by and subjugated through direct referral to public utility. This even anticipates the role of the market economy as the "invisible hand" — wherein the social agent is typically motivated by a combination of "private" utility and the natural moral sense shared by a "civilised" common populace — with "regular civil government" and occasional maintenance from legislators functioning as indispensable components, all of which serves to reconcile what is just and proper for individuals with what is good for the public as a whole.

References

[1] Berry, Christopher J. 2013. *The Idea of Commercial Society in the Scottish Enlightenment*. Edinburgh: University Press.

[2] Lieberman, David. 2006. "Adam Smith on Justice, Rights, and Law." in *The Cambridge Companion to Adam Smith*, edited by Knud Haakonssen, 214-45. Cambridge: Cambridge University Press.

[3] Haakonssen, Knud. 1981. *The Science of a Legislator: The Natural Jurisprudence of David Hume & Adam Smith*. Cambridge: Cambridge University Press.

[4] _____. 2003. "Natural jurisprudence and the theory of justice." In *The Cambridge Companion to the Scottish Enlightenment*, edited by Alexander Broadie, 205-21. Cambridge: Cambridge University Press.

[5] Hont, I and M. Ignatieff. 2005. "Needs and justice in the *Wealth of Nations.*" In *Jealousy of Trade: International Competition and the Nation-State in Historical Perspective*, edited by Istvan Hont. Cambridge, MA: The Belknap Press of Harvard University Press.

[6] Hume, David. 1998. *An Enquiry Concerning the Principles of Morals*, edited by Tom L. Beauchamp. Oxford: Oxford Clarendon Press.

[7] Hutcheson, Francis. 2005. *System of Moral Philosophy,* 2 vols, London: Continuum Classic Texts.

[8] Hutcheson, Francis. 2007. *A Short Introduction to Moral Philosophy.* Edited *by* L. Turco*.* Indianapolis: Liberty Press.

[9] Smith, Adam. 1976. “The Theory of Moral Sentiments.” In *The Glasgow Edition of the Works and Correspondences of Adam Smith*, edited by D. D. Raphael and A. L. Macfie. Oxford: Clarendon Press/ Indianapolis: The Liberty Fund. [TMS].

[10] Smith, Adam. 1976. “*An Inquiry into the Nature and Causes of The Wealth of Nations*.” In *The Glasgow Edition of the Works and Correspondences of Adam Smith*, edited by R. H. Campbell, A. S. Skinner and W. B. Todd. Oxford: Clarendon Press/ Indianapolis: The Liberty Fund. [WN].

[11] Smith, Adam. 1978. “Lectures on Jurisprudence.” In *The Glasgow Edition of the Works and Correspondences of Adam Smith,* Report of 1762-3, edited by R. L. Meek, D. D. Raphael, and P. G. Stein. Oxford: Clarendon Press/ Indianapolis: The Liberty Fund. [LJA].

[12] Smith, Adam. 1978. “Lectures on Jurisprudence.” In *The Glasgow Edition of the Works and Correspondences of Adam Smith,* Report dated 1766, edited by R. L. Meek, D. D. Raphael, and P. G. Stein. Oxford: Clarendon Press/ Indianapolis: The Liberty Fund. [LJB].

[13] Winch, Donald. 1978. *Adam Smith’s Politics: An Essay in Historiographic Revision*. Cambridge: Cambridge University Press.

Pigou's Theory on Welfare Economics in the Narrow and Broader Senses: Based upon the Indirect Utilitarian Strategy

Satoshi Yamazaki, Kochi University, Japan

Abstract

It is typically believed that A.C. Pigou's theories on welfare economics is the incarnation of utilitarian moral principle proposed by Bentham and Sidgwick. For instance, Y. Edgeworth once observed that Pigou drew inspiration from Sidgwick concerning wealth and welfare, and that the good which philanthropists and the public sector should seek to realize is defined by Pigou in accordance with Sidgwick's utilitarianism. However, the interpretation of Pigou's ethics has recently become rather controversial, since several recent studies have attempted to correct the typical understanding from their respective points of view.

This presentation presents a new understanding of Pigou's welfare theory as a whole system based on my own examinations of his works, which include points such as: (i) an investigation of Pigou's extensive works (encompassing major and minor documents) and a reconstruction of his basic moral principle, *ideal utilitarianism*, (ii) an examination of the rigorous practical application of his ethics on his welfare economics, (iii) an analysis of his (implicit) need concept and its close relationship to the idea of the national minimum (or "safety net," to use a rather modern term), and (iv) a clarification of his opinion that the national minimum should take priority over any other expedience, comprised chiefly of subjective satisfaction (utility).

Integrating the issues introduced above, this presentation proposes a notion of welfare economics in the *narrow* and the *broader* sense considering Pigou's work. His welfare economics in the broader sense encompasses two major criteria: desire satisfaction and need satisfaction. Moreover, under certain conditions, Pigou admits that the satisfaction of needs takes priority over pleasure and utility, which means that the prescription of the safety net does not rely on utility maximization. How should we explain Pigou's position? This article explores a new exposition of Pigou's moral strategy in the light of the notion of *indirect* utilitarianism.

Introduction

This article presents a reconstruction and new interpretation of A.C. Pigou's welfare theory as a whole based on the accumulated results concerning Pigou that I have made for the past few years. The concept of the presentation can be summarized as an interpretation of Pigou's welfare economics in the *narrow* and the *broader* senses. First, as I have mentioned previously (Yamazaki 2011, 1n), defining his work on welfare economics could be problematic. It seems slightly nearsighted for us to identify his economic thought based only on his major works such as *Wealth and Welfare* (1912 (*WW*)) and *Economics of Welfare* (1932 (*EW*)). Reflecting on his original intention—his notion that economics is merely instrumen-

tal to national well-being and welfare—we should not commit ourselves to such a near-sighted view. We can certainly address welfare economics in the broader sense as well as a narrower sense based on Pigou's works, since he based his entire economic thought on not only his major works, but also minor works including his essays and discourses, which are not necessarily systemized. In this article, we endeavor to explore Pigou's welfare economics in both senses.

I A Brief Review of My Preceding Studies on Pigou's Ethics and Welfare Economics

As a preliminary study to lead into the main argument in the next section, we briefly survey the points that I have previously made concerning Pigou in my previous articles.

First, Pigou's ethics were examined based on his early—but critical—documents, which had rarely been referred to in other studies of his theories (Yamazaki 2002). Until recently, his ethics have been more or less regarded as traditional and as typical hedonistic utilitarianism. For instance, Edgeworth (1913), Hutchison (1953), Schumpeter (1954), Blaug (1978), O'Donnell (1979), and Collard (1981, 1996) placed Pigou in the utilitarianism stream of the Benthamite or Sidgwickian traditions. However, these scholars did not necessarily consider Pigou's thought from a strictly ethical or philosophical point of view. Their arguments seem to merely be supplementary observations incidental to economics. I have shown that—contrasting to conventional understanding—Pigou's utilitarianism differs from that of Bentham, J. S. Mill, or Sidgwick in some crucial respects, and that Pigou should be regarded as an ideal (non-hedonistic) utilitarian, a position held by G.E. Moore and H. Rashdall.

Second, his theories concerning economic and non-economic welfare – the key concepts in his work on welfare economics – were re-examined (Yamazaki 2011, 2012). According to Pigou, economic welfare forms part of overall welfare. This begs the question; what are the contents of his theories on non-economic welfare? It is generally accepted that non-economic welfare is indirectly related to economics through economic welfare. Nevertheless, is there any direct relationship between them? This is the second problem I have also addressed. It is generally assumed that economic welfare is obtained through the desire satisfaction principle. However, I have shown that the contrasting *need* satisfaction principle exists in Pigou's thinking. Additionally, crucial aspects of his welfare economics theory depend on this second principle. From a theoretical perspective, those parts of welfare that are accomplished through the satisfaction of needs do not completely coincide with economic welfare (utility or subjective satisfactions). Contrary to conventional understanding,

I argued that certain aspects of non-economic welfare are intended to be promoted directly in Pigou's welfare economics (Yamazaki 2011).

Third, I observed that Pigou's concept of needs is closely related to his concept of the national minimum—or safety net—in welfare economics (Yamazaki 2011, 2012). Although this is not clear from studying his main works (e.g. *WW* and *EW*) alone, a broader investigation indicates that his argument concerning the national minimum is not so much based on desire satisfaction, but on need satisfaction, by which he intended to promote people's non-economic welfare (e.g. character and ethical personality). Overall, we can interpret Pigou's concept of the national minimum to be founded on the satisfaction of objective needs, rather than on individuals' subjective satisfaction represented by economic welfare. This constitutes quite a different understanding of Pigou's work from conventional thinking.

Lastly, some practical issues were addressed, for instance the priority of enforcing the safety net for the poor in social policy (Yamazaki 2012). Taking a rough view, welfare economics based on utilitarianism does not secure the safety of the weak or of social minorities, but rather, a sacrifice on their part for an increment in social welfare may be acceptable, since its ultimate criterion is mere summation of the total utility. Surely this is a typical denunciation, but I have already stated that this pattern does not hold true for Pigou's work in terms of the interpretation of the *indirec*t strategy of utilitarianism. Since this last point is of extreme importance, let us briefly revisit my previous exposition (Yamazaki 2012) below.

Figure 1 presents Pigou's distribution criterion as described by Collard (1981, 111-12) based on Pigou's double propositions (production and distribution in *EW*).[1] The society consists of the poor and the rich only. The vertical and the horizontal axes represent the income level of the poor and the rich respectively. The intersections of these axes and the oblique line indicate the amount of social income, and each point on the oblique line represents a corresponding pattern of social income distribution for both sides. The intersection of the oblique line and the 45-degree line from the origin of the coordinate axes indicates equal distribution. We can now set an arbitral point K0 (a typical economic disparity) on the oblique line (distribution pattern) and consider a socially better set than K0 based on Pigou's prescriptive criteria. Pigou's specific distribution standard lies in his second proposition in *EW*: other factors (especially the amount of national income) being unchanged, a more equal distribution of wealth is conducive to the aggregation of social economic welfare (social betterment). Notably, interpersonal comparison and the law of diminishing marginal

[1] Figure 1 is quoted from Collard (1981, 112), while Figure 1-a and 1-b are my own modifications of the original Figure 1. Although I do not address these here, Collard also represents Pareto's and Rawls's distribution criteria using similar figures.

utility underlie this statement. If other factors change, a so-called "disharmony" (i.e. the more equal distribution hinders the growth of national income) may occur (Pigou 1932, 645), which creates uncertainty concerning whether we can obtain resultant social betterment.

The shaded areas including the edges—except K0—in Figure 1 shows Pigou-superior set to the initial K0 (Collard 1981, 111-12). The "?" inside the triangle corresponds to the foregoing "disharmony," which does not necessarily guarantee social improvement due to the trade-off between the reduction of national income and the enhancement the poor's welfare. As I have already pointed out, the Pigou-superior distribution in Figure 1, as drawn by Collard, is based upon the criterion of economic welfare (subjective desire satisfaction) alone. Neither objective needs nor the national minimum are considered. Contrastingly, noticing Pigou's contentions that "a *national minimum* is understood as an objective minimum of conditions in all departments of life, below which the fortunes of no citizen are allowed to fall" (Pigou 1912, xxvi. Italics original), that "It is the duty of a civilized state to lay down certain minimum conditions in every department of life, below which it refuses to allow any of its free citizens to fall" (Pigou 1914, 36), and that "It is generally agreed in modern communities that some minimum standard of life must be established … below which no citizen or family shall be allowed to fall" (Pigou 1952, 203), I introduced such an income level, which just satisfies the objective needs or the minimum standard represented in the diagram as the dotted horizontal line. As can be seen, there are two cases to be examined: in one case, the dotted line is under K0 (Figure 1-a), while in the other, it is above K0 (Figure 1-b). Concerning the former case, the initial K0 standard—even though it is applicable to the poor—is beyond the minimum requirement. Therefore, the welfare situation of the poor does not seem to be an object of social mandatory relief. Therefore, the result of Figure 1-a is identical to the original Figure 1. However, Figure 1-b, in which the initial K0 is beneath the dotted line, is completely different. As stated above, Pigou claimed that the level under the dotted line corresponds to such a condition that no one is allowed to fall into. Therefore, if the need criterion in Pigou's work is adequately considered, Pigou-superior distribution criteria can be reasonably modified from the original in Figure 1 and expanded to accommodate social justice more than naïve utilitarianism prescribes. However, Figure 1-b implies that we are ready to admit to a possible reduction in national income to benefit the unfortunate who fall under the minimum. Considering naïve utilitarian calculation, such a case cannot be approved, as it may entail a consequential shrinkage in the social summation of utility. Pigou, however, eloquently asserts:

> After all … so long as progress in technique continues, production may be expected to expand in any event. The offset to increased fairness in distribution is thus likely to be, not a

catastrophe for production, but at the worst, some slowing down in its rate of increase; and that is not a disaster. (Pigou 1955, 87)

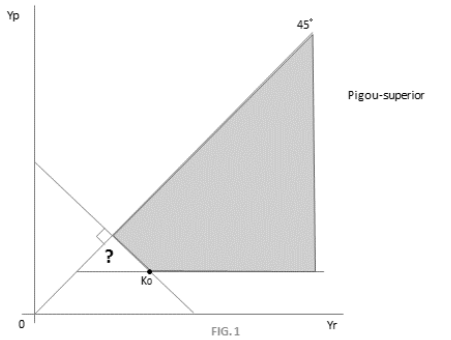


FIG. 1

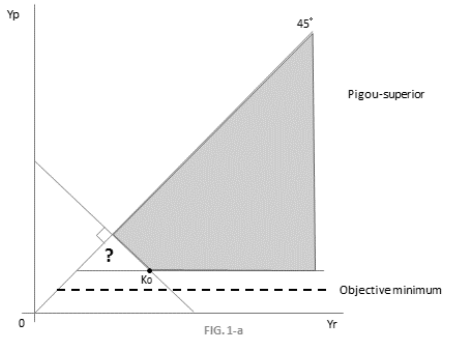


FIG. 1-a

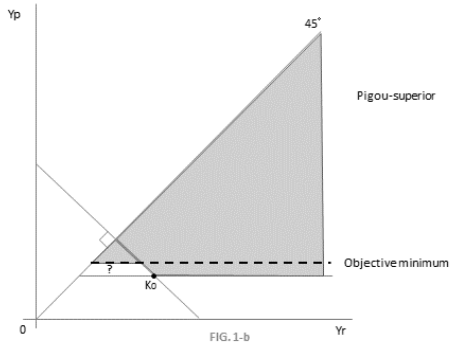


FIG. 1-b

In summary, the above indicates that in a case where the satisfaction of basic needs and other interests (or preferences) compete with one another, the former must be socially prioritized. This is called Pigou's "basic rule" in welfare economics. However, how can this basic rule be demonstrated, and how it is compatible (or incompatible) with the utilitarian principle? Presumably, as I have indicated, Pigou's argument can be placed under the heading

of *indirect* utilitarian prescription.[2] Essentially, Pigou does not acknowledge that every good must be accompanied by obligation (Pigou 1965, 13). Obligations are not directly linked to the amount of welfare (intrinsic good). Even according to a certain interpretation of utilitarianism, moral obligations are not necessarily and directly prescribed by reference to the simple summation of utility. For instance, Kelly stated the following:

> While it is correct to argue that all authoritative reasons for action must be reducible to act-utilitarian ones, this does not imply that either individual or legislator is under a direct obligation to pursue the maximum social well-being in all circumstances. (Kelly 1990, 254)

Moral obligation directly concerns "vital interests" (such as security) for well-being. Such an item "is distinguished on the grounds that it is a necessary condition of the formation of any conception of well-being" (Kelly 1989, 75). Naturally, in Pigou's work, vital interests correspond to needs and the national minimum. Arguably, that "necessary condition" indicates that it is irrelevant to the amount of the intrinsic value. Even if the value of the necessary condition is trivial (compared to, for instance, other sublime virtues), neither preference satisfaction nor a virtuous life can be realized without the condition. Therefore, it must be prioritized for everyone, regardless of its degree of value. While the "rightness" of actions is determined by the consequential outcome (welfare or utility), the obligatoriness is not. Since the minimum (need satisfaction) is defined as a moral obligation, it needs to be prioritized before other considerations that merely maximize total utility (expedience). As an example, J.O. Urmson once observed:

> … though there can be very tricky problems of duty, they do not naturally present themselves as problems whose solution depends upon an exact determination of an ultimate end; while the moral principles that come most readily to mind—truth-telling; promise-keeping; abstinence from murder … and the like make a nice discrimination of the supreme good seem irrelevant. We do not need to debate whether it is Moore's string of intrinsic goods or Mill's happiness that is achieved by conformity to such principles; it is enough to see that without them social life would be impossible and any life would be indeed solitary, poor, nasty, brutish, and short … Such considerations … have led some utilitarians to treat avoidance of the *summum malum* rather than the achievement of the *summum bonum* as the foundation of morality. (Urmson 1958, 208-209. Italics original)

Therefore, Urmson has an affinity for utilitarianism (it is more suitable to accommodate his claims than any other moral principle). Pigou recognized that the policy of the minimum

[2] This observation is an extension of the corresponding parts of my work (Yamazaki 2012).

standard can eventually be justified by considerations of "the compelling obligation of humanity" (Pigou 1914, 37) which is likewise irrelevant to direct maximization of welfare.[3]

II The Structure of Pigou's Welfare Economics in the Narrow and the Broader Senses

Next, we are approaching the culmination of this article.

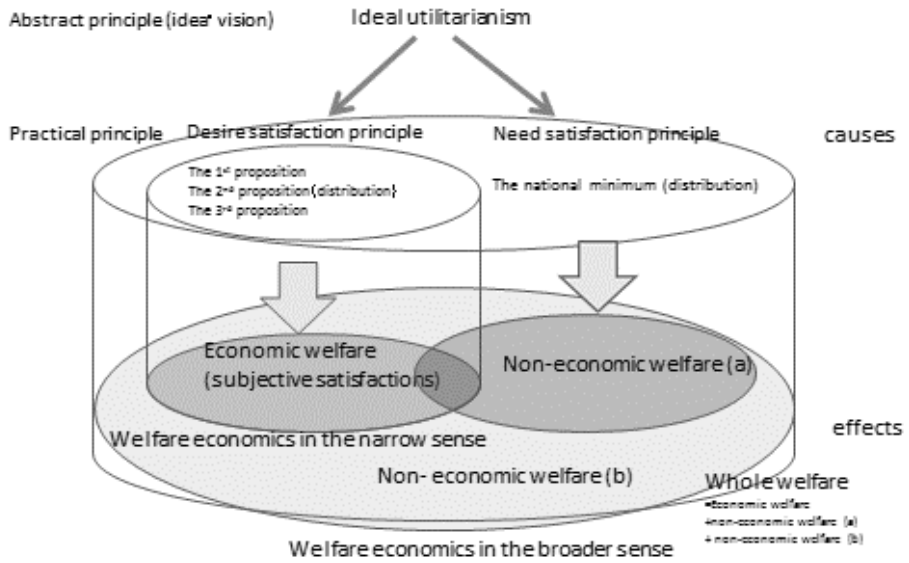


Figure 2

Figure 2 shows a conceptual framework of the whole system of Pigou's welfare economics based on my own interpretations and findings concerning his works. We now review each point in the figure while referring to the observations provided above.

[3] Although Urmson holds the affinity with utilitarianism, his reasoning of justification of obligation appears to somewhat differ from that of Kelly's. For while Kelly tries to found the validity of an absolute and prioritized obligation within utilitarian axiology (welfarism) at the level of normative ethics, Urmson's reasoning of foundation seems to go beyond normative ethics and extend to meta-ethics. In that sense, the foundation like Kelly's may be called an intrinsic justification to utilitarianism, whereas Urmson's, an extrinsic one.

The vision refers to his basic moral principle, which we argue is non-hedonistic with multiple ingredients of welfare, called "refined hedonism," using Baldwin's term (Baldwin 1990, 132).

As Pigou stated, the purpose of welfare economics is to create measures to promote welfare in an easier way, and there are clearly at least two major practical principles in his system: *desires* and *needs*.

However, almost all existing studies concerning his welfare economics theories have thus far only featured the former aspect. We call this welfare economics in the *narrow* sense, the criterion of which is exclusively subjective utility (economic satisfaction).

On the other hand, Pigou's welfare economics can be conceived in the *broader* sense by exploring his implicit but firm concept of needs satisfaction. This is one of the most crucial points of this presentation.

Of course, the satisfaction of both needs and desires are mere practical measures, and are therefore not intrinsically but instrumentally crucial. Pigou's final aim is the consequences of those prescriptions. Desires are supposed to achieve economic satisfaction, while needs are intended to lead to non-economic welfare, including certain virtuous elements. However, of all basic human needs, the primary needs in particular refer directly to existence or security or health, which are necessary conditions for a formation of welfare. From that perspective, the satisfaction of the primary needs should be considered as an indirect prescription to the promotion of welfare. On the other hand, of course, Pigou intended to realize certain welfare elements directly according to the need principle (say, moral character or quality of people through education and training). Moreover, as has been clearly illustrated, Pigou actually prioritizes needs satisfaction over other considerations.

A remaining problem is how to incorporate this need principle into his utilitarian framework. For this task, we suggest referring to the indirect strategy of utilitarianism. By referring to Kelly and Urmson, we focused on the notion of necessary conditions for welfare. That essentially means no security, no happy life. These necessary conditions that Pigou regards as primary needs may be—in themselves—less valuable than other supreme factors (either virtuous or aesthetic factors). However, when a concept is obligatory, it does not necessarily proportionally correspond to its holding value. It is clear that Pigou has considered this.

To conclude, everyone's primary needs ought to be socially prioritized over any other considerations, and the justification for this is not subject to simple direct welfare maximization. Pigou has introduced a lexical order between obligation and non-obligation, and the

former corresponds to basic needs. We may call this his indirect strategy of utilitarianism, of which his broader welfare economics is constituted.

Eventually, according to Pigou's notion, if all citizens are well-educated, we do not need to consider welfare economics in the broader sense. For educated people would arrange economic decision so that their subjective satisfaction will be concurrent with their needs satisfaction (there are little gulf between desires and needs (or, economic virtue and human virtue) in Pigou's (1906, 1907) terms). However, as Pigou (1906, 379-80) has asserted, even if non-educated people generally could be the best judge of what they do want, they might not be the best judge of what they *ought* to want. In order to reduce the gap between subjective desires and objective needs, Pigou implicitly recourses on needs principle and prescribes those priority, which has led us to conceive his welfare economics in the broader sense.

Acknowledgement

This study is supported by the research grant JSPS (KAKENHI: 15K03383).

References

[1] Baldwin, T. 1990. *G. E. Moore*. London: Routledge.

[2] Blaug, M. 1978. *Economic Theory in Retrospect.* 3rd edition. Cambridge: Cambridge University Press.

[3] Collard, D. 1981. "Pigou." In *Pioneers of Modern Economics in Britain*, edited by D. P. O'Brien, and John R. Presley, 105-139. London: Macmillan.

[4] _____. 1996. "Pigou and Future Generations." *Cambridge Journal of Economics* 20: 585-97.

[5] Edgeworth, F. Y. 1913. "*Wealth and Welfare* by A.C. Pigou." *Economic Journal* 23 (89): 62-70.

[6] Hutchison, T. W. 1953. *A Review of Economic Doctrines 1870-1929.* Oxford: Clarendon Press.

[7] Kelly, J. 1989. "Utilitarianism and Distributive Justice: The Civil Law and the Foundations of Bentham's Economic Thought." *Utilitas* 1: 62-81.

[8] _____. 1990. "Utilitarian Strategies in Bentham and John Stuart Mill." *Utilitas* 2: 245-66.

[9] O'Donnell, M. G. 1979. "Pigou: An Extension of Sidgwickian Thought." *History of Political Economy* 11: 588-605.

[10] Pigou, A. C. 1901. *Robert Browning as a Religious Teacher*. London: C. J. Clay and Sons.

[11] _____. 1906. "The Unity of Political and Economic Science." *Economic Journal* 16 (63): 372-80.

[12] _____. 1907. "Memorandum on Some Economic Aspects and Effects of Poor Law Relief" *Appendix vol. 9. 1910, Minutes of Evidence, Royal Commission on the Poor Laws and Relief of Distress*, Cd.5068, 981-1000. London: His Majesty of Stationary Office and Wyman and Sons. Ltd.

[13] _____. 1908. *The Problem of Theism, and Other Essays*. London: Macmillan.

[14] _____. 1912. *Wealth and Welfare*. London: Macmillan.

[15] _____. 1914. "Some Aspects of the Housing Problems." In *Lectures on Housing*, written by B. Rowntree, and A. C. Pigou. Manchester: Manchester University Press.

[16] _____. 1932. *Economics of Welfare*. 4th edition. London: Macmillan.

[17] _____. 1952. *Essays in Economics*. 2nd edition. London: Macmillan.

[18] _____. 1955. *Income Revisited: Being a Sequel to Income*. London: Macmillan.

[19] _____. 1965. *Essays in Applied Economics*. 2nd edition. London: Frank Cass.

[20] Schumpeter, J. A. 1954. *History of Economic Analysis.* New York: Oxford University Press.

[21] Urmson, J. O. 1958. "Saints and Heroes." In *Essays in Moral Philosophy*, edited by A. I. Melden, 198-216. Washington: University of Washington Press.

[22] Yamazaki, S. 2002. "Pigou as the Ideal Utilitarian." *Annals of the Society for the History of Economic Thought* 41: 35-47 [in Japanese].

[23] _____. 2011. *Pigou no Rinri Shisou to Kousei Keizaigaku: Fukushi, Seigi, Yuseigaku [Pigou's Ethical Thought and Welfare Economics: Well-being, Justice, and Eugenics]*. [In Japanese]. Kyoto: Showado.

[24] _____. 2012. "Need and Distribution in Pigou's Economic Thinking." *Paper presented at 12th Conference of International Society for Utilitarian Studies*, New York, August 9-11, 2012. http://www.stern.nyu.edu/experience-stern/about/departments-centers-initiatives/academic-departments/business-society-program/isus-2012-conference/papers/index.htm (accessed June 27, 2013).

Two Ways to Satisfy (and No Way to Satisfy Utilitarians)

Alexandra Zinke, Karlsruhe Institute of Technology, Germany

Abstract

Preference utilitarianism holds that an action is morally good iff it maximizes overall preference satisfaction. In principle, there are two ways to satisfy preferences: either you alter the facts such that they fit the subject's preferences, or you change the subject's preferences such that they fit the facts. While standard preference utilitarianism focuses on the first strategy, the present paper will explore the prospects and limits of the second strategy. I will firstly argue that there are cases in which it seems morally right to aim at preference satisfaction by preference change, but secondly acknowledge that an action that induces a global change of preferences doesn't necessarily seem morally right. The real philosophical challenge is to distinguish those cases where altering a subject's preferences is morally right from those where it isn't. The paper ends with a skeptical outlook on the possibility of justifying the distinction on purely preference-utilitarian grounds.

Introduction

Rationality demands the maximization of one's own welfare. According to utilitarianism, morality demands the maximization of overall welfare. *Preference* utilitarianism subscribes to a desire-fulfillment theory of welfare (also known as *preferentism*): a subject's welfare increases with the fulfillment of her desires / the satisfaction of her preferences. Thus, the core idea of preference utilitarianism can be stated as follows: an action is morally good iff it maximizes overall preference satisfaction. Or, in its prescriptive reformulation: One should choose an action that maximizes overall welfare.[1] Preference utilitarianism will be presupposed throughout this paper.

Let us begin by examining the central notion of preference utilitarianism, *satisfaction (or fulfillment) of preferences*. We will say that a preference is satisfied iff the content of the preference is realized: *S*'s preference that *p* is satisfied iff *p*. The subject need not know about the satisfaction of the preference or experience any feelings of fulfillment. The notion of satisfaction is of course not restricted to the satisfaction of preferences but also applies

[1] Here and in what follows, I use "should", "right", etc. in their moral, not their prudential reading.

to all other pro-attitudes, e.g., to wants, desires, wishes, etc.[2] A pro-attitude is satisfied iff it its matched by the world. Preference utilitarianism thus says that actions should establish a match between the content of a pro-attitude and the world. How can they do so? *Prima facie*, there are two ways to establish this fit: one could change the world such that it fits the pro-attitudes, or one could change the attitudes such that they fit the world.[3]

Usually, preference utilitarianism is read in the manner of world-to-preference direction of fit: we should change the (objective, not preference-related) facts such that the world fits the *actual* preferences of the subjects. If Ann prefers the apple to the banana, we should offer her the apple rather than make her prefer the banana. And, to take a somewhat more serious example, if Ben is starving to death, we should give him food rather than make him want to die. Let me call this reading of the initial utilitarian thesis *world-directed utilitarianism*.[4] More precisely, according to *radical world-directed utilitarianism*, an action is morally good iff it maximizes the overall satisfaction of the *given*, i.e., *actual*, preferences. We call this theory *radical* world-directed utilitarianism as it exclusively values preference satisfaction by changes of the objective facts.

It is important to stress, however, that preference utilitarianism as initially stated is neutral with respect to the two strategies of preference satisfaction: nothing but the overall amount of preference satisfaction counts. Utilitarianism itself is silent about *how* the fulfillment of preferences should be achieved. As John Rawls says, if preference satisfaction is all that matters, then we must be "ready to consider any new convictions and aims, and even to abandon attachments and loyalties, when doing this promises a life of greater overall satisfaction" (Rawls 1982, 181). If all that matters is the amount of preference satisfaction, we could make Ann prefer the banana she already has, instead of supplying her with the factually desired apple. And, instead of giving him food, we can at least try selling to Ben the relief found in finally experiencing the eternal tranquility that only death can yield. Let *radical preference-directed utilitarianism* be the thesis that an action is morally good iff it

[2] Talk of satisfied preferences seems a bit awkward, as preferring appears to be a three-place relation between a subject and two objects: *S* prefers *a* to *b.* However, in this paper I will be a bit sloppy and sometimes use "prefer" as a binary relation ("*S* prefers that *p*") and sometimes as a three-place, comparative relation ("*S* prefers *a* to *b*"). Furthermore, I will use "*S* desires/wishes/wants that *p*" interchangeably with "*S* prefers that *p*". Nothing of significance will hinge on this.

[3] Of course there is also a third way: one could combine the two strategies and change both. However, I will here concentrate on the two more conservative strategies of manipulating only one side.

[4] As mental attitudes in general, and preferences in particular, are also ‚parts of the world', this label is not quite accurate. It is intended to stress the contrast between changing the preferences (i.e. mental entities) themselves and changing the facts at which the preferences are directed (which are often, though not necessarily, non-mental facts).

maximizes overall satisfaction of preferences by changing the preferences such that they fit the actual facts. Again, the theory is *radical* as it exclusively values the generation of preference satisfaction by change of preferences, not by change of objective facts.

Radical preference-directed utilitarianism seems to be a nonstarter. If there is a way to satisfy a given preference (without violating any other preferences), then that action seems morally good – at least from the assumed utilitarian perspective. I will not attempt to defend radical preference-directed utilitarianism. But we can think of a weaker form of preference-directed utilitarianism, *liberal preference utilitarianism*, which allows for both ways of maximizing preference satisfaction: it says that an action is morally good iff it maximizes overall preference satisfaction – independently of whether this is attained by changing facts or preferences. Liberal preference utilitarianism will be the view defended here.

The first section will argue by way of example that radical world-directed utilitarianism is wrong: there are at least some cases in which it seems morally good to change an agent's preferences to ones that are satisfied by the world as it is. The second section addresses some potential problems for preference satisfaction by preference change. It defends liberal preference utilitarianism, but also argues that the theory must be supplemented by a principle that distinguishes cases in which preference satisfaction by preference change is a legitimate option from those where it isn't. The paper ends with the skeptical worry that a distinction of these cases cannot be motivated by purely preference-utilitarian means. Preference utilitarianism thus provides at best an incomplete theory of morally good actions.

I A Case for Preference-Directed Utilitarianism

I will present two types of cases in which it seems intuitively morally good to establish preference satisfaction by preference change. A note of clarification: There is a huge debate about whether the satisfaction of all preferences or only of the intrinsic ones counts, and about whether actual or ideal preferences are the target. I bracket this discussion as I think that my cases apply also to versions of preference utilitarianism that concentrate on intrinsic and ideal preferences: we should sometimes even change ideal intrinsic preferences.[5]

[5] We lack a precise account of ideal preferences, but the following characterization by Arneson might be helpful: "My ideally considered preferences are those I would have if I were to engage in thoroughgoing deliberation about my preferences with full pertinent information, in a calm mood, while thinking clearly and making no reasoning errors." (Arneson 1989, 83)

I.a Unrealistic Preferences

Ann, your beloved teenage daughter, deeply desires to become the next big pop star. Unfortunately her voice is terrible rather than terrific. Whenever you listen to her, you become more convinced that her dream will forever remain unfulfilled. What should you do? It is practically impossible for you to change the worldly facts such that your daughter's preferences have a chance of becoming satisfied. No singing lessons will help. The only possible way to make her have fulfilled preferences is by changing them. If you are still aiming at maximizing preference satisfaction, you should try to alter her preferences.[6] You could show her different aims in life, foster her interest in painting or sports so that she will forget about the pop star business, or maybe you should introduce her to punk music.

If it is practically impossible to satisfy a given preference, we should try to reach preference satisfaction by changing preferences. Unrealistic preferences provide the first sample case in which is seems right to change a subject's preferences.[7]

I.b Conflicting Preferences

Ann has grown up and is now planning her honeymoon. Her true love Ben wants to go to the sea, while she prefers the mountains. Money is sparse, so they cannot do both; love is intense, so they definitely want to go together. They consult you about what to do. What should you do?

Given the circumstances, it is metaphysically impossible to satisfy both Ann's and Ben's preferences. If you are striving for a maximization of preference satisfaction, you should try to change Ann's or Ben's preferences (or both). This will probably be no easy task, but it seems to be the way to go. Only once Ann's and Ben's preferences are in harmony will it be possible to satisfy those of both of them. Conflicting preferences, i.e., preferences that cannot be satisfied simultaneously, provide my second sample case in which it seems morally right to change a subject's preferences.[8]

[6] For reasons of simplicity, we here ignore the preferences of all other moral subjects, e.g. your possible preferences about Ann's preferences.

[7] What should Ann herself do in the above situation? If we follow the above line of reasoning, she should adapt her preferences. See also Bruckner 2009 for a defense of this intuition with respect to a similar case.

[8] Typical cases involving "ill preferences" or "perverse desires" can also be described as cases of conflicting preferences: if Cen desires to torture the cat, Cen's and the cat's preferences are in conflict.

I think that the two presented cases support the view that it is sometimes morally right – at least from a utilitarian perspective – to strive for preference satisfaction by preference change. If that is correct, radical world-directed preference utilitarianism is wrong. As always, however, moral intuitions might diverge. Some readers might have different views on some or all of these cases. Let me observer however, that a defense of radical world-directed utilitarianism requires some justification for the primacy of actual or given preferences over not-yet-actual ones. The core principle that preference satisfaction is of (moral) value has an immediate intuitive appeal that the more sophisticated principle, which exclusively focuses on, and holds fixed, given preferences, lacks. From a purely preference-utilitarian perspective, what should be wrong with adapting preferences to the world – at least sometimes?

In the next section, I will discuss two possible objections to preference satisfaction by preference change. I will reject the first objection, but acknowledge that the second objection points to the limits of preference satisfaction by preference change. We end up with a modest form of liberal preference utilitarianism.

II Objections to Preference-Directed Utilitarianism

The first objection to preference-directed utilitarianism employs the notion of higher-order preferences, i.e., preferences about one's own preferences. We can think of preferences as ordered in a (possibly infinite) hierarchy. The preferences of order 1 are directed at the world. Preferences of order 1 are, e.g., the preference for an apple, the preference for becoming the next big pop star, or the preference to spend time in the mountains. But the subject will possibly also have preferences that have preferences of order 1 as their contents. For example, the subject might have the second-order preference that her first-order preferences will soon be satisfied, or the second-order preference that no one changes her first-order preference to go to the mountains. Then again, there can be preferences of order 2, etc., *ad infinitum*. In general, preferences of order $n + 1$ will concern preferences of at most order n. (Real agents will often not explicitly entertain many higher-order preferences. However, first, we can also allow for implicit preferences; second, we here consider somewhat idealized agents; and third, and most importantly, we aim at making the *prima facie* objection to preference-directed utilitarianism as strong as possible.)

Objection (higher-order preferences): It is plausible to assume that at least some agents have higher-order preferences that (at least some of) their lower-order preferences are not to be interfered with. Ann wants to become the next big pop star and wants nobody to

change that preference of hers. And I have a very strong preference that no neuroscientist changes my preference not to commit suicide today. Thus we can usually not improve overall preference satisfaction by changing preferences of a lower order as this will violate strong higher-order preferences.

Reply: This objection to preference-directed utilitarianism applies only to an impoverished version of the theory. Of course, if an agent has relevant higher-order preferences, e.g., the second-order preference *B* not to change the first-order preference *A*, then one should not interfere with *A* in isolation but change *B* first. We must always begin by changing the relevant preferences of the highest order. Thus, before changing Ann's first-order preference to become a pop star, we must change her second-order preference that no one interfere with her first-order pop star preference. (If the hierarchy of preferences is infinite and there is no highest preference to begin with we should change all relevant preferences simultaneously.)[9,10]

Preference satisfaction by preference change, understood correctly, does not violate any higher-order preferences: they are not violated, because they are changed first. However, let me now develop another, more fundamental, objection to preference-directed utilitarianism. It shows that in many cases, realizing preference satisfaction by preference change seems intuitively morally wrong (or at least not morally right).

There is a trivial two-step way to maximize preference satisfaction by preference change: we first delete all unsatisfied preferences – this eliminates any mismatch between preferences and facts – and then generate maximally strong preferences such that they are satisfied by the world as it is, thereby maximizing the overall amount of preference satisfaction. Thus we make the agents maximally desire whatever is actually the case – and only this. If the number of fish in the Amazonas is even, then we should make Ann strongly desire this; if the last dinosaur died on a Tuesday, we should make Ann have a strong preference for

[9] Of course, from a practical perspective this is quite demanding. However, overdemandingness objections seem irrelevant as long as we are discussing only evaluative, not prescriptive, moral principles.

[10] Let me point out a remaining worry: I have assumed that our preferences are ordered in a hierarchy. This excludes the possibility of self-referential preferences like the preference that this very preference not be interfered with, or the very general preference that there be no interference with any preference – including this one. Such self-referential preferences cannot be located at any level in the hierarchy. If the conception of self-referential preferences makes sense and an agent has the preference that there be no interference with this very preference, one cannot change it without violating it. Nevertheless, the negative impact of violating this one preference might be countervailed by the satisfaction of all other preferences, so that even self-referential preferences do not necessarily block maximizing the total amount of preference satisfaction by preference change.

this fact, and so on. I suspect that this seems counterintuitive to many proponents of preference utilitarianism. Let me trigger intuitions a bit with the help of a variant of a well-known thought experiment.

Objection (the preference adjustment machine): Let there be a machine, call it the *preference adjustment machine*, that changes all preferences of an agent such that they fit the facts: it deletes all unsatisfied preferences and creates all preferences that fit the facts. Once an agent is plugged into the machine, she prefers maximally whatever is the case. If everybody is plugged to the machine, the machine creates a brave new world with beings who all maximally desire the same: the world as it is. The result is a maximum total amount of preference satisfaction.[11] Thus, according to liberal preference-directed utilitarianism, we should all plug or be plugged into the machine. Even more: you are morally obliged to plug to the machine, not only yourself, but anybody. Again, this consequence might seem counterintuitive to many. At least, it often does so to me.[12, 13] (If this consequence doesn't seem devastating to you, that's fine! You seem to be a proponent of liberal preference-directed utilitarianism, and the rest of the paper will be of no interest to you.)

Reply: I do not want to reject this objection to liberal preference utilitarianism, but rather wish to make precise what exactly it shows. It does not show that we *never* value preference satisfaction by preference change, but it suggests that we *sometimes*, or typically, tend to value preference satisfaction by a change of worldly facts higher than by preference change. If this is correct, then liberal preference-utilitarianism must be supplemented by a principle that distinguishes between cases in which we can maximize preference satisfaction by preference change from cases in which we shouldn't do so. We need a choice principle, or *Preference Principle*, telling us in which cases preference satisfaction by a change of worldly facts is to be preferred over preference satisfaction by preference change. Without such a Preference Principle, liberal preference utilitarianism provides only an incomplete theory of morally good actions.

[11] Of course the total amount of preference satisfaction grows further if the machine additionally creates new bearers of preferences (i.e., new subjects), but let us here concentrate on maximizing the amount of preference satisfaction for already existing beings.

[12] Note that this intuition doesn't fade even given a more restrictive notion of preference change that only allows for changing the content of already existing preferences and does not allow the creation of new preferences.

[13] For a similar, though less radical though experiment, see Parfit 1984, 496: "I am about to make your life go better. I shall inject you with an addictive drug. From now on, you will wake each morning with an extremely strong desire to have another injection of this drug. […] This is no cause for concern, since I shall give you ample supplies of this drug. Every morning, you will be able at once to fulfil this desire."

We cannot here discuss different proposals for a Preference Principle. Yet let me exemplarily introduce one, if only for the purpose of illustration. The following Preference Principle suggests itself:

Preference Principle: Satisfy the actual (intrinsic and ideal) preferences first. If impossible (e.g., because of highly unrealistic or contradicting preferences), change the preferences.

The suggested Preference Principle gives priority to the satisfaction of given preferences, but suggests changing the preferences if this is the only plausible possibility leading to preference satisfaction. It thereby captures both the intuition that there are cases in which we can, or should, attain preference satisfaction by preference change, and the intuition that we need not plug ourselves or others into the preference adjustment machine. Thus there seems to be an easy way to complete liberal preference utilitarianism with a suitable Preference Principle.

However, let me stress that the above Preference Principle (and, I fear, most variants of it)[14] does not seem to allow for a justification within preference utilitarianism. From the perspective of preference utilitarianism, there is no reason why we should opt for the satisfaction of actual preferences first. Preference utilitarianism, as stated above, exclusively aims at maximizing welfare, where maximizing welfare is understood as maximizing preference satisfaction. The source of a rationale for the Preference Principle, however, seems to rely on considerations surrounding the "autonomy of the subject" the "subject's identity", "oneself being the author of one's preferences", or something along these lines. In a purely preference-utilitarian worldview, there is no place – or at least no natural place – for valuing autonomy or the like; the only moral good is proclaimed to be a maximization of satisfied preferences.

III Conclusion

There are two ways to obtain a satisfied preference: by changing the world such that it fits the actual preferences, or by changing the preferences such that they fit the worldly facts. The common reading of preference utilitarianism focuses on changing the world and leaving the preferences intact. I have defended liberal preference utilitarianism, which allows for

[14] For instance, one could suggest a principle employing the distinction between deliberate and unconscious preference adaption (see, e.g., Elster 1983 and Bovens 1992), or propose a principle referring to Bruckner's notion of "reflectively endorsed" preference change (Bruckner 2009).

both ways of preference satisfaction. More precisely, I have suggested that there are cases in which preference satisfaction by preference change seems morally right, but also stressed that a universal adjustment of preferences to reality doesn't seem morally right. If you share these intuitions but still want to stick to preference utilitarianism, you must supplement your theory with a Preference Principle that distinguishes these situations. I doubt that such a principle can be justified within a purely preference-utilitarian framework.

Acknowledgement

I am grateful to Christopher v. Bülow, Wolfgang Freitag, Peter Königs and Christian Seidel for very helpful comments on an earlier version of this paper.

References

[1] Arneson, Richard J. 1989. "Equality and Equal Opportunity for Welfare." *Philosophical Studies* 56: 77-93.

[2] Bovens, Luc. 1992. "Sour Grapes and Character Planning." *The Journal of Philosophy* 89: 57–78.

[3] Bruckner, Donald W. 2009. "In Defense of Adaptive Preferences." *Philosophical Studies* 142: 307-24.

[4] Elster, Jon. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. New York: Cambridge University Press.

[5] Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

[6] Rawls, John. 1982. "Social Unity and Primary Goods." In *Utilitarianism and Beyond*, edited by Amartya Sen and Bernard Williams, 159-85. Cambridge: Cambridge University Press.

Hare's Utilitarianism, Varner's Animals

A Panel Discussion of Gary Varner: *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism* (Oxford University Press, 2012)

Abstract

This panel discussion of Gary Varner's book, *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism* (2012), consists of commentaries by four critics: Gary Comstock, Susana Monsó, Alastair Norcross, and Adam Shriver; followed by a response from the author. Varner's book is a detailed application of R.M. Hare's two-level version of utilitarianism to thinking about issues in animal ethics. Although Hare was Peter Singer's dissertation advisor, Hare never published a systematic discussion of what his theory implied about animal ethics. In the book, Varner examines how far Harean, two-level utilitarianism supports conclusions that Singer has argued for. A central theme of the panel discussion is how the cognitive capacities of individuals affect the relative "moral significance" of their experiences and their lives.

In this panel, the comments of three critics were followed by a response by the author. After a brief overview of the book's contents, the comments of the critics are reproduced below in the order that they were presented: Alastair Norcross, Adam Shriver, Susana Monsó, and panel organizer Gary Comstock. The references from the commentaries and the author's response are gathered together at the end.

Overview of the Book

Gary Varner, University of Texas, USA

The inspiration for my 2012 book *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism* was a graduate seminar that I taught many years ago on the work of Peter Singer. In that seminar, we began by reading R.M. Hare's *Moral Thinking: Its Levels, Method and Point* (1982), and during the semester I repeatedly asked us to consider carefully to what extent Hare's two-level utilitarianism would have the implications for animal ethics that Singer argues for (since Hare himself never systematically explored that question).

There are three parts to the book. The first gives an overview of Hare's version of two-level utilitarianism, reconstructs his argument to the principle of utility from the logic of moral discourse, and expands on his treatment of various "intuitive level system" rules that real-world utilitarians need for the conduct of daily life (which I refer to as "ILS rules" and which include laws, codes of professional ethics, a societal "common morality," and "personal moralities" of individuals). Part one also provides Harean responses to a range of standard objections to utilitarianism.

In part two, I argue that good utilitarian reasons can be given for recognizing a distinction among "persons," "near-persons," and "the merely sentient," where "persons" are defined as individuals with a biographical sense of self, and "near-persons" are defined as lacking that *biographical* sense of self, but as having a fairly robust, conscious sense of their past and future.

Finally, part three discusses various complications that would be involved in modifying a society's various types of ILS rules over time, as background ecological, technological, economic, and social conditions change, and gives one topical illustration: How conceptions of "humane sustainable agriculture" can and should change over time.

On that note, let me add that I have a sequel in the works, with the working title *Sustaining Animals: Envisioning Humane, Sustainable Communities*, which will discuss alternative developments of ILS rules regarding other areas of animal ethics, such as pets and working animals, wildlife scientific research on animals, and so on.

References

[1] Varner, Gary. 2012. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism*. New York: Oxford University Press.

On the Moral Significance of Persons, Near-Persons, and the Merely Sentient

Alastair Norcross, University of Colorado, USA

Gary Varner's *Personhood, Ethics, and Animal Cognition* is a thorough, clear, and persuasive defense of Hare's two-level utilitarianism, and an exploration of that approach with respect to what morality, both at the critical and intuitive levels, has to say about animals. I have long been a fan of the approach, and have defended a closely related approach (Peter Railton's (1984) "sophisticated consequentialism") in several places (Norcross 1997, 2010, and 2012). Because space is short, I will not linger over exposition, but will assume that my readers are familiar with at least the basic idea of the two-level approach. My concern in this short piece is with what Varner has to say about how the different cognitive abilities of different sentient creatures (including humans) should shape the intuitive-level system (ILS) rules concerning those different creatures.

In chapter 7, Varner argues that autonoetic consciousness gives the lives of those who possess it both "special moral significance" compared with the lives of those who are merely sentient, and "greater moral significance". Furthermore, different degrees and kinds of autonoetic consciousness give greater or lesser moral significance to the lives of those who possess the various degrees and kinds. So, the paradigm case of a creature with autonoetic consciousness is a person, who possesses a biographical sense of self. Other creatures, with lesser degrees of autonoetic consciousness may be "near-persons", whose lives have greater moral significance than the "merely sentient", but lesser than those of persons. So, what is the moral significance of autonoetic consciousness?

> From a utilitarian perspective, the most basic reason for thinking that autonoetic consciousness adds value to the life of an individual turns on the fact that the abilities to consciously remember the past and to consciously anticipate the future allow the individual to reexperience good (and bad) states of consciousness and to anticipate (and dread) future experiences. (Varner 2012, 162)

As I said, Varner's account of personhood centers on the possession of a biographical sense of self. He largely adopts Schechtman's "narrative self-constitution" view, which is a version of the view that persons are "storytelling beings" whose lives "can be richer and more complex than those of beings that lack the ability to tell stories" (Varner 2012, 139). So, roughly

(given space constraints), the idea is that the lives of some creatures possess greater moral significance (and special moral significance) than the lives of others, because there are more ways (respects) in which those creatures can be both benefitted and harmed than can others. This appeals to a "principle of inclusiveness", that Varner explains as follows:

> if we know that an experience A contains some value, and that experience B contains all of that value and more, then we know that experience B contains more value than experience A. (Varner 2012, 163)

There is obviously a lot to be said about Varner's nuanced and detailed account of the narrative self-constitution view of personhood, and the significance of various degrees of autonoetic consciousness. I want to focus here on a couple of worries about critical-level distinctions between kinds of moral significance, and a worry about justifying particular intuitive-level rules concerning different sentient creatures.

First, there is Varner's claim that the lives of persons, for example, possess "greater" moral significance than the lives of non-persons. He stresses, in chapter one, that he doesn't mean by this that their lives are preferable, but merely that they can be harmed and benefitted in ways that non-persons cannot, and thus that we should take "special care" in our dealings with them. I don't take issue with the underlying point, but I don't think it justifies the claim of "greater" moral significance. After all, there are many differences between different sentient creatures that affect the ways in which they can be harmed and benefitted. Birds, for example, can be harmed by having their wings clipped. We, not having wings, cannot be harmed in that way. Various forms of color-blindness can render certain kinds of aesthetic pleasures and pains inaccessible to some people. We wouldn't want to say that the lives of people with those forms of color-blindness had less moral significance than the lives of other people, even though it would be true that certain combinations of colors that might give pleasure or pain to other people would have no effect on them, and thus that we needn't take the same care in exposing them to these combinations as we would with others. This might seem like a terminological quibble. As I said, Varner doesn't claim that "greater moral significance" implies "more valuable." But I think there is a real danger that talk of greater moral significance will lead us to assume precisely that. It would be preferable to claim, simply, that the lives of persons have different moral significance from the lives of non-persons.

Second, I have a worry about Varner's appeal to the principle of inclusiveness. The idea seems to be that persons have the same range of valuable and disvaluable experiences as near-person, and more besides, and likewise for near-persons compared with non-persons.

That is, the greater cognitive abilities of persons serve to add to the range of morally relevant experiences. But couldn't a heightened degree of autonoetic consciousness also *detract* from a range of morally relevant experiences? Varner describes merely sentient creatures as beings who "live in the present." Perhaps the experiences of such creatures have a kind of value or disvalue that is utterly unlike anything we experience, and is, in fact, inaccessible to us, precisely because we possess the degree of autonoetic consciousness that we do. This is similar to a worry I have regarding John Stuart Mill's appeal to "competent judges" in evaluating different kinds of pleasure (1957, ch. 2). On Mill's account, the preferences of those who have experienced both kinds of pleasure are a reliable guide to the comparative values of two pleasures. But, it is possible (and, judging by my own experience, actual) that the experience of one kind of pleasure can itself change the experience (and maybe even the memory) of a different kind of pleasure. Thus the experience by me of both pleasures A and B may render the experience of pleasure A by someone who has only experienced A inaccessible to me. The same may be true of the mere ability to experience various kinds of pleasures (and pains). It is tempting, and certainly pleasing to our own sense of importance, to talk of the added "richness" and "complexity" of our mental lives as making possible greater value (while also recognizing the possibility of greater disvalue). But if, as I have briefly here suggested, the principle of inclusiveness doesn't apply, we simply have no reason to believe that our "richer" or "more complex" experiences possess more value (either positive or negative) than the (we assume) simpler experiences of creatures without autonoetic consciousness (or with a lesser degree of it than we possess).

Finally, when it comes to ILS rules regarding animals, Varner suggests that "With regard to merely sentient animals, … good intuitive level rules will generally treat them as replaceable" (Varner 2012, 288), and thus will allow what Varner calls "humane, slaughter-based agriculture." Singer, on the other hand, despite agreeing that our critical-level principles should regard merely sentient animals as replaceable, argues that the best ILS rules will forbid even humane, slaughter-based agriculture, on the grounds that this would foster the kind of attitude towards animals (regarding them as resources for our exploitation) that would likely encourage (or rather reinforce) a lack of consideration of the interests that such animals do have. It would be very difficult, in other words, for most people to maintain the view that animals are replaceable, without also failing to consider the interests they do have in not being made to suffer. Varner's response is that this move is in tension with Singer's dismissal of what he (Varner) sees as a similar objection to euthanasia for certain human infants. Singer argues that humans have shown themselves to be perfectly capable of recognizing important distinctions amongst different humans. But there are clearly important disanalogies here. In the case of allowing the killing of certain categories of humans, we are (certainly in modern societies) going against a strong presumption against killing all humans.

The burden of proof is on those who wish to carve out an exception to the rule against killing. In the case of killing animals, even now, the burden of proof is usually assumed to be on those who argue for the impermissibility of such killing. So ingrained is the attitude that killing animals is a morally trivial matter, that even a philosopher such as Don Marquis recently revised his famous argument against abortion (1989), on the grounds that the original argument might (just might) imply that the killing of squirrels was a morally serious matter (2015). Slippery slope arguments depend on the plausibility of the empirical claims that underlie them. Though I don't have the space to explore the issue here, it seems pretty clear to me, at least, that Singer is correct to claim that the slippery slope argument against certain limited forms of infant euthanasia rests on shaky empirical grounds, and also correct to claim that the slippery slope argument against ILS rules permitting humane slaughter-based agriculture rests on solid empirical foundations. I will add, as one final point, that the environmental and health-based reasons against animal agriculture seem to me to be individually decisive in any case, so the issue is overdetermined in favor of ILS rules prohibiting the practice.

References

[1] Mill, John Stuart. 1861/1957. *Utilitarianism*. Indianapolis: Bobbs-Merrill.

[2] Norcross, Alastair. 1997. "Consequentialism and Commitment." *Pacific Philosophical Quarterly* 78 (4): 380-403.

[3] _____. 2010. "Act-Utilitarianism and Promissory Obligation." In *Promises and Agreements: Philosophical Essays*, edited by Hanoch Sheinman, 217-236. Oxford/ New York: Oxford University Press.

[4] _____. 2012. "Consequentialism and Friendship." In *Thinking About Friendship: Historical and Contemporary Perspectives*, edited by Damian Caluori, 161-79. Basingstoke: Palgrave Macmillan.

[5] Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13 (2): 134-71.

[6] Varner, Gary. 2012. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism*. New York: Oxford University Press.

Commentary on Varner's *Personhood, Ethics, and Animal Cognition*

Adam Shriver, University of Oxford, UK

Gary Varner's *Personhood, Ethics, and Animal Cognition* expands upon ideas expressed in earlier essays and his first book *In Nature's Interests?* (2002) to more fully develop (and partially revise) a moral hierarchy that nicely captures several deep-seated intuitions about the value of sentience and personhood for moral status. His discussion in both books of the evidence for sentience and interests in different species is among the best in combining careful analytic thought with relevant scientific evidence, and due to the clarity of thought his work provides a useful starting point for those discussing sentience across species and its moral implications.

But sentience is only the baseline for moral considerability, and Varner applies a similarly careful analysis to the concept of personhood and a new category he introduces, near-personhood, in hopes of explaining the widely-shared view that there is something unique about the value of human life. Since speculating on the specialness of humans is a frequent theme of Western philosophy, it's important to be clear exactly what Varner is arguing. Some might believe that the pain of a human being is more morally important than a similar pain in a nonhuman animal; this, however, is not Varner's position. Rather, while essentially accepting something along the lines of Singer's Principle of Equal Consideration of Interests, which would entail that similar hedonic experiences should be treated similarly regardless of whom they occur in (and regardless of the species of the experiencing being), Varner nevertheless suggests that the overall life of a typical human has more value than that of a typical nonhuman because such a life is more "morally charged" in both positive and negative directions due to unique types of interests possessed by persons that can be layered on top of those available to near-persons and the merely sentient.

Though I am broadly sympathetic to Varner's arguments and overall project, I want to challenge one particular aspect of his account; namely, the suggestions that near-personhood and personhood can make a life go *worse* for an individual than the lives of merely sentient organisms. I believe that the "moral charge" is only in one direction; the capacities Varner discusses can make lives go morally better, but since the badness of suffering is what ultimately determines how badly bad lives can go and since there's no reason to believe that persons can suffer more than nonpersons, it follows that persons do not have lives that are

more "charged" in the negative direction. I'll briefly mention later why I think this is important.

Varner gives several different arguments for the claim that persons' lives are more morally charged than near-persons, which are in turn more morally charged than the merely sentient. We can divide these arguments into what I'll call experiential arguments and desire/interest-based arguments. The experiential arguments fail to establish that personhood increases the charge in either a positive or negative direction. Varner suggests that due to the capacity to project oneself into the future and past in the case of near-persons and to conceive one's life as a whole in the case of persons, these groups can consciously remember positive or negative experiences as well as consciously anticipate positive or negative experiences, thus "layering" an additional level of morally significant valance on top of already existing experiences.

However, the idea of layering doesn't capture how our experiences typically go. At any particular time, I have a limited capacity of attention which also limits my capacity for hedonic experience; if I direct my attention towards a pleasant memory, I will have less attention available for focusing on current experiences. Isn't this, after all, why so many self-help practitioners encourage us to "live in the moment"? If I'm eating a tasty vegan meal, but then start remembering a day at the beach, I will presumably start enjoying the meal less than if I had just focused on the experience of eating. There are some cases where having a memory of something similar to a current activity can help to enhance the experience of that activity (and Varner gives a few examples in the book), but it doesn't seem to me that the "pleasure enhancement" of such cases are strong enough that it can't simply be outweighed by very intense experiences that lack layering. So there's reason to be sceptical that conscious rememberings or anticipations importantly expand the overall capacity for positive or negative experiences in sentient beings at any particular moment and in aggregations of such moments.

That leaves the preference and interest based arguments, of which there are two types. One type of argument, similar to that above, suggests that both the capacity to anticipate the future and an autobiographical sense of self can give organisms additional types of desires that can be layered on top of the types of desires shared with the merely sentient. The other argument, which is unique to persons, is that only persons have "lives as a whole" which constitute a radically different type of interest from that possessed by non-persons.

As noted above, I'm broadly sympathetic to these arguments. But it's worth examining how these additional types of interests relate to the overall well-being of organisms. I agree with

Varner that being able to choose what type of life as a whole one pursues and to live according to that choice can make a life go better overall than it could for an organism that lacks that capacity. Moreover, I follow him in thinking that because persons have more to aspire to, it can be more "tragic" for a human life to be terminated prior to reaching its full potential compared to a non-person. However, it does not follow from these ideas that a person's life is in fact more "morally charged" in both positive and negative directions rather than just one.

To see this, imagine a world full entirely of persons who are incapable of having negative experiences. Their overall experiences are generally positive, but they have intense hopes and aspirations that they often fall short of achieving. In other words, they have many desires about the short and medium term future and about their lives as a whole that go unfulfilled, but they never feel negative experiences as a result of these desires being thwarted. Can these persons' lives ever reach a point where they fall below the threshold of a life worth living? Is it ever the case that it would be better, from their own perspective, for them not to have existed at all despite their overall pleasure?

I think the answer to both of these questions is "no." The badness of our desires about the future and about our lives as a whole being thwarted depends on (A) the loss of the goodness that would have obtained if they had been satisfied and (B) the negative feeling that results from knowing that the desires were thwarted. But if you take away (B), the feeling, then this thwarting is no longer the type of thing that pushes a life below a neutral point by itself. And we should remove (B) from consideration since, as I argued above, rememberings and anticipations don't seem to expend the overall capacity for valanced positive or negative experiences and there's no reason to believe that persons overall moment-to-moment experiences are more intense than those of the merely sentient.

In other words, while I agree that the types of capacities persons and near-persons have can make lives go better, I'm sceptical that they can make lives go worse than those of the merely sentient. Suffering is what makes life go badly, and—as Singer (1990) puts it—in suffering the animals are our equals. More sophisticated cognitive capacities may lead to different causes of suffering, different reasons for suffering, and perhaps even different flavors of suffering, but not to different magnitudes of suffering.

This is important when we think about the practical implications of Varner's view. Modern animal research is based on the claim that the harms caused to animals by invasive research is outweighed by the potential benefits to humans. In particular, consider the use of animal models in biomedical research for negative hedonic experiences such as pain, anxiety, and depression where it is clear that as a society we would not accept these experiences being

induced in humans merely for the sake of potential medical breakthroughs. Varner's claim that persons' lives have additional value beyond that of the merely sentient and that of near-persons has some implications for how we think of these tradeoffs, but even more dramatic implications in contexts such as these would follow from the claim that humans are capable of living far worse lives than those of other organisms. As such, I think it's important to pay attention to whether the moral charge truly increases in both directions, or only increases the potential for the goodness of lives. We can acknowledge that there's something special about the positive value of persons without making the (in my view mistaken) claim that persons' capacity for ill-being is also greater than that of other sentient animals.

References

[1] Varner, Gary. 2002. *In Nature's Interests? Interests, Animal Rights, and Environmental Ethics*. New York: Oxford University Press.

[2] _____. 2012. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism*. New York: Oxford University Press.

Treating Animals as the Sort of Thing They Are: Commentary on Gary Varner's *Personhood, Ethics, and Animal Cognition*

Susana Monsó, Messerli Research Institute, University of Veterinary Medicine Vienna, Austria

Gary Varner's *Personhood, Ethics, and Animal Cognition* has been my first encounter with Hare's ethical theory. As a convinced anti-utilitarian, I found Hare's views as unappealing and counter-intuitive as I expected to find them. However, the descriptive aspect of his theory makes it especially difficult to argue against him. After all, the reactions Hare's claims trigger in me are exactly the ones he would have predicted me to have, given the sort of ILS rules that I have been educated into adopting. How can you argue against someone whose theory predicts and incorporates the very objections you would like to raise?

In this commentary, I cannot do proper justice to the complexity and argumentative detail of Varner's *Personhood, Ethics, and Animal Cognition*, let alone attempt to refute Hare's ethical theory. Instead, I would like to accept the game that Hare and Varner are proposing, and attempt to raise an objection 'from within.' The idea is to capture one of the things that I find lacking in Varner's account of how we should treat animals, while at the same time avoiding the Harean standard response to objections. Thus, I will try to avoid any appeal to my intuitions and instead argue from the logic of moral discourse.

According to Varner's account of Hare, this author derived his ethical theory from the logical properties that all moral judgements share, namely, *universalizability* (the requirement to judge similar cases similarly), *overridingness* (the idea that moral norms override other types of norms), and *prescriptiveness* (the notion that sincerely assenting to a moral judgement implies acting accordingly and approving of others acting accordingly; Varner 2012, 12–13). From these three properties, Hare derives the principle of utility, the idea that the right thing to do in any circumstances is whatever would maximise aggregate happiness (Varner 2012, 71). I believe that this principle does not follow from these three logical properties (or, at least, I was not convinced by Varner's arguments to the effect), but what I want to focus on is the idea that happiness should be viewed as the sole prudential value that must be maximized. I will argue that pain and suffering, as well as pleasure and enjoyment, do not exhaust the harms and benefits that humans and other animals can be subjected to.

And I will argue that further harms and benefits can be derived from the very logic of moral discourse.

To see this, let me begin by pointing out that there is another logical requirement that moral judgements must accommodate, and which is not mentioned by Varner, even though he implicitly follows it. This is the requirement that I, for want of a better term, shall call *moral appropriateness*, and it consists of the idea that *the treatment that is morally appropriate for each being depends on the sort of being it is*, that is, the treatment of a being prescribed by moral judgements must take into account the morally relevant properties of that being. Moral appropriateness is a very basic notion in ethics. Regardless of whether one is a deontologist, a utilitarian, or a virtue ethicist, what one must do is considered to be intimately related to whether what one is dealing with is an inanimate object, a sentient being, an end-in-itself, and so on. The moral appropriateness requirement can even be seen as implicit in the logical requirement of universalizability. Indeed, the commitment to judge similar cases similarly incorporates the idea that beings with different morally relevant properties may call for different treatment, and that the properties that are deemed morally relevant in one case should be deemed morally relevant in all other cases.

As noted, Varner himself incorporates the moral appropriateness requirement. It is present, for instance, in the schema of the notion of person that he adopts: "an individual who deserves special treatment or respect ... *because* he, she, or it has some capacity or capacities ..." (Varner 2012, 6, emphasis in the original). The emphasis on the word 'because' suggests a commitment to the idea that the possession of certain capacities that are morally relevant grounds an entitlement to special treatment—an entitlement that is not present in beings that lack these morally relevant capacities. And indeed, Varner devotes a large portion of his book to distinguishing between persons, near-persons, and the merely sentient, precisely because he believes that these are morally relevant properties, and that the treatment that is morally appropriate is different for beings with different morally relevant properties.

The logical requirement of moral appropriateness is so fundamental that one can consider it to entail what Varner, following Hare, calls an ILS principle. This is an ILS rule that is not overrideable, and so functions as a deontological principle. Varner only speaks of two ILS principles in relation to animals: the universalizability principle (similar cases must be treated similarly) and the principle of respect for sentient life (sentient animals should not be killed unnecessarily). However, we could speak of a third ILS principle that would be intended to incorporate the moral appropriateness requirement. It could be formulated as:

all beings have a right to treatment that is appropriate to the sort of thing they are, or alternatively: *all beings have a right to have their morally relevant properties taken into account*.

With the moral appropriateness principle at hand, I can begin to articulate what I consider to be lacking in Varner's account of how we should treat animals. I will concentrate on the case of farm animals, which is the one that Varner deals with most extensively. Varner considers that the available evidence — at the time he wrote this book — suggests that farm animals are merely sentient animals, and so that they do not qualify as near-persons. Whether or not this corresponds to the current state of the evidence is unimportant for present purposes, for Varner's distinction between near-persons and the merely sentient has no real practical implications. This is because Varner considers that the well-being of both near-persons and the merely sentient is "purely a function of how much positive and negative conscious states they contain" (Varner 2012, 172). This means that Varner is assuming that the well-being of these animals can only be improved or worsened by the presence or absence of positive and negative subjective experiences. Now, when one treats an animal as though she could only be benefitted or harmed by the presence or absence of positive or negative experiences, one is effectively reducing that animal to a container of experiences, a 'mere receptacle of value,' as Tom Regan would put it (see, for instance: Regan 2004, 208-10.). If the animal in question is more than this, if she has other morally relevant properties, this means that we are not treating her as the sort of being she is—we are failing to follow the moral appropriateness principle.

This appears to be the case for farm animals, or at least those who are slaughtered in the largest numbers: pigs, cattle, and chickens. The available evidence suggests that these animals can not only feel pleasure and pain, but that they also have distinct personalities, preferences, tastes, feelings, complex social relationships — in short, they have *individuality* (see Marino 2017; Marino and Allen 2017; and Marino and Colvin 2015 for an up-to-date review of the relevant evidence). This individuality is a morally relevant property, for it is what makes them non-replaceable in a very literal sense: they are unique individuals and no other animal can fill their specific place. Their individuality grounds their non-replaceability in the same way it does for us. Indeed, as soon as we consider ourselves as individuals, rather than receptacles of positive and negative experiences, our non-replaceability becomes rather obvious. If I were to be killed because, say, a doctor decided to harvest my organs to save five different patients, my replacement with another human who would lead a happy life would be no consolation to any of those who know me and love me, because my uniqueness, my individuality, is lost and will never come back.

This individuality opens the door to a specific type of harm that is not captured in terms of experiential welfare: this is the harm that comes from objectifying or commodifying an animal who is a subject, an individual. This is a harm that farm animals are routinely subjected to. Indeed, the treatment of farm animals exhibits the seven symptoms of objectification identified by Martha Nussbaum (1995, 257):

1. Instrumentality: farm animals are treated as tools for the satisfaction of our food preferences.

2. Denial of autonomy: we determine the contents of farm animals' lives from beginning to end, denying them any opportunity for self-determination.

3. Inertness: farm animals are constrained in their movements and interactions with one another, as though they lacked agency and activity.

4. Fungibility: farm animals are treated as interchangeable with one another.

5. Violability: farm animals are treated as though they lacked boundary-integrity, as something that it is permitted to break up or break into. This is most obvious in the practice of slaughter, but is also present in other practices such as the castration of piglets or the de-beaking of chicks.

6. Ownership: farm animals are treated as things that can be owned, bought, and sold.

7. Denial of subjectivity: farm animals are treated as something whose experiences and feelings need not be taken into account, at least not for their own sake and only in those cases in which negative welfare impacts on productivity or on profitability.

The objectification of farm animals constitutes a form of harm because the animals' individuality is a morally relevant property that is not being taken into account, let alone respected. The animals are not being treated as unique, irreplaceable individuals with their own preferences and personalities, but as mere objects, as interchangeable units in the food production process. They are not being treated as the sort of thing they are, thus violating the moral appropriateness principle. Further, objectification is an objective form of harm that is not dependent on the presence of certain negative mental states. This is one of the reasons why the cognitive disenhancement of farm animals to deprive them of the capacity to feel pain or to see is not, contrary to what Varner (2012, 276-78) says, entirely beneficial for the animals. While cognitive disenhancement eliminates one dimension of harm (negative mental states), it is another form of objectification, since these animals' negative mental states also contribute to their individuality, and by eliminating them we are stripping them

of some of their uniqueness. Thus, such cognitive disenhancement would also be a harm imposed on these animals, even if it doesn't result in any specific negative mental state.

Typically, utilitarians do not accept that objective harms exist. However, I suspect that Varner needs to incorporate objective harms for his theory to work. If only subjective harm counts, then nothing can ensure that the lives of persons are more 'morally charged' than those of animals. While Varner argues strongly for the idea that our biographical sense of self entails that we have more value and disvalue in our lives than animals do, if we think of value and disvalue as merely a function of positive and negative mental states, then nothing guarantees this. If the value of achieving my life-long dream is merely a function of how content it makes me feel, there is no way for me to know whether this value is superior, equivalent, or inferior to value of the pleasure of my dog as he basks in the sun with not a care in the world. The way for Varner to escape this is to incorporate an objective dimension, which indeed he seems to do. He asserts, for example, that for a person, "how well his life goes is not completely addressed by asking how good it felt, on the whole, to live that life" and that what is in a person's best interests is "to live a good story, to be a certain kind of person, to achieve certain things" (Varner 2012, 172). But if we accept the existence of objective harms and benefits in the case of humans, why not accept it in the case of animals?

References

[1] Marino, Lori. 2017. "Thinking Chickens: A Review of Cognition, Emotion, and Behavior in the Domestic Chicken." *Animal Cognition* 20 (2): 127–47.

[2] _____, and Kristin Allen. 2017. "The Psychology of Cows." *Animal Behavior and Cognition* 4 (4): 474-98.

[3] _____, and Christina M. Colvin. 2015. "Thinking Pigs: A Comparative Review of Cognition, Emotion, and Personality in Sus Domesticus." *International Journal of Comparative Psychology* 28 (1). https://escholarship.org/uc/item/8sx4s79c.

[4] Nussbaum, Martha C. 1995. "Objectification." *Philosophy & Public Affairs* 24 (4): 249–91.

[5] Regan, Tom. 2004. *The Case for Animal Rights. Updated with a New Preface edition.* Berkeley: University of California Press.

[6] Varner, Gary. 2012. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two-Level Utilitarianism*. New York: Oxford University Press.

Varner on Animals: Room for Far-Persons?

Gary Comstock, North Carolina State University Raleigh, USA

In distinguishing near-persons from persons and the merely sentient, Gary Varner (2012) supplies an important set of categories to classify the moral statuses of three different kinds of animals:

> Persons, such as typically developing adult humans
>
> Near-persons, such as the great apes, cetaceans, elephants and perhaps corvids, parrots, scrub jays and others
>
> The merely sentient, such as, perhaps, fish

In which category should we put companion and so-called "food animals:" dogs, cats, cows, pigs, and chickens? Varner thinks there is not yet enough empirical evidence to decide the question. He asserts, and I agree, that domesticated animals lack the more complex psychologies of the great apes. Does this mean, then, that cows are merely sentient? Given the gaps in our current state of knowledge, Varner hesitates, and does not place domesticated animals in any category. Should he have included domesticated animals as near-persons?

I think not. Domesticated animals have their own space, somewhere between near-persons and the merely sentient. We need a new category, which I'll call far-persons (Comstock 2017b). To define far-persons I'll help myself to Varner's conceptual tools. In so doing, I take myself to be filling in part of the middle of what he describes as a continuum.

A person, writes Varner, is an individual whose biographical sense of self entitles them to special treatment. A near-person lacks a biographical sense of self but has a robust autonoetic consciousness: a rich sense of their past, present, and future. Near-persons do more than simply respond to stimuli, the defining characteristic of the merely sentient. Near-persons lack ground projects, the hallmark beliefs and desires persons have with which they aspire to make something out of their lives as a whole. But near-persons have conscious beliefs and desires, and executive control of them. In this regard they differ from the merely sentient who have only the ability to enjoy pleasant stimuli while responding defensively to noxious stimuli.

If I'm right, domesticated animals are neither near-persons nor merely sentient. Take cows, for example. Cows, unlike chimpanzees and other near-persons, lack a robust autonoetic consciousness. They are not able to think about themselves, remember the past episodically, or make plans for tomorrow. They are not near-persons.

However, neither do they seem to be merely sentient. They do not seem to live entirely in the moment. They seem to have concepts, cognitive representations, beliefs and desires with which to form explicit, if short-term, memories and prospections. They are able to look forward as far as several minutes into the future, to form hypotheses about how to achieve goals within that timeframe, and flexibly to change strategies if pursuing one strategy initially fails (Comstock 2017a). If cows are not simply pushed or pulled in this or that direction by external forces, if they have executive control and the ability to inhibit urges, then their behavior cannot be explained in terms of stimulus and response.

I think cows are far-persons, individuals with lyrical experiences, experiences with "minute" temporal horizons that last at most two or three minutes. Lyrical experiences can be intense and profoundly pleasurable or painful, but they do not persist in memory and cannot be achieved by long-term planning. Far-persons have four basic emotions: happiness, sadness, anger, and fear or surprise (Jack, Garrod, and Schyns 2014). They can take pride in their successes (or be frustrated by their failures) to learn simple new skills. They are sentient beings who can recognize faces and vocalizations of conspecific friends and enemies, and human friends and enemies. That is, they know their foes and how to avoid them, and they know their friends and how to please them (think licking and grooming). Far-persons have a unified perspective—a point of view of their own—but they lack long-term conscious memories and aspirations. They lack second-order beliefs about their own minds, and have no beliefs about the beliefs of others. Their beliefs are all "first order," that is, directed outward at objects in the world. They can exercise executive control over some of their desires on the basis of reasons, act autonomously as they pursue their aims, communicate their desires to others, and deceive others. But they cannot read others' minds. Although they lack explicit memories of things that happened weeks or months ago, they have explicit memories of what happened a few seconds or minutes ago.

Varner does not address the moral status of domesticated animals in detail. He observes that he has not yet seen behavioral evidence that these animals have a robust autonoetic consciousness, and he advises that we err on the side of caution when adopting policies about how to treat them. Given the state of inquiry, Varner's position is reasonable; domesticated animals should not be listed among possible candidates for near-personhood. But in what category should we place them? If we take our clues from Varner's analysis of wild squirrels, we might infer that he thinks domesticated animals are merely sentient.

Varner argues that squirrels do not plan consciously for the future, do not have episodic memories, and lack theory of mind. There is no evidence he writes, and I concur, that squirrels recognize themselves in mirrors or have any other of the distinguishing behavioral features of near-persons. They have implicit memories and unconscious anticipations of the future, but these capacities are not sufficient to sponsor explicit, conscious future planning. A squirrel hoarding acorns, Varner writes, consciously desires "to get each acorn into its stash" but "is completely unconscious of the purpose of its hoarding behavior" (Varner 2012, 164). Since it is not conscious of the long-term goal it

> can achieve no sense of satisfaction when it has stashed enough acorns. It cannot, in effect, say 'There, I've accomplished *that*! Since *that* (the goal of laying up enough acorns for the winter) is something of which it is not conscious. (Varner 2012, 164)

I think Varner's right that the squirrel does not consciously plan for the long-distance future, but what about the short-term? Varner allows that the squirrel can "achieve a sense of satisfaction" from achieving a short-term objective, for example, getting *this* acorn into *that* stash. A squirrel cannot get satisfaction at the end of the day from having put in an honest day's labor, as it were, because it does not possess the concept of "a day's labor." But the squirrel may well have these three mental states (Carruthers 2008):

> [BELIEF] if this acorn is cached, then it can be eaten when hungry
>
> [BELIEF] if this acorn is not cached, then it cannot be eaten when hungry
>
> [DESIRE] eat this acorn when hungry

These three mental states automatically produce caching behavior. If the caching behavior is blocked by, say, a heavy rock that has fallen on the cache, the squirrel may acquire another set of relevant beliefs and desires and this subsequent set may enable the squirrel to negotiate the obstacle and complete the task. Perhaps forming and acting on the novel hypothesis enables the squirrel to succeed in achieving its original goal (to *eat this acorn when hungry*). If so, and if the squirrel recruits frontal cortical brain structures to inhibit and control its thoughts—as does the squirrel's close rodent cousin, the rat (Narayanan and Laubach 2017)—then there is little reason to deny that the behavior is conscious.

And that would be a lyrical experience: autonomously adopted, intentionally planned, rationally executed, potentially involving emotions and facial recognition, a potential source of pleasure. Even if the experience is short-term and stretches little more than two dozen seconds from beginning to end, it may be a conscious plan that requires the animal to

employ concepts, beliefs, and desires. Table 1 summarizes my proposal about far-persons' mental capacities.

I have argued for what I initially thought was a friendly amendment to Varner's insightful three-fold scheme. However, upon reflection, the argument now seems to me to require a significant revision of Varner's lowest category. For, if I am right, there are no "merely sentient" animals. The animals I had previously thought would fall into this category are, I now think, either far-persons or not sentient at all. Dogs and cats, for example, are far-persons, not merely sentient. Worms and insects (and perhaps mollusks and even fish), on the other hand, probably lack a standpoint from which to integrate inputs or characterize the various aspects of the world. Lacking a unified perspective, they cannot coherently be said to desire that their future be free of any present noxious stimulus. For without a perspective, they cannot have desires, beliefs, or temporal horizons. If there are such animals, as there certainly appear to be, then all of them drop out of the "merely sentient" category because they are only "merely responsive" not "merely sentient." And if, as I suspect, all of the other animals in the merely sentient category (such as cows and pigs) are actually far-persons, then it's possible that the category itself is empty.

FAR-PERSONS *Lyrical consciousness* Simple minds	1. Unified perspective	Centrally integrates multi-modal perceptual inputs from the immediate environment providing a characterization of the various aspects of the external world, information that becomes the basis for beliefs about how objects are arrayed relative to one's standpoint
	2. Short temporal horizons	Has conscious short-term prospections stretching no more than a few dozen seconds into the future; unconscious or procedural long term memories reaching back weeks into the past; and conscious or episodic intermediate term memories reaching back a few dozen minutes into the past
	3. Concepts (semantics)	Has first-order beliefs, that is, positive and negative attitudes toward the contents of propositions (if the bee believes "this is the hive" then it has a positive attitude toward the proposition "this is the hive"); has desires, that is, a disposition to take pleasure in some immediate future state of the world; can receive and express information (e.g., invitations, warnings) visually or orally*; can identify some others as friend or foe (e.g., facial recognition)*
	4. Causal reasoning, reversal learning	*Has object permanence;* understands cause and effect; able to form alternative hypotheses and choose rationally among them to achieve a goal; adapts behavior relatively quickly in response to changes in familiar reward patterns
	5. Sentience and emotion	Has phenomenal experiences; feels pleasure, pain, and basic emotions (anger, happiness, surprise, disgust, sadness, and fear); takes an egoistic interest in one's immediate welfare

NEAR-PERSONS *Autonoetic consciousness* Complex minds	6. Autonomous agency	Has feelings of freedom; feels frustrated when one's desires are thwarted; feels one has executive control of one's decisions; has second-order beliefs (beliefs about one's beliefs)
	7. Intermediate temporal horizons	Conscious intermediate term prospections stretching no more than a few hours into the future; conscious or episodic long-term memories of a few days into the past
	8. Grammar (syntax)	Understands changes in semantic meaning when the same concepts in one expression are re-arranged into a second expression
	9. Theory of mind, self-conscious	Understands the behavior of other selves as motivated by psychological states similar to one's own; recognize oneself as a conscious individual extended temporally across the hours; pass the mirror mark test
	10. Empathy, altruism	Shares others' feelings, engages in social relationships; takes an allocentric interest in the welfare of others in one's in-group (family, kin, tribe)
PERSONS *Narrative consciousness* Moral minds	11. Autobiographical control	Has ground projects, can shape long stretches of one's life into stories of one's own making
	12. Long temporal horizons	Has conscious long-term prospections stretching years into the future; and conscious long-term memories stretching decades into the past
	13. Narrative	Understands oneself in narrative terms, as a character in a plot with moods and settings; employs one of three mechanisms to control and enact one's chosen story: language, pictures, or music
	14. Moral rights	Has moral self-governance, the feeling that one is in control of oneself; can inhibit "instinctual" impulses to act on principles no rational and fair-minded person could reject; possesses negative rights against others that they not violate one's life or liberty
	15. Moral obligations	Has a sense of justice; takes a universalizable interest in the welfare of those outside one's in-group; assumes the second-person standpoint, recognizes that many individuals outside one's in-group deserve equal respectful treatment and should not be used only as a means to one's ends; understands actions as blameworthy or praiseworthy

TABLE: FAR-PERSONS' PSYCHOLOGICAL CAPACITIES

References

[1] Carruthers, Peter. 2008. "Meta-cognition in Animals: A Skeptical Look." *Mind & Language* 23 (1): 58-89.

[2] Comstock, Gary. 2017a. "Concerning Cattle: Behavioral and Neuroscientific Evidence for Pain, Desire, and Self-Consciousness." In *Oxford Handbook of Food Ethics*, edited by Anne Barnhill, Mark Budolfson, and Tyler Doggett, 139-69. Oxford: Oxford University Press.

[3] _____. 2017b. “Far-Persons.” In *Ethical and Political Approaches to Nonhuman Animal Issues*, edited by Andrew Woodhall, and Gabriel Garmendia da Trindade, 39-71. Cham, Switzerland: Palgrave Macmillan.

[4] Jack, Rachael E., Oliver G. B. Garrod, and Philippe G. Schyns. 2014. “Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time.” *Current Biology* 24 (2): 187-92.

[5] Narayanan, Nandakumar S., and Mark Laubach. 2017. “Inhibitory Control: Mapping Medial Frontal Cortex.” *Current Biology* 27 (4): R148-50. DOI: 10.1016/j.cub.2017.01.010.

[6] Varner, Gary. 2012. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare’s Two-Level Utilitarianism*. New York: Oxford University Press.

Replies to Norcross, Shriver, Monsó, and Comstock

Gary Varner, University of Texas, USA

Given the limits that the commentators and I all agreed to for our contributions to this panel, I will focus on a limited number of the points that the four commentators have made.

A theoretical concern that Alastair Norcross raises is an argument against my use of Ralph Barton Perry's principle of inclusiveness. He doubts that the principle applies to comparisons of the experiences of persons and non-persons for the same reason he doubts that it applies to John Stuart Mill's comparison of the experiences of "the competent judges" to those who are not currently "capable of appreciating" what Mill concludes are qualitatively superior pleasures. The reason is that acquaintance with new experiences can alter how one experiences previously enjoyed activities, and Norcross says that introspectively, he thinks that this is true. The extent to which this is true is, I think, an empirical question, and my own introspection doesn't jibe with Norcross'. I tend to think that if I shed what Mill referred to as "any feeling of moral obligation to prefer one pleasure to another" (Mill 1957, 12), then I can immerse myself in the previously enjoyed activity and not have my appreciation of it perverted. I often find my students saying that they think that they prefer the dissatisfied human's life because they're expected to, and that's what I say in response. But whether or not one really *can* shed the socialization in question is a very big *if!*

Norcross also worries that allowing humane slaughter of merely sentient animals will "likely encourage (or rather reinforce) a lack of consideration of [their] interests," a point that Susana Monsó also worries about. When I responded to this line of argument in the book, I noted that Singer replies to such slippery slope arguments against human euthanasia by saying that, as Norcross puts it, "humans have shown themselves to be perfectly capable of recognizing important distinctions amongst different human beings." But, Norcross suggests, in the case of humans, this is because

> the burden of proof is on those who wish to carve out an exception to the rule against killing, [whereas] in the case of animals, even now, the burden of proof is usually assumed to be on those who argue for the impermissibility of such killing.

I take this kind of worry seriously, and to the extent that it is well-founded, I agree that it should push us in the direction of erring on the side of caution in the formulation of ILS

rules. Hare's type of utilitarianism acknowledges that the weaknesses of human beings should be taken seriously when formulating ILS rules, and I may be dead wrong about our abilities to resist this.[1]

Adam Shriver raises issues about both of my arguments for the relatively "charged" nature of persons' lives over those of near-persons, and of near-persons' lives over those of the merely sentient.

First, he argues that my "experiential" argument works in neither the positive nor the negative direction. This is because conscious organisms may have "limited capacity for attention available for focusing on current experiences," so he is "sceptical that conscious rememberings or anticipations expand the overall capacity for [either] positive or negative experiences." In response I will say only that he may well be correct, but I think this is an empirical question about human psychology and neuroscience. It is one that Shriver, among other scholars, is particularly well positioned to address in the future, however, given his strongly interdisciplinary preparation that includes the study of neuroscience.

Regarding my "preference- and interest- based arguments" for increased moral charge, Shriver argues that they work in only the negative direction. He agrees with me that experiencing the satisfaction of having what I call one's interest in one's life-as-a-whole can *add* value to one's life, but, he says, "if you take away [that] feeling [of satisfaction], then this thwarting is no longer the type of thing that [decreases the value of one's life.]" I wasn't clear exactly what Shriver meant by this. For my point was that if one takes a conscious interest in how one's life-as-a-whole goes, and it goes badly, then that layers dissatisfaction on top of the various dissatisfactions that a non-person (i.e. one incapable of taking an interest in how its like-as-a-whole goes) the dissatisfactions that a non-person is capable of. But some of what Shriver says in that part of his commentary suggests another interpretation of Shriver's concern that is explicitly raised at the end of Susana Monsó's comments, regarding concerning what are commonly called "objective harms."

Monsó observes that "Typically, utilitarians do not accept that objective harms exist," and I long *thought* (but notice the past-perfect tense there) I long *thought* that I wanted to say

[1] Norcross also argues that the fact that some individuals (e.g. color-blind people) can be harmed in ways that others (e.g. normally sighted people) cannot, doesn't suffice to show that the former's lives have greater moral significance than the latter. In the interests of time, I'm not addressing this directly. I think it may be of a piece with his concern at the end about slippery slope arguments and ILS rules.

that nothing harms an individual unless and until its valenced[2] phenomenally conscious states are affected. So a tumor in my lung doesn't harm me at all if I'm killed by a bus before it affects me (in the animal science literature, this is Ian Duncan's (1996) conception of animal welfare as "all about feelings"). And although certain cognitive abilities allow one to be harmed by a diagnosis of cancer before the tumor adversely affects my physical health, an individual without the ability to understand the diagnosis isn't harmed until it adversely affects their physical health. In a similar vein, I have also long *thought* that I didn't want to endorse objective list views, according to which achieving certain things or exercising certain capacities is objectively good for an individual regardless of what attitudes the individual has towards those achievements or exercises. Achieving certain things, exercising capacities, and having biological functions in place are all things that show up on various "objective list theories" of the good.

But I use the past perfect tense, because since I adopted a biographical conception of personhood for the first time in this book, I've been tempted to say that when it comes to persons, they *can* be harmed (*and benefitted*) by things that they don't even know about. That does seem possible if one embraces a biography-based account of how well one's life-as-a-whole goes, by opening the door to saying that events after one's death can affect that interest. In particular, as I noted at one point in the book, it could mean that events after one's death can impact how well your life story played out. In the book, I didn't try to sort this issue out any further, claiming that since the book is about our treatment of non-human animals, and I argued in the book that no non-humans have an interest in how their "lives-as-a-whole" go, I could punt on the issue (Varner 2012, 137, fn. 2). But on reflection inspired by Monsó's (and implicitly) Shriver's comments, I can't continue to punt. For if, as Monsó suggests, I'm committed to some kind of objective list view with regard to persons, then I can't continue to ignore the type of objection that she raises: that there may be ways of harming *non*-persons that can't be captured in terms of their subjective experiences. So I need to decide whether or not I want to stick to my idea that what is in persons' best interests is a function of the life stories that they have chosen à la Marya Schechtman (1996). To be consistent, I need to either drop that account of why persons have a special kind of interest that non-persons lack, or explain why objective list accounts make sense for persons but *not* for non-persons.

[2] I say "valenced" here, because I think it is possible to have non-valenced phenomenologically conscious states, e.g. an experience of seeing blue has a phenomenal feel, but in and of itself it has no positive or negative valence. Without positive or negative valences, however, I don't think that such phenomenologically conscious states matter, morally speaking.

Finally, regarding Gary Comstock's comments, in the book I said that "merely sentient" animals, which live entirely in the present, might be a "hypothetical construct." The reason is that there may be no animals that are sentient that do not experience physical pain,[3] and pain is clinically defined as something that one desires to end. That is a desire for things to change in the future, if only the very *immediate* future (Varner 2012, 22). At the end of his commentary, Comstock speculates that the category of "merely sentient" animals may be empty for related reasons.

I also emphasized in the book that there is probably a "continuum" of conscious awareness of the future and past (see Varner 2012, 165 and 22), so that a simple distinction between "near-persons" and "persons" cuts with too dull a knife. Comstock's work on "far-persons" provides a more fine-grained distinction. In both his comments here, and in a 2017 paper, Comstock argues that, for reasons parallel to those that I gave for recognizing near-persons as a distinct category from both persons and the merely sentient, we should also recognize his category of far-persons. While far-persons lack autonoetic consciousness that is as "robust" as that of near-persons, far-persons have what he describes as "*lyrical* experiences" that transcend the immediate present of the merely sentient. He stipulates that the temporal horizon of far-persons' autonoetic experiences is less than "beyond the onset of the next sleep cycle" (2017b, 46), in contrast to near-persons' autonoetic experiences that extend beyond the present day or (as he puts it in the 2017b paper) a given "sleep cycle."[4] Just like near-persons, far-persons would lack an interest in how their lives-as-a-whole go, but just like near-persons, the autonoetic experiences of far-persons transcend the immediate present, and this allows the same layering of value on top of merely sentient animals' experience of the present. So if that is a good argument for the conclusion that near-persons' lives are more "morally charged" than those of the merely sentient, it's also a good argument for recognizing that the lives of far-persons are more "morally charged" than those of the merely sentient, while they are *less* "morally charged" than those of both near-persons and persons. Both the "experiential" and "preference- and interest-based arguments" that Shriver described in his comments would apply to far-persons, so I definitely

[3] The one exception is humans with congenital insensitivity to pain (CIP). The condition is highly maladaptive, however, as documented in Melody Gilbert's 2005 film *A Life Without Pain* (http://alifewithoutpain.com; Varner 2012, 106), and this means that no non-human animals with the condition would survive to adulthood.

[4] He does this, I think, because in my book I discussed the "planning for breakfast" experiments that have been conducted on great apes and scrub jays.

agree with Comstock's general argument for using his category of far-persons in the formulation of ILS rules.[5]

By way of concluding, I want to thank Comstock for organizing the ISUS panel on my book, and Norcross, Monsó, and Shriver for joining as additional commentators. As I trust will be apparent from my responses, their comments will be genuinely helpful to me as I continue working on the sequel, and I plan to acknowledge their help improving that book.[6]

References

[1] Comstock, Gary. 2017b. "Far-Persons." In *Ethical and Political Approaches to Nonhuman Animal Issues*, edited by Andrew Woodhall, and Gabriel Garmendia da Trindade, 39-71. Cham, Switzerland: Palgrave Macmillan.

[2] _____. 2021. "Varner on Animals: Room for Far-Persons?" In *Utility, Progress, Technology (Proceedings of the XVth Conference of the International Society for Utilitarian Studies)*, edited by M. Schefczyk and C. Schmidt-Petri, 353-358. Karlsruhe: KIT Scientific Publishing.

[3] Duncan, Ian. 1996. "Animal Welfare Defined in Terms of Feelings." *Acta Agriculturae Scandinavica*, Section A, Animal Science Supplement 27: 29-35.

[4] Mill, John Stuart. 1861/1957. *Utilitarianism*. Indianapolis: Bobbs-Merrill.

[5] Monsó, Susana. 2021. "Treating Animals as the Sort of Thing They Are: Commentary on Gary Varner's *Personhood, Ethics, and Animal Cognition.*" In *Utility, Progress, Technology (Proceedings of the XVth Conference of the International Society for Utilitarian Studies)*, edited by M. Schefczyk and C. Schmidt-Petri, 347-351. Karlsruhe: KIT Scientific Publishing.

[5] A related caveat: just as I said in the book that distinguishing between near-persons and the merely sentient might not be justified for the ILS rules of societies under certain ecological conditions (e.g. pre-contact First Americans), the same goes for distinguishing between far-persons and both near-persons and persons. Under present conditions in affluent Western countries, however, I'm inclined to agree with Comstock that we can (as it were) afford to do so, e.g. by moving faster toward not raising mammals (and perhaps not birds) in slaughter-based agriculture.

[6] Also, I welcome questions about parts of their comments that I didn't explicitly address. I can be reached at: g-varner@tamu.edu.

[6] Norcross, Alastair. 2021. “On the Moral Significance of Persons, Near-Persons, and the Merely Sentient.” In *Utility, Progress, Technology (Proceedings of the XVth Conference of the International Society for Utilitarian Studies)*, edited by M. Schefczyk and C. Schmidt-Petri, 339-342. Karlsruhe: KIT Scientific Publishing.

[7] Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca, NY: Cornell University Press.

[8] Shriver, Adam. 2021. “Commentary on Varner’s *Personhood, Ethics, and Animal Cognition*.” In *Utility, Progress, Technology (Proceedings of the XVth Conference of the International Society for Utilitarian Studies)*, edited by M. Schefczyk and C. Schmidt-Petri, 343-346. Karlsruhe: KIT Scientific Publishing.

[9] Varner, Gary. 2012. *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare’s Two-Level Utilitarianism*. New York: Oxford University Press.

This volume collects selected papers delivered at the 15th Conference of the International Society for Utilitarian Studies, which was held at Karlsruhe Institute of Technology in July 2018. It includes papers dealing with the past, present, and future of utilitarianism – the theory that human happiness is the fundamental moral value – as well as on its applications to animal ethics, population ethics, and the future of humanity, among other topics.

ISBN 978-3-7315-1108-3