Stockholm University

Moral Lessons from Psychology

Contemporary Themes in Psychological Research
and their Relevance for Ethical Theory

Henrik Ahlenius

Doctoral Thesis in Practical Philosophy at Stockholm University, Sweden 2020

Moral Lessons from Psychology

Contemporary Themes in Psychological Research and their Relevance for Ethical Theory

Henrik Ahlenius

Academic dissertation for the Degree of Doctor of Philosophy in Practical Philosophy at Stockholm University to be publicly defended on Friday 18 December 2020 at 13.00 online via Zoom, public link is available at the department website.

Abstract

The thesis investigates the implications for moral philosophy of research in psychology. In addition to an introduction and concluding remarks, the thesis consists of four chapters, each exploring various more specific challenges or inputs to moral philosophy from cognitive, social, personality, developmental, and evolutionary psychology. Chapter 1 explores and clarifies the issue of whether or not morality is innate. The chapter's general conclusion is that evolution has equipped us with a basic suite of emotions that shape our moral judgments in important ways. Chapter 2 presents and investigates the challenge presented to deontological ethics by Joshua Greene's so-called dual process theory. The chapter partly agrees with his conclusion that the dual process view neutralizes some common criticisms against utilitarianism founded on deontological intuitions, but also points to avenues left to explore for deontologists. Chapter 3 focuses on Katarzyna de Lazari-Radek and Peter Singer's suggestion that utilitarianism is less vulnerable to so-called evolutionary debunking than other moral theories. The chapter is by and large critical of their attempt. In the final chapter 4, attention is directed at the issue of whether or not social psychology has shown that people lack stable character traits, and hence that the virtue ethical view is premised on false or tenuous assumptions. Though this so-called situationist challenge at one time seemed like a serious threat to virtue ethics, the chapter argues for a moderate position, pointing to the fragility of much of the empirical research invoked to substantiate this challenge while also suggesting revisions to the virtue ethical view as such.

Keywords: *consequentialism, deontology, emotion, ethics, evolution, innate, moral judgment, moral philosophy, psychology, utilitarianism, virtue.*

Stockholm 2020
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-185993

ISBN 978-91-7911-354-4
ISBN 978-91-7911-355-1

Department of Philosophy

Stockholm University, 106 91 Stockholm

MORAL LESSONS FROM PSYCHOLOGY

Henrik Ahlenius

Stockholm
University

Moral Lessons from Psychology

Contemporary Themes in Psychological Research
and their Relevance for Ethical Theory

Henrik Ahlenius

ISBN print 978-91-7911-354-4
ISBN PDF 978-91-7911-355-1

Printed in Sweden by Universitetsservice US-AB, Stockholm 2020

To my beloved Lisa
and to our kids
Stella and Tom

I had a strange dream, or half-waking vision, not long ago. I found myself at the top of a mountain in the mist, feeling very pleased with myself, not just for having climbed the mountain, but for having achieved my life's ambition, to find a way of answering moral questions rationally. But as I was preening myself on this achievement, the mist began to clear, and I saw that I was surrounded on the mountaintop by the graves of all those other philosophers, great and small, who had had the same ambition, and thought they had achieved it. And I have come to see, reflecting on my dream, that, ever since, the hard-working philosophical worms had been nibbling away at their systems and showing that the achievement was an illusion.

Richard Hare

Preface

The goddamn thesis is finally finished! I started graduate school so long ago it has become an embarrassment. Former students of mine finished before I did – something that is, our head of department informed me in brazen denial of the thesis' major presupposition, "unnatural and, *hence*, wrong". A friend told me, in yet another misguided attempt at cheering me up, that a possible upside to working for such a long time on the dissertation is that my chapter on the evolutionary origins of moral thinking must be the first ever Pleistocene eyewitness account of what really happened back then. Oh well. Anyway, here it is.

I wish to thank the many friends and current and previous colleagues who have provided support over the years. These include Sama Agahi, Jonas Åkerman, Gustav Alexandrie, Per Algander, Simon Allzén, Erik Angner, Gustaf Arrhenius, Andrea Asker Svedberg, Conrad Bakka, Lars Bergström, Katharina Berndt Rasmussen, Stina Björkholm, Greg Bognar, Björn Brunnander, William Bülow O'Nils, Åsa Burman, Staffan Carlshamre, Åsa Carlsson, Jens Dam Ziska, Hege Dypedokk Johnsen, Jonathan Egeland Harouny, Karin Enflo, Björn Eriksson, Romy Eskens, Daan Evers, Maria Forsberg, Anna Petronella Foultier, Lisa Furberg, Kathrin Glüer-Pagin, Jimmy Goodrich, Johan Gustafsson, Gösta Grönroos, Sören Häggqvist, Bob Hartman, Anandi Hattiangadi, Lisa Hecht, Mattias Högström, Madeleine Hyde, Mats Ingelström, Mikael Janvid, François Jaquet, Sofia Jeppsson, Eric Johannesson, Hana Kalpak, Karl Karlander, Mirre Khan Oidermaa, Ulrik Kihlbom, Simon Knutsson, Palle Leth, Johan Lindberg, Sandra Lindgren, Anders Lundstedt, Hans Mathlein, Andreas Mauz, Victor Moberger, Niklas Möller, Tara Nanavazadeh, Pavlo Narvaja, Jonas Nordebrand, Karl Nygren, the late Ragnar Ohlsson, Niklas Olsson Yaouzis, Sara Packalén, Peter Pagin, Martin Peterson, Anna Petrén, Mikael Pettersson, Dag Prawitz, Marcel Quarfood, Daniel Ramöller, Emma Runestig, Peter Ryman, Håkan Salwén, Stefan Schubert, Levi Spectre, Henning Strandin, Maria Svedberg, Gunnar Svensson, Kjell Svensson, Nils Sylvan, Claudio Tamburrini, Torbjörn Tännsjö, Folke Tersman, Amanda Thorell, Olle Torpman, Hans-Jörgen Ulfstedt, Emma Wallin, and Åsa Wikforss.

An early version of chapter two was presented at University of Gothenburg's Department of Philosophy, Linguistics, and Theory of Science. Many thanks to organizer Ingmar Persson and all the others who participated in those discussions.

I was most fortunate to be able to spend a semester working with Gilbert Harman at lovely Princeton University. I thank Gil and the many new friends I made while there, in particular Mark Budolfson, Angela Mendelovici, Philipp Koralus, and Jack Spencer. My stay at Princeton was made possible by grants from The Swedish Foundation for International Cooperation in Research and Higher Education (STINT) and from the Anders Karitz Foundation. I am very grateful to both of these institutions for their support.

For many years, I've had a second academic home at Karolinska Institute's ethics group within the Department for Learning, Informatics, Management, and Ethics (LIME). I was given the opportunity to present an early version of chapter four there, and I wish to thank my many good friends at LIME: Gert Helgesson, Annelie Jonsson, Niklas Juth, Petter Karlsson, Anna Lindblad, Niels Lynøe, Tomas Månsson, and Manne Sjöstrand.

A third and more recent academic home has been DIS where I've taught a course on medical ethics for American students spending a semester in Stockholm. Many thanks to Anne Bachmann, Louise Bagger Iversen, Kim Bergqvist, Jim Breen, Susana Dietrich, Natalia Landázuri Sáenz, Tina Mangieri, Mark Peters, Steve Turner, and many other friends and colleagues at both the Stockholm and Copenhagen offices.

Jens Johansson was opponent at my mock viva in 2019, and I thank him for his many valuable suggestions. After that, Gunnar Björnsson and Krister Bykvist formed the departmental internal assessment committee and pointed to several remaining shortcomings which I've done my best to rectify. Thank you both.

Thank you also to associate professor Charlotte Alm of Stockholm University's Department of Psychology, who took the time to discuss virtue ethics from a social psychology perspective with me. She seemed baffled and intrigued that these results from the 1960s and 70s had been taken to such extremes in some quarters of moral philosophy. Her nuanced views on the person-situation issue reinforced my suspicion that there was something fishy about that debate as it had taken place in philosophy.

Special thanks to my main supervisor Jonas Olson whose advice, patience, and constructive criticisms over the years have been immensely helpful. Thanks to him the thesis rose again from a dormant state and I was finally able to bring it to completion. And then there's co-supervisor Frans Svensson. Whenever I've produced something, he's read it more or less the same day, sending back detailed feedback and words of encouragement and guidance. An expert on virtue ethics, he never tried to dissuade me from my initial (and too hasty) acceptance of the situationist critique of character traits but instead quietly demonstrated that some individuals just are reliably helpful, honest, and kind. Reflecting on his role in this project I am reminded of the story of Lund University chemistry professor Charlotta Turner whose Yazidi grad student got stuck in a rough spot in ISIS occupied territory of Iraq while he was trying to save his family from genocide.[1] Turner had the university security folks enlist a mercenary commando squad to extract him and get them all back to Sweden and his research. While militarily engaging the world's most blood-spattered terrorist organization on behalf of her grad student was a pretty nice thing to do, it does not quite match the assistance Frans has provided. There are many things that need to align for someone to be able to finish a complicated and protracted enterprise such as a PhD thesis but I know for sure I would not have been able to do it if it weren't for him. Thank you.

Thanks to a fortunate mix of unconditional love and ignorance of academia, both my family by birth and my family by choice have stood by my side over the years. Thank you to my mother Inga-Britt, to my sisters Karin and Marianne, to my mother-in-law Margaretha, and most of all to Lisa.

[1] www.nbcnews.com/news/world/how-swedish-professor-helped-rescue-grad-student-isis-controlled-iraq-n947866

Introduction

"That's an empirical question" uttered in a philosophy seminar usually means the end of discussion. The development of philosophy since its beginning in ancient times has been one of dropping topic after topic, giving birth to new fields as they cluster into coherent wholes with research questions and methods of their own. What is left is a set of "eternal" questions or issues. Some even think philosophical questions by their very nature lack answers, or perhaps that they have answers – only not ones we can ever find. In his introduction to *A History of Western Philosophy*, Bertrand Russell movingly characterized philosophy as a "*No Man's Land"* between science and theology: "Like theology", he wrote, "it consists of speculations on matters as to which definite knowledge has, so far, been unascertainable; but like science, it appeals to human reason rather than to authority, whether that of tradition or that of revelation."[2]

You can think of the separation of science and philosophy as one brought about in response to what kind of *evidence* is thought to be the relevant kind. If the question you pose can be answered by making an observation, or performing a more controlled experiment, we think of that question as belonging to the natural or social sciences. But, as Russell, pointed out, there are some questions which are not answerable just by accumulating more data or observation, and which nonetheless at least *seem* to make a lot of sense and be approachable applying, well, reason. Some of these questions are mathematical – and then there is this motley bag of issues we call philosophical: What is truth? What is knowledge? What has value? What makes acts right or wrong? And so it is that we think of philosophical problems as problems we can ponder just using our intellectual capacities to think clearly, make distinctions, making valid inferences etcetera. Even though philosophical problems may connect to scientific issues, just doing more and more careful science will not directly answer the philosophical problems.

While philosophers these days would typically think that the topics they are considering cannot be answered by careful observation and experimentation, and so is not reducible to some scientific inquiry, for our predecessors, the line was not as sharp. As Kwame Anthony Appiah put it:

[2] Russell, 1946 p. 13.

> You would have had a difficult time explaining to most of the canonical philosophers that *this* part of their work was *echt* philosophy and *that* part of their work was not. Trying to separate out the 'metaphysical' from the 'psychological' elements in this corpus is like trying to peel a raspberry.[3]

I am pointing this out to illustrate how an interest in psychological and other empirical matters is not a new trend in philosophy but rather a return to the classic way of doing things: incorporating domains of knowledge and modes of inquiry to answer whatever questions we find interesting. So although "experimental philosophy" is a recent label to describe attempts to use scientific methods to shed light on or answer philosophical questions, the broader research program as such is not new, and if anything attempts to dissociate philosophy from other types of inquiry is the exception.

This is a thesis on moral philosophy, and there are special reasons why, in this field, there is a long tradition of emphasizing the split between matters empirical and matters moral. The most famous wedge to be driven in between science and ethics was formulated by David Hume in 1739:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surprised to find, that instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but is, however, of the last consequence. For as this *ought*, or *ought not*, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. But as authors do not commonly use this precaution, I shall presume to recommend it to the readers; and am persuaded, that this small attention would subvert all the vulgar systems of morality, and let us see, that the distinction of vice and virtue is not founded merely on the relations of objects, nor is perceived by reason.[4]

[3] Appiah 2008, p. 13.

[4] Hume 1739/40, p. 469, While it is true to say this so-called Hume's law introduces a wedge between the is-talk of science and the ought-talk of ethics, Hume's project was also to incorporate ethics into science, subtitling his book "An attempt to introduce the experimental method of reasoning into moral subjects".

This has become known as Hume's law, which in slogan form says that you cannot derive an ought from an is, or more formally that no moral conclusion can be derived in a deductively valid fashion from a set of non-moral premises. As far as I am aware, no successful counter example to this law has ever been produced, and this thesis is not attempting it either. The significance of Hume's statement was perhaps not obvious to coming generation of thinkers, for it would not take long before some of them indeed wanted to derive an ought from an is. The fact that Hume himself so casually blended normative and descriptive talk did not help either, and it is indeed unclear if what he meant to say in that passage is equivalent to what we now think of as Hume's law.

More than a 100 years after Hume's *Treatise*, Charles Darwin published his *On the Origin of Species by Means of Natural Selection* – and people immediately started to think it had moral implications of all sorts. One idea associated (perhaps somewhat unfairly[5]) with philosopher and intellectual jack of all trades Herbert Spencer says that to ascribe to a behavior or set of motives that they are "good" is equivalent to saying they are "more evolved". Suggestions of this sort, i.e. that there is an identity between the evaluative and the natural, was attacked by G.E. Moore who in his *Principia Ethica* considered it an instance of "The Naturalistic fallacy".[6] According to Moore, "good" referred to a non-natural, irreducibly normative, quality, which could never be identical to natural properties such as "pleasurable" or "more evolved". If what was good and what was more evolved were one and the same thing, asking "This behavior is more evolved, but is it good?" would be just as silly as asking "This behavior is more evolved, but is it more evolved?" Clearly, Moore, remarked, the first question is "open", i.e. it is not obvious what the answer is, whereas the second is not open but ill-posed or has a trivial answer.

There has been much debate around Moore's proposals, both concerning what it really means to say a question is "open" and just what kind of identity – semantic or ontological – is ruled out by his analysis. But the general lesson stuck: stay away from incorporating biology and psychology into philosophical ideas about right and wrong. The separation was further deepened by the logical positivist movement

[5] Weinstein 2019.

[6] Moore 1903, sections 10-14.

which came about around this time.[7] So for a large part of the twentieth century, moral philosophy proceeded in a more *a priori* fashion, trying to steer clear of the empirical fields.

1 The resurgence in empirically oriented philosophy

The terms "experimental philosophy" or "X-phi" refer to a trend or research paradigm emerging within analytic philosophy around the turn of the millennium. If there is a founding father or specific individual in the modern era with which this approach is associated it is certainly Stephen Stich, who through his own work and that done by his many grad students was crucial for the expansion of this line of inquiry within philosophy. In *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*, Stich raised a warning that there may be much more diversity in how various populations think about philosophical problems than had hitherto been assumed.[8] Together with his then grad student Jonathan Weinberg and his former grad student Shaun Nichols, Stich some years later did an empirical study on the intuitions about knowledge ascriptions in so-called Gettier cases, finding that American students and students in East Asia systematically varied in their assessments.[9] In addition to cultural variations in people's intuitions about knowledge ascriptions, one study suggested there is also systematic cultural variation on views discussed in philosophy of language, with Westerners and East Asians thinking differently about references and proper names.[10] Other studies found gender differences in intuitions about philosophical cases, or that personality traits may predict views on free will and responsibility.[11]

[7] Ayer 1936; Stevenson 1944.

[8] Stich 1990.

[9] Weinberg, Nichols & Stich 2001.Though this finding gave rise to fruitful discussions prompting philosophers to reflect on priors or assumptions, it has not stood up well and more recent surveys seem to find no cultural differences in intuitions of this sort. See Machery et al. 2015 and Machery et al. 2017 for these later studies. Gettier cases are cases where a person has a justified true belief and it is still not clear she has knowledge. See Gettier 1963. An enormous literature exists around this. For an overview of the issue, see Ichikawa and Steup 2018.

[10] Machery et al. 2004.

[11] For the gender differences issue, see Buckwalter & Stich 2014. Their findings are criticized in Seyedsayamdost 2015 and Adleberg, Thompson & Nahmias 2015. For personality traits as predictors of judgments on freedom and responsibility, see Feltz & Cokely 2009.

These are all fascinating findings themselves, if only from the point of view of psychology. But for philosophers they seemed unsettling in that they called into question the reliability of our capacity to assess philosophical problems using intuition and considered judgment.[12] Admittedly, we do not really have to do experiments to get us thinking about such problems. Gerry Cohen noted that Oxford philosophers by and large accepted the analytic-synthetic distinction whereas Harvard philosophers by and large did not. Philosophers at both these prestigious universities are clearly as smart and well-informed, so why would it be that there is such a divide structured along this arbitrary line?[13] Roger White noted that he probably holds many of the philosophical beliefs he holds because, more or less by coincidence, he went to a certain grad school with a certain advisor.[14] Had he ended up in another grad school and another advisor he would have held other beliefs, beliefs incompatible with the beliefs he now holds. And those beliefs would have seemed to him just as self-evident and secure as his current ones do. Origin stories like that has spurred a fruitful interest in philosophical methodology and so-called metaphilosophy.[15] They evoke a theme we will have reason to come back to many times, namely the notion that some background stories of how a belief came about seem to *vindicate* the belief in question whereas others seem do *debunk* or *undermine* the belief in question.

2 The rise of the new empirical moral philosophy

At about the time as Stich and his grad students at Rutgers University surveyed ordinary people's linguistic and epistemic intuitions, philosophy grad student Joshua Greene over at neighboring Princeton University came up with a new approach to thinking about the so-called trolley dilemmas. They can be worded somewhat differently, but here are three examples:

[12] In a paper ominously titled "The Rise and Fall of Experimental Philosophy", Antti Kauppinen argued against the value of experimental approaches for moving philosophy forward. His criticism focuses on conceptual analysis and examining so-called folk-psychological conceptions of such things as responsibility, knowledge etcetera. Since this thesis does not examine or rely on that kind of data it will not be relevant to address the concerns he raises.

[13] This observation comes from Cohen's *If You're an Egalitarian, How Come You're So Rich?* but I have only read of it in White, below.

[14] White 2010.

[15] See Williamson 2007 and 2020.

Switch, Bystander or Spur
A runaway trolley is heading towards a group of five people further down the tracks. If nothing is done they will be hit and killed. A bystander positioned at a switch can redirect the trolley onto a sidetrack, where one person will be hit and killed. Would it be morally permissible for the bystander to redirect the trolley in order to save the five?

Push, Footbridge, or Fat Man
A runaway trolley is heading towards a group of five people further down the tracks. If nothing is done they will be hit and killed. A bystander positioned at a footbridge spanning the tracks considers jumping in front of the trolley to stop it and save the lives of the five people on the tracks. He realizes he weighs too little, and so his sacrifice will be of no use. Next to him, however, stands a very large person. By pushing that person off the footbridge, the trolley would come to a stop, although the large person would be killed. Would it be morally permissible for the bystander to push the large person from the footbridge in order to save the five?

Loop
A runaway trolley is heading towards a group of five people further down the tracks. If nothing is done they will be hit and killed. A bystander positioned at a switch can redirect the trolley onto a sidetrack, which loops back to the main track. On this loop stands a large person who is heavy enough to bring the trolley to a stop, although the large person would die as a result. Would it be morally permissible for the bystander to redirect the trolley onto the loop track in order to save the five?

Philippa Foot formulated the first of these dilemmas in a 1967 paper on the ethics of abortion and the doctrine of double effect.[16] Judith Jarvis Thomson later created the *Push* and *Loop* versions,[17] and after that

[16] Foot 1967. In Foot's version, the person acting in the first case was the *driver* of the trolley, not a bystander. I do not think this matters as long as we continue to make the following assumptions: 1) the trolley is out of control due to an *accident*, not negligence or sabotage; 2) if nothing is done the five will be killed, and whatever is necessary to save them demands an *action* of some sort: switching a lever or pushing a person. If we think of the situation as involving a bystander at the switch rather than a driver of the trolley, we may perhaps think of the bystander as more of either interfering or letting the natural chain of events play out, whereas if we think about the problem from the point of view of a driver, it seems more plausible to think of her as responsible for either outcome; there is no default. Personally, I do not think that matters, and in the recent literature the difference seems to have been obfuscated. Even if you are the driver, it remains true that, just as in the bystander case, inaction is an option.

[17] Thomson 1985. Thomson, who came up with the *Footbridge* and *Loop* versions, refers to the individual whose weight is sufficient to stop the trolley as "Fat Man". Most

there have been yet more versions concocted, but these three constitute the basic three ingredients of what has been called 'trolleyology'.[18] Philosophers have used these dilemmas to probe into matters such as whether or not it matters that a death is brought about by action or inaction, if using someone as a means is always wrong, if there is a difference in intending for someone to die versus foreseeing that they will die and so on. The more philosophers thought about these cases, especially if they were of a basically nonutilitarian persuasion, the more complicated matters started to appear. Is there any coherent whole which could provide as justified and make sense of our intuitions about them?

Greene had the suspicion that the explanation people have such a hard time formulating that theory X, has to do the psychological responses the various cases give rise to in us. He suggested that if we want to understand the philosophical disagreement around such cases, and the more theoretical underpinnings which they are expressions of, we need to uncover the psychology behind it all. So he took the trolley problems to a psychology professor and asked if he could help him look into the brains of people while they are processing these moral dilemmas. The result was published in the prestigious journal *Science* and reached an audience far bigger than the little teacup that is academic philosophy.[19] The trolley problems started to become a staple of not just many academic disciplines – philosophy, psychology, neuroscience, behavioral economics, cultural anthropology – but also of popular culture.

There has since been an explosion of interest in empirical moral psychology, often with a (tedious) focus on the trolley problems. This interest has also coincided with, and gotten strength from, a contemporaneous celebration of philosophy applied to the social and political

writers today refer to the person instead as "large", "heavy", a "bodybuilder" or someone wearing a heavy backpack. Philosophers are so habituated to pondering thought experiments of this sort, they all assume nothing else besides the two options described is possible to do. They also assume the question is not what to do given *uncertainty* about what actions plausibly lead to which outcomes. It is unclear if people surveyed in various polls and experiments also consistently make these assumptions. People do have a tendency to avoid the problem by thinking there is third solution or inserting doubts about the plausibility of the causal claims the dilemmas presuppose. These methodological worries are discussed in Ahlenius & Tännsjö 2012.

[18] The term 'trolleyology' (in this sense) originates with Appiah 2008, p. 89. For a comprehensive and very readable history of the trolley dilemmas, see Edmunds 2014.

[19] Greene et al. 2001.

domains, since it has been suggested that something like figuring out solutions to the trolley dilemmas is involved in programming the behavior of self-driving cars. Should the car act so as to maximize survival chances of its occupants regardless of the costs to others? If the only way the car can avoid crashing into a kindergarten is to crash in to a pedestrian, should it? This meant the trolley dilemmas went from being a set of obscure thought experiments in academic philosophy to something that seemed to be a key factor in an emerging technological and logistical development potentially affecting billions.[20] And moral philosophy was at the center of it all.

Many additional factors together help explain the increased interest in empirical approaches to moral philosophy. Greene was lucky that the *fMRI* technique was just being developed and that Princeton housed one of the first used for non-medical purposes. Hume would certainly have wanted such a machine, but it was not available to him. The time was simply right. Another factor is that the world's most famous philosopher, Peter Singer, wrote favorably about Greene's findings and conclusions, making them well-known to a wide audience both within and outside of academic philosophy.[21]

In December of 2002, Daniel Kahneman (together with Vernon Smith) received the Nobel Memorial Prize in Economic Sciences, for "for having integrated insights from psychological research into economic science".[22] So it is fair to say there was at the time a general interest in recruiting insights from psychology to other neighboring fields, and the pictures emerging of human psychology both from Kahneman's and Greene's research are quite similar, and Greene explicitly builds on insights from Kahneman (and his deceased collaborator Amos Tversky). Both behavioral economics and experimental moral philosophy are research programs which seek to incorporate an empirically adequate, as opposed to an assumed or idealized, view of human psychology into their respective bodies of research.

Greene's work is unquestionably the main cause the trolley dilemmas escaped the academic discussions of moral philosophers and made their appearance in psychology, comparative anthropology, discussions about the ethics of self-driving cars, magazines and even tv

[20] See Nyholm & Smids 2016, and Kauppinen forthcoming. For a large survey of people's ideas about how such vehicles should be programmed, see Awad et al. 2018.

[21] Singer 2005 and, for a wider audience in a syndicated column, Singer 2007.

[22] www.nobelprize.org/prizes/economic-sciences/2002/kahneman/facts/

series.[23] After Greene's experiments, the trolley dilemmas became the hub of empirically oriented ethics, and in many ways the public face of ethics.[24] Being such thankful ways of introducing the longstanding conflict between utilitarian and deontological ways of addressing moral problems, they have proved irresistible tools in both teaching and in the empirical study of moral judgment. To an extent I think their popularity makes moral philosophy seem silly and out of touch with real-world problems, not to mention real world psychology. For all their merits, it would be absurd and impoverished to believe we can come to understand and assess all of moral thinking by exclusively dwelling on these contrived scenarios. Having said that, I am methodologically promiscuous and welcome confronting moral theories with any kind of evidence we may think is of relevance in making up our minds about their plausibility. Real cases, imagined cases, practically impossible but logically possible cases, consistency with other views we believe we have reason to believe etcetera – all of these facets are legitimate checkpoints when doing moral philosophy. To my awareness, there is pretty widespread agreement on this stance actually, but it is worth mentioning to anyone who has peeked in on ethics or experimental philosophy and walked away with the impression we are *only* thinking about weird thought experiments. We shouldn't, and we aren't.

3 The plan of this thesis

As you can see, there is at present a return to psychological and biological issues bordering on moral philosophy. This renewed interest is often not in the business of anchoring morality in, or somehow deriving it from, biology, but is rather employed in a skeptical or destructive enterprise: to use findings and ideas from evolutionary theory or psychology to one way or the other *undermine* various views in moral philosophy.[25] This way of employing findings in psychology and evolutionary theory will be a recurring theme throughout this book.

The thesis is a monograph, but perhaps a somewhat eclectic one. The various chapters are bound together by un overarching question:

[23] I am thinking of *The Good Place* and *Orange Is the New Black*, as discussed in Elizabeth Yuko's 2017 *Atlantic* piece.

[24] See for instance Davis 2015.

[25] In works of a more popular kind, not written by academic philosophers, one can see subtitles as "How Science Can Determine Human Values" (to Sam Harris' *The Moral Landscape*.

what can the philosophical field of ethical theory learn from absorbing what is being done in evolutionary, cognitive, social, and developmental psychology, where morality is studied not primarily from the point of view of right and wrong but as a way of coming to understand its role in human life? But the thesis is less eclectic than first impressions perhaps convey. Although spanning many seemingly disparate issues in moral philosophy – the ongoing discussion between consequentialists and deontologists, the possibility of justifying some ethical claims by giving them status of self-evident truths, the plausibility of thinking of ethics as centered around notions of virtue and vice etcetera – there are common links tying all of these debates together, namely the psychobiological underpinnings of our moral emotions, thoughts and behaviors. Claims of innateness are implied both in the debates covered in chapters 2 and 3, on the philosophical relevance of neuroscientific and evolutionary debunking approaches to moral judgments respectively. And the seemingly separate debate, addressed in chapter 4, on the challenge from social psychology to virtue ethics, is likewise linked, or that is my contention at least, to the psycho-evolutionary findings and theories which hold center stage in the innateness chapter.

4 Moral innateness

Chapter 1 examines the basis of our capacity for moral cognition. Is thinking in moral terms something humans do sort of as a spin-off effect of having language and living together, or is it rather "hard-wired"? This inquiry takes me into evolutionary psychology and biological anthropology. There are many options in this debate, with various grand psycho-evolutionary models about the nature of moral judgments and the moral emotions or, for short, morality. I will not take a definitive stance among the options, but will present and defend enough to make it likely that moral innateness in a specific enough sense is plausible. Though here and there in the chapter I reveal my adaptationist inclinations, I believe what I ultimately try to establish is not too outlandish and may be arrived at from various starting points, namely the idea that evolution has provided us with a set of emotional responses which lead us to moralize in some ways rather than others. This psycho-evolutionary background, and the more particular conclusion about emotions, will then be of relevance for the ensuing two chapters, which link moral judgment and cognition to the debate in

ethical theory, in particular the conflict between consequentialist and deontologist normative theories.

5 What pushes our moral buttons? The neuroscience of moral judgment

In chapter 2, I describe the immensely influential research by Joshua Greene, the philosopher who borrowed a brain scanner to see what happens inside our heads when we engage in moral problem-solving. Greene and others then invoked these findings in the debate over which is the more plausible ethical position, deontology or utilitarianism. As we will see it is not by trying to bypass Hume, but by using these empirical data in a more indirect way, that he aims to undermine intuitive support for deontology and remove some of the resistance to utilitarianism.

Greene draws from a wealth of evidence – brain imaging, research on heuristics and biases, evolutionary theory, surveys, anthropology, forensic psychiatry and other kinds of data – to defend a theory he calls the dual process theory of moral judgment. The dual process theory, roughly, says that there are two kinds of mental processes responsible for producing moral judgments. One component occurs quick and automatic and the other is more time-consuming and deliberate. The two forces, as it were, fight it out within us, leading us to sometimes accept verdicts produced by an unreflective automatic process, and other times to side with the more cognitively demanding verdict of the second type of processing. This empirical theory of moral psychology, in turn, Greene believes, may be invoked to undermine some of the intuitive support for deontological approaches to ethics while leaving utilitarian approaches relatively unharmed. The last fifteen or so years have seen an intense discussion, both from the point of view of science and moral philosophy, accepting, assessing and rebutting this challenge. In this chapter I look at the dual process view and its relevance for moral philosophy. While being largely sympathetic to Greene's claims, defending him against some of the criticism formulated by Selim Berker, I argue that a certain degree of restraint is called for.

6 Bracketing our evolved psychology

Bringing chapters 1 and 2 together to discuss the implications of an evolved tendency to moralize in a certain way and the ongoing debate over utilitarianism and its contenders, I turn in chapter 3 to a proposal by Katarzyna de Lazari-Radek and Peter Singer on the lines that we

can get at ethical truths using reason to counteract some of our evolved biases. Lazari-Radek and Singer fruitfully try to merge two projects: that of answering the evolutionary debunking of moral beliefs made famous by e.g. Sharon Street, on the one hand, and, on the other, that of adjudicating between what they call the demands of ethics and the demands of rational egoism. Answering this latter challenge in favor of the ethical primacy is also an answer to the evolutionary challenge more generally, they claim. And not only that, the special way, inspired by Henry Sidgwick, that this conflict is resolved also shows, they claim, that the kind of general evolutionary debunking of ethical beliefs made popular by Street and others does not target the kind of impartial ethical beliefs undergirding utilitarianism but only a subset of ethical beliefs more conducive to deontological modes of thinking in ethics. As much as I would want this to be successful, my assessment is largely negative. My skepticism towards this project mainly stem from two sources. One source has to do with their reliance on an intuitionist moral epistemology in contrast to a more coherentist one, which makes more explicit use of a reflective equilibrium style of justification, as opposed to a foundationalist. The others source of skepticism towards their specific project has to do with the difficulties of insulating one kind of ethical judgments (those favoring utilitarianism) from other kinds of ethical judgments. Once we put on the skeptic glasses of evolutionary debunking, it is hard avoid its corrosive effects on *all* ethical judgments, deontological or otherwise. The current discussion on evolutionary debunking strategies is a flourishing field, with most contributions approaching the matter from a metaethical point of view. Since Lazari-Radek and Singer explicitly invoke evolutionary debunking (and its remedy) to support a specific normative theory, viz. utilitarianism, I focus in this chapter on that form of response to the debunking challenge.

7 Lack of character?

In chapter 4, I turn to a different way of using psychological research to influence a debate in moral philosophy: the attack on virtue ethics using social psychology. According to this critique, which has been voiced notably by John Doris and Gilbert Harman, the behavior of human beings is simply not the result of them having different *character traits*. Instead, according to the challenge at hand, the dynamics of the social situation, including many features we are not even aware of, explain what makes us do what we do.

Though I agree that some of the defenses that virtue ethicists have offered could have been better, I also try to help them, by offering the evolutionary-biological account of human behavior presented in earlier chapters. What we know of human psychology anchored in such an understanding speaks against thinking of us as autumn leaves flown around by the wind. And a more updated reading of the scientific evidence bears this out too. Still, one may worry that my psychobiological attempts to save virtue ethics from situationist social psychology was just taking it out of the frying pan and into the fire. I believe that the situationist challenge, if correct, would have been worse for virtue ethics and that there remains good hope that it can be wedded to a modern and accurate psychology. The overall conclusions is that, because virtue ethics is the normative ethical view most imbued with assumptions about human psychology, it is particularly vulnerable to what the empirical science of psychology actually warrants.

8 Why this?

There are many issues in philosophy where an experimental angle has been applied to shed light on entrenched debates. In this thesis, I focus on empirical work with a direct bearing on debates in ethical theory. I wanted to create an invigorating collision between such work and the basic contenders in normative ethics – consequentialism, deontology, and virtue ethics – to examine if these debates could in any way be resolved or at least be advanced as a result. There has been other experimental work done on matters of relevance for ethics, such as freedom of the will and moral responsibility, or the links between accepting a moral view and being disposed to act on it. That kind of research is of course highly interesting but does not directly challenge any of these basic moral outlooks.

1 Is Morality Innate?

In a study published in *Nature* 2003, Capuchin monkeys were trained to return from their cage a token and receive a piece of cucumber as reward. Fans of cucumber, the monkey kept returning the tokens again and again. When a conspecific in a neighboring cage all of a sudden received a grape (which is sweeter and more highly valued) for the same task, the recipient of cucumber would demonstratively throw away the cucumber and refuse to continue the exercise. The tendency not to accept the cucumber was even stronger when the neighbor received the grape without returning the token (i.e. getting the reward without the effort).[26]

So, monkeys do not like it when they are rewarded comparatively less. When presenting a video of the spurned monkey's refusal to play along, de Waal quipped that the sequence was "basically the [occupy] Wall street protest" (i.e. reaction to perceived injustice).[27] Such captivating and human-like behavior in closely related non-human animals inspire reflection on the psycho-emotional bases and evolutionary origins of the more advanced but related moral psychology that we see in humans. Do these observations support, as the authors of the study suggest, "an early evolutionary origin of inequity aversion"?[28]

Is there, then, a biological basis for our proclivity to evaluate the behavior of others and ourselves in moral terms? In a sense, the answer is trivially yes. We already knew that, say, pet lizards sharing much of the environment of human children do not come to moralize the behavior of self and others the way their human mini masters come to do. Something about human and lizard biology accounts for that difference. Granted, lizards are very alien creatures. What about closer relatives, such as other primates? As we saw above, some researchers talk about the building blocks of moral cognition being present there.[29] And if, in some sense yet to be specified, moral capacity is innate, finding its root among our closest relatives would be an important piece of

[26] Brosnan & de Waal 2003.

[27] https://youtu.be/meiU6TxysCg.

[28] Brosnan & de Waal 2003, p. 297.

[29] Frans de Waal has famously argued for the evolutionary continuities of human morality and its roots in primate psychology. In addition to the study on Capuchin monkeys and unequal pay, see Procter et al. 2013. See also Henrich & Silk 2013, Heaney, Russell, and Taylor 2017, and Brosnan & de Waal 2014.

evidence. This chapter examines various interpretations of the claim that morality is "innate", and defends what has been called a strong or even "immodest" version of innateness.[30] I do this by looking at work from developmental, evolutionary, and cognitive psychology, as well as ethology. I present the contours of two current grand theories of moral innateness, viz. Jonathan Haidt's "Moral foundations" and John Mikhail's "Universal grammar", arguing that all these various sources give us plausible grounds for accepting the immodest view.

Why would philosophers be concerned with the possible biological bases of morality? Well for one, there is simply the attraction of Wilfred Sellars' dictum that philosophy is about coming to "understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term".[31] Ethics – understood as the traditional fields of metaethics, normative theory, and applied ethics – is of course an integral part of philosophy, but following Sellars it is also the business of philosophy to come to understand how this field hang together with all the phenomena (psychological, biological, linguistic etcetera) surrounding the traditionally "pure" domains of philosophical inquiry. Another reason is that philosophers can contribute to the advancement of this research by providing their expertise, both about moral philosophy and a general competence in conceptual analysis, clarity and rigor of reasoning. Additionally, we want to find out if there are any *implications* for moral philosophy depending on the answer to the innateness question. Maybe there are no interesting relations between moral philosophy and the issue of the proposed biological capacity to think and feel in moral terms. Or perhaps, quite to the contrary, these findings, depending on what they are more specifically, could serve to vindicate, refute, undermine or support a given theory or view in normative or meta ethics. Such ancillary questions will be addressed in later chapters. In this chapter I will describe the varying ideas that may be called moral innateness, and will try to make it plausible that there is indeed a sense in which moral innateness is true.

The issue of the biological underpinnings (or lack thereof) of morality is part of a larger discussion about the makeup of human psychology – an exploration of human nature if you will. Although it is a given that we approach this topic with the understanding that our psychology, like our organs and bodily functions, is the product of millions of

[30] Prinz 2009, p. 168. See more below.

[31] Sellars 1962, p. 35.

years of evolution, it is a contested question just how finely chiseled the mechanisms of our psychology and behavioral repertoire are. The research program called Evolutionary Psychology (an offshoot of sociobiology) is usually taken to be committed to the idea of *massive modularity*, i.e. the notion that our psyches are made up of a multitude of specialized, genetically grounded algorithms or computational mechanisms that evolved to solve certain problems faced by our ancestors.[32] One such module, or set of modules, might very well be a moral appraisal system, and indeed evolutionary psychologist typically are moral nativists.[33] "Module" has a physical, spatial ring to it. But the expression does not refer to neural anatomy, but to *functional specialization*. Alternatively, one takes an "empiricist" view of the mind, postulating fewer and more broadly competent problem solving faculties that need much empirical input. Empiricists about the mind typically deny moral innateness (and indeed the innateness of many other psychological or behavioral traits).[34] These two competing accounts will be the recurring leitmotif of this chapter.

From an evolutionary point of view, there are two plausible major pathways taking us to what we now think of as morality: sympathy and cooperation.[35] Sympathy is a psychological adaptation instilled in our lineage to solve the challenge of prolonged care of offspring. Human children, compared to all other animals, are uniquely helpless and slow to develop. Taking care of them is tolling and could not be done by someone who is not wired to feel strongly for their wellbeing. So, while all mammals need mechanisms that make parents tend to their offspring, the parental care seen in humans is beyond anything else known on this planet. The other key component is not about care and sympathy, but about how we as intensely social creatures are dependent on others with whom we do *not* always have a psychological bond of bigheartedness. Our species' great dependence on cooperation is thus the second fundamental factor undergirding our moral psychology.

[32] Downes 2014.

[33] I will use 'innateness' and 'nativism' as synonyms.

[34] See Prinz 2012 for a compelling book-length empiricist case. More on Prinz later.

[35] For the neurobiology of bonding and mammalian sympathy, see Churchland 2011 and 2019. For examples of accounts emphasizing cooperation, see Curry 2016 as well as Curry, Mullins, and Whitehouse 2019.

Humans have had the same brain size and mental capacities for many, many tens of thousands of years. We started to use these resources to pursue science and other intellectual enterprises only recently. They were not developed to solve those kinds of things in our evolutionary past. The most plausible account is that we owe our outstanding cognitive capacities largely to reap the benefits of cooperation. The brain circuity and mental capacities needed for language in turn are part of this account. Language is a social phenomenon and likely arose in large part because it helped our ancestors coordinate action as well as to disseminate information on things like the trustworthiness or reproductive status of ourselves and others.[36] There is thus a crucial interdependence between these three very important aspects of our species: language, cooperation, and morality.

1 The meaning of innateness

We have already seen that saying morality has a biological basis is not saying much since everything we do does. A more restricted query is asking if *morality* is *innate*. Alas, these terms are imprecise and used by different thinkers in different ways. Trying to get a clearer grasp of the central terms is thus a good start. The first commonsensical suggestion is that "innate" connotes a degree of hardwiring and means something like "genetic in origin and robustly independent of environmental influence". On this view, a person's blindness may be innate: say she lacked the genes that encoded for fetal development of the optic nerves; and, of course, no amount of training or other form of environmental influence would make her see. Though blindness may be genetic in origin for a given individual, stereovision in the human species is an adaptation.[37] In psychiatry, there is a discussion on how to account for conditions such as autism spectrum disorders. People used to believe autism was caused by a certain kind of parenting style, but it has become more common to now believe it is genetic in origin.[38] The upshot, then, is that autism might be *innate*. Though such a usage of the term is perfectly legitimate, I think this implication of what I labeled the commonsensical view makes apparent that we really want to ask a more general question of the human species, and not just about

[36] Berwick, Chomsky et al 2013; Dunbar 1996; Jones 2016; Al-Ubaydli, Jones, and Weel 2013; Proto, Rustichini, and Sofianos 2019.

[37] For further discussion and other uses of "innate", see Mameli & Bateson 2011.

[38] See Sandin 2014 and Malik et al. 2019.

the biology of some individual.[39] Given this wider focus, it is more useful to think of innateness in the present context as a shared species-typical psychological architecture, i.e. the traits we have in virtue of being humans. To get a grip on what kind of thing that may be, it helps to think of such qualities in terms of *biological adaptations*.

An adaptation is a trait or a feature of an organism, which has been developed under the pressure of natural selection. Erections and ears are adaptations, as are teeth and opposable thumbs. Capacity to read and write are not adaptations but we recruit cognitive machinery which are in order to fulfill such evolutionary recent tasks. Asking whether or not morality is innate, then, is asking whether or not morality is a trait whose emergence in our lineage is to be explained by the reproductive success it conferred on our ancestors over those lacking that trait in their surroundings, or whether it is to be explained more along the lines we would explain our capacity to read and write. These latter capacities are features of our psychology, but we do not typically think they themselves have been the target of selection pressures. That may be because they are too recent to have had an influence on our genome, or because they cannot be meaningfully separated from a larger bundle of capacities of which they are a part, or constitute a non-adaptive but neutral side effect of adaptive capacities. These latter kind of phenomena are called, following Stephen Jay Gould and Richard Lewontin, *spandrels*, a term they borrowed from architecture.[40] There, a spandrel refers to the approximately triangular area created between two arches or between an arch and the wall, for instance in a church. This surface area can be used for decoration or to write a message, but that is not why they exist. They exist because they are a side-effects of that particular way of erecting a high building. Moral nativists think our moral capacity is more like vision and less like reading and writing; more like the arches and less like the spandrels they accidentally give rise to.

[39] I side with Joyce 2006 in thinking of "innate" from the point of view of its role in evolutionary history, rather than disentangling at the level of a specific individual whether or not a given trait has a biological or environmental root. Mogensen 2014 points to usages of "innate" that do not involve adaptation, as when we say that a given dysfunction in an individual is innate. This usage is perfectly acceptable, and authors need to be explicit about what they stipulate central terms to mean. In this work, "innate" refers to a species-wide adaptive trait.

[40] Gould and Lewontin 1979.

2 *Innateness in non-moral domains*

By measuring looking time and other signs of puzzlement or surprise, contrasted with habituation, psychologist have tried to operationalize the extent to which human infants' perception and mental processing of the world come equipped with certain innate ideas or structuring principles. There seems to be a number of constraints or set of regulatory expectations that are innate, and thus "known" by infants, for instance that solid bodies cannot pass through one another, that a moving physical object cannot cause another physical object to move unless they come into contact, or that objects move along continuous trajectories and cannot disappear and rematerialize. Using limited visual cues of partly hidden and partly observable movements, infants will form expectations as to whether or not what they have seen is the movement of one and the same or two different objects. Infants also distinguish between objects and animate agents. The common feature is that these expectations are not conclusions drawn from experience and inductive reasoning, but are present prior to any experience or reasoning.[41]

Babies also understand the bodily movements of other humans in a special way. In one study, an infant would be observing a person reaching for either a teddy bear or a ball placed next to one another. Next, an experimenter switched the placing of the objects. When the first person once again reached, the infant expected them to reach for the same object, not the same place. When a rod or metallic claw did the reaching, this expectation was absent.[42] This suggests infants have a competence which allows them to infer *intentions* of other people, while realizing non-animate objects, while moving, do not have intentions.

With the example of understanding other people's intentions we've moved into the realm of interpersonal relations, i.e. the social world. Navigating this territory is likewise premised on the existence of specialized problem-solving capacities of our minds. Perhaps the most intensely studied candidate is *cheater detection*. Cheating occurs when an agent takes the benefit of a social exchange but does not satisfy the requirement that the benefit was premised on.

The suggestion that we have a specialized cheater-detecting mechanism grew out of comparisons of our reasoning abilities concerning conditionals in general as contrasted to social exchange conditionals in

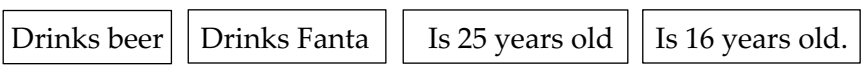
[41] See Baillargeon 1987, Spelke 1990, and Spelke & Kinzler 2007. A very readable overview of some of this research is given in Bloom 2013.

[42] Woodward 1998.

particular. Identifying a cheater might be thought of as applying a simple conditional scheme: If *P* (you pay the price), then *Q* (you can have the goods). Psychologist Peter Wason wanted to see how well people perform at falsifying conditional hypotheses. He showed participants the following cards with either a letter or a number on the side facing up:

Each card has a letter on one side and a number on the other. He then asked them to check the veracity of the rule "If a card has a *D* on one side, it has a *3* on the other" (a pretty basic *P* implies *Q* statement) by turning the card(s) that need to be turned, and only that/them. Which card(s) do you turn? Most subjects chose either the *D* card or the *D* card and the *3* card. The correct answer is the *D* card and the *7* card. The *3* card and the F card are irrelevant for the conditional rule and so do not need to be turned, but people are very bad at picking out the *7* card as a possible violator of the rule.[43]

Now let us look at a structurally similar check on a conditional rule. This time the rule is "If someone is drinking beer that person needs to be 18 or older." Which of the following cards (representing bar guests) do you need to turn to check if the rule has been violated:



Here most people effortlessly come up with the right solution: we need to know how old the beer drinker is and what the 16-year-old is drinking. It does not matter what the 25-year-old is drinking and we do not need to know how old the Fanta drinker is.

The notion that there are specialized psychological appraisal systems for detecting cheating in social exchanges also gains support from studies on patients presenting with positive schizophrenic symptoms. Such patients will have deficits in their general reasoning abilities, but exhibit intact capacities when the content is switched to social exchange.[44] The upshot is that though we are dealing with a pretty simple piece of conditional reasoning, we do not approach the problems

[43] The first study by Wason 1968 was about even numbers and colors of cards. The contrast to reasoning about cheating was introduced in Cosmides & Tooby 1992.

[44] Kornreich et al 2017. For an overview, see Cosmides & Tooby 2015.

using a domain-general reasoning ability. Instead, depending on whether or not there is a content of possible social exchange gone sour, we are able to pick out the suspects. And, importantly, cheater detection, not general logical thinking, solves it for us. The same kind of tasks have been tinkered with to control for possible confounding factors such as familiarity with the situation described, readability etcetera. The pattern remains the same: we are expert cheater detectors and amateur logicians. Findings like these support the notion that at least part of our capacity to identify and assess social interactions are due to specialized mental faculties rather than a consequence of general reasoning-abilities. More recent research has corroborated and expanded on these cognitive adaptations for monitoring social exchange and neighboring areas. Again, people exhibit an elevated capacity, compared to descriptive controls, to identify violations of socio-moral codes pertaining to helping behavior, maintaining coalitions, and submitting to authority. [45]

3 Morality – content and capacity

When someone suggests morality is or is not innate, just what is involved? In addition to simply denying any such innateness, there then seems to be three possibilities, with an increasingly "full" pre-equipped capacity claimed:

1) *Non-nativism:* Just as human beings do not possess a dedicated set of proclivities and competencies to talk about staplers or chocolate bars, there is no innate competence or proclivity to make moral judgments, and the domains we moralize, and the content of our moral judgments, are entirely up to the surrounding environment.
2) *Weak nativism*: Human beings possess a dedicated set of proclivities and competencies to moralize, but the domains we moralize, and the content of our moral judgments, are up to the surrounding environment.
3) *Strong nativism:* Human beings do not only possess a dedicated set of proclivities and competencies to moralize, we also have an innate tendency to make and accept moral judgment with a certain content, i.e. the content is strongly constrained.[46]

[45] See Sivan, Curry, and van Lissa 2018.

[46] Following Prinz 2014, p 105.

To ask if morality is innate, then, is to ask if a) we have an adaptive capacity to make moral judgments of a rather open-ended character or b) a capacity to make moral judgments with a strongly constrained or directed content or none of these. Of course empiricism in this debate does not deny the obvious, namely that we do make moral judgments. It just claims this is something learned and made possible by general capacities that we have and which did not evolve to make moral judgments.

Since the strong version makes additional claims compared to the weak versions, i.e. that we are wired not only to moralize but to moralize a certain way, it would appear that acceptance of the weak version is easier to justify as compared to the strong version. Jesse Prinz even used to call the strong version "immodest", suggesting only a fanciful assessment of the available evidence can lead anyone to accept it.[47] But as a matter of evolutionary chronology it is hard to understand how weak or capacity nativism, which is here taken as the more modest claim, might have evolved sans a pre-linguistic content, a domain our ancestors tended to moralize. And, importantly, a domain over which we moralize could not make all kinds of moral judgments made within it equally adaptive. A domain matters because it matters what we believe and do within that domain. There is good reason to believe we would not be a moralizing species if we were not first a species with strong emotional reactions to, e.g., social interactions. Therefore, I think the notion that morality could be innate in the sense that we make moral judgments and it is then an open matter what kinds of moral judgment we would be making is *prima facie* implausible.

Long before anyone on this planet had ever made a moral judgment, our ancestors first evolved a set of emotional and behavioral propensities in response to various challenges, opportunities and threats. These propensities may cluster into various domains, such as cooperation, care or deference. Only later emerged a linguistic competence to moralize what goes on in these domains. But the content came before the capacity. We come equipped with a set of innate biases that make some moral judgments seem more attractive than others. If this were not the case it would be equally easy to train children to come to accept the norm that people who help others deserve punishment, as it would be to train children to accept the norm that people who harm

[47] Prinz 2009, p. 168.

others deserve punishment.[48] There would be no natural psychological pull towards the norm that helpers deserve praise and harmers blame, and presumably no tendency to *like* those who offered help more than those who caused harm. The acceptance of some moral judgments over others is clearly adaptive. Individuals who were prone to making moral judgments of the sort "It is okay to kill your babies" were less likely to thrive and reproduce than beings disposed to accept judgments like "It is good to return favors and seek trustworthy partners". Admittedly, in order to ground the specifically adaptationist claim about moral judgments, it is not enough to show that moral judgments are shaped by our emotional and behavioral dispositions. One also needs the additional claim that making moral judgments influences action and attitudes in ways that go beyond what the dispositions alone do. I will address two ways of linking moral thinking to behavior in section 5, viz. the idea that morality functions as a commitment device and as social signal.

4 The emotional bases of moral judgment

Among empirically oriented researchers of moral psychology, a growing consensus has emerged over the last fifteen to twenty years or so concerning the profound role of emotions in moral cognition. The evidence comes from functional imaging of the brain, research on psychopathy and Autism spectrum disorders, on patients with partial brain damages, from emotional priming induced under hypnosis, the influence of disgust on moral judgments, motivated reasoning, and many more.[49] Of course, the idea that emotions play an important role for moral judgment is much older than that. What has happened, though, is that this view has been supplemented with a wealth of evidence not available to thinkers like David Hume.[50]

[48] In personal communication, developmental psychologist Kiley Hamlin recalled of such an attempted study that "we essentially tried to have an experimenter praise harmers and boo at helpers – we didn't get very far because kids were super distracted by the experimenter and didn't seem to process the show at all." But see Van de Vondervoort & Hamlin 2019, and DesChamps, Eason, and Sommerville 2016 for study designs attempting to answer these questions.

[49] For some overviews, see Prinz 2009, Greene 2013, Sinnott-Armstrong (ed) 2008a and 2008c. Criticism of some of the proposed links between disgust and moral judgment can be found in Ghelfi et al. 2020.

[50] These empirical data concern the causal processes that give rise to moral judgments and are compatible with cognitivist ideas on the semantics of moral judgment. See Joyce 2008 for further discussion.

In the preceding section, I argued that content must have come before capacity, partly because our moral judgments have something to do with what it is adaptive to approve and disapprove of. To see why this is so, consider the following account from a non-human context, again by primatologist Frans de Waal. The chimpanzees Puist and Luit had a longstanding habit of helping each other. One day, Puist was attacked by a third individual, Nikkie. Puist turned to Luit in search of support, but Luit did nothing to avert the attack. Subsequently, Puist furiously attacked not Nikkie, but Luit, who had failed to reciprocate previously offered assistance. If you believe that fury in response to the defection of a coalition partner is more adaptive than delight, you should believe some contents of our moral judgments are more likely than others.

Both human and nonhuman animals experience emotions of a variety of sorts, propelling or inhibiting them as appropriate towards adaptive behavior. They need to drink and eat, avoid trauma and contaminated food sources, find a mate to reproduce with, and stay clear of predators and many other dangers. That is why they experience emotions such as hunger, thirst, disgust, sexual desire, and fear. The creatures humans evolved from were capable of experiencing emotions long before they were human, and indeed long before they were even social animals. Some of the emotions that evolved prior to anything we can call morality also regulate social behavior, and these emotions were the coopted building blocks upon which morality later arose.

How did it arise? Recall the indignant Capuchin monkey. She was perfectly happy working for cucumbers – until it dawned upon her someone else got grapes for the same job. Experiencing and displaying anger at unfavorable distributions is an adaptive mechanism for avoiding exploitation. The strong negative reaction makes it more likely that the individual Capuchin monkey will not stay in or accept bad deals, and the reaction, what Brosnan and de Waal call an "inequity aversion", is also a signal to others the she will not. It is important that the monkey not only experiences something like offense, but she *displays* it to others as well. Moral emotions serve to regulate behavior, that of the individual as well as that of others. And the same is true of the story of the three chimps of course: Puist's anger at and ensuing attack on the traitor send a clear message that failures to reciprocate will not be tolerated.

5 *Morality as commitment device and social signal*

Why might a moral sense, i.e. a capacity to moralize, be an adaptive feature, something that would be enhancing the reproductive success of our ancestors? Is it not enough that we have, as the capuchin monkeys and the chimps, pro-social and punitive emotions? Why would this thing moralizing enter the stage?

Humans as physical creatures are not very impressive. Not the fastest, strongest, smallest or largest of animals. But we have our brainpower and our spectacular capacity to cooperate and plan ahead. These have been the keys to our success. We have already seen how cheater detection is an inbuilt feature of our psyches. Why would that be? Because we rely on cooperation to survive. Hence, being trustworthy, and being known to be so, and being able to assess the trustworthiness of others, are important traits in an organism whose success is tied up with the joint efforts of others. But sometimes you just do not feel like chipping in. There is always temptation to forfeit the cost part and still reap the benefits produced by the generosity and diligence of others. But others are quite remarkable at spotting cheaters and will quickly terminate all reciprocal interactions with you if you cheat. And that is not good for you. Cooperating is not as good as cheating undetected, parasitizing on others, but definitely better than ostracism. So humans (and other social animals) are locked in a sort of arms race, stag hunt dilemma type of scenario where we need to constantly monitor our incentive to cooperate, the tendency of others to cheat if they can get away with it, and so forth.[51]

Morality would have been a successful strategy to cope with that kind of situation. Adopting certain moral convictions removes options. I have never been tempted to break in to my neighbors and steal from them when they are away. The thought has never crossed my mind. I am pretty sure I could get away with it, but I am just not into that sort of thing. This way moral convictions shape behavior, by either pushing us to do the things we think we should but lack immediate motivation to do, or making us refrain from doing what is tempting

[51] A stag hunt dilemma is a situation where an individual can choose to reap a small benefit from solitary action (killing a hare) or a larger benefit (killing a stag) requiring the cooperation of others, without being able to monitor or influence whether or not others will actually participate in the cooperation; often ends up with all individually going for the lesser hare reward. The expression originates with Rousseau, and is given a comprehensive treatment in Skyrms 2003.

but not what we should, or, as in the breaking-in case, simply by removing options from conscious consideration. In many cases, moral convictions may be unnecessary, as the behavior may be unattractive or unattractive in any event. But for other instances, adopting a set of moral convictions will help the individual with providing a behavioral auto-pilot removing some options and considering or suggesting others. That way, the sense that you are morally accountable provides a powerful motivational corset in the marshmallow test that is life.

Acceptance of norms also signals to others that you are committed to a certain course of action, and can reliably be predicted to act in certain ways. That is a valuable asset. You can be trusted, not just to do what is beneficial for you, but what you think you should do regardless of that, as well as to oversee that the people you interact with do too. If you are the kind of person who would not go back to a market vendor because the raspberries they sold you had molds on them, you are *more* likely to receive raspberries with molds on them than someone who *is* the kind of person who would go back and point that out and is *known* to be such a person. But raspberries do not cost much, is it worth the hassle? It is the *principle*, they shouldn't be fooling people like that.[52]

How does morality succeed in doing all of this motivation-boosting and interpersonal policing? The trick is the apperception that morality is calling us from outside of ourselves, providing an external nudge or non-conditional instruction. This takes us to the question what it is, more precisely, to, as I have been calling it, "moralize", or "make moral judgments". Here I think it is wise to give a rather general characterization, one that does not at this stage tilt or exclude options further down the road when we want to assess the philosophical implications of the possibility of moral innateness. For that reason, I am reluctant to say, for instance, that moral judgments are beliefs, and that utterances of moral sentences are assertions, i.e. expressing something either true or false. Nor that they are *mere* expressions of a favorable or a disapproving attitude. What is important is that moral judgments are taken to have an authority or validity which is independent of the wants of the speaker, i.e. they are categorical. Moral judgments come with *oomph*, i.e. a perceived inescapable normative clout, as Richard Joyce named the phenomenon.[53] He provides the following checklist, stating

[52] See Frank 1988 for more on this line of reasoning.

[53] Joyce 2006, pp. 62-4; 199-209. See also Olson 2014. Perceived or not, one might think that moral judgments *are* inescapable, i.e. binding whether the agent recognizes

we would expect to tick off at least some of them before we call some human activity making moral judgments:

> Moral judgments centrally govern interpersonal relations; they seem designed to combat rampant individualism in particular.
>
> Moral judgments are often ways of expressing attitudes—e.g., favor or disapproval— but at the same time they bear all the hallmarks of being assertions (i.e., they also express beliefs).
>
> Moral judgments purport to be practical considerations irrespective of the interests/ends of those to whom they are directed.
>
> Moral judgments purport to be inescapable—there is no "opting out."
>
> Moral judgments purport to transcend human conventions.
>
> Moral judgments imply notions of desert and justice (a system of "punishments and rewards").
>
> For creatures like us, the emotion of guilt (or "a moral conscience") is an important mechanism for regulating one's moral conduct.[54]

6 The Moral Foundations Theory

There are several grand models or theories attempting to systematize and make sense of human moral psychology. Probably the most discussed is the so-called Moral Foundations Theory developed by Jonathan Haidt and Jesse Graham, using work done by anthropologist Richard Shweder.[55] The theory sets out to list the basic allegedly innate "taste buds" that shape moral thinking and that accounts for both similarities and differences between cultures and individuals. Each of these taste buds refers to a domain or set of problems that we tend to moralize, and the conjunction of them all is what we may think of as the moral foundation. Each domain arose in response to a set of adaptive challenges for our ancestors, and the emerging psychology makes some moral judgments and ideas of virtue more likely. For example, early humans had a lot to gain from cooperation, but also risked being exploited. Successful interactions will lead to emotions like gratitude, while having been taken advantage of will trigger anger. We think of

that or not. In the present discussion, I am primarily interested in the psychological part

[54] Joyce 2007, p 262.

[55] Following an innate human tendency to focus on visible high-status individuals, in the henceforth I will simply refer to it as Haidt's view. See Haidt 2012.

character traits relevant for this domain as virtues such as trustworthiness, fairness etcetera. The suggestion that there is a class of "triggers" with corresponding moral taste buds in us is tantamount to strong nativism. On Haidt and Graham's view, human morality can be plausibly broken down to these five fundamental building blocks: care/harm; fairness/cheating; loyalty/betrayal; authority/subversion, and finally sanctity/degradation.

The set of adaptive challenges and their corresponding solutions in our psychological make-up is summarized in the chart below.[56] We need not here go into, and critically assess, the precise taxonomy. The important thing is the overall conclusion that moral judgments are underpinned by a suite of emotional responses put in place by evolution to regulate behavior, including signaling to others.

	Care/ harm	**Fairness/ cheating**	**Loyalty/ betrayal**	**Authority/ Subversion**	**Sanctity/ degradation**
Adaptive challenge	Protect and care for children	Reap benefits of two-way partnerships	Form cohesive coalitions	Forge beneficial relationships within hierarchies	Avoid contaminants
Original triggers	Suffering, distress, neediness expressed by one's child	Cheating, cooperation, deception	Threat or challenge to the group	Signs of dominance and submission	Waste products, diseased people
Current triggers	Baby seals, cute cartoon characters	Marital fidelity, broken vending machines	Sports teams, nations	Bosses, respected professionals	Taboo ideas (communism, racism)
Characteristic emotions	Compassion	Anger, gratitude, guilt	Group pride, rage at traitors	Respect, fear	Disgust
Relevant virtues	Caring, kindness	Fairness, justice, trustworthiness	Loyalty, patriotism, self-sacrifice	Obedience, deference	Temperance, chastity, piety, cleanliness

[56] Chart from Haidt 2012, p. 125.

7 *Universal Moral Grammar*

Claims of innateness in a certain domain typically arise from the suggestion that our competence in that domain outmatches, and cannot plausibly be accounted for by, the exposure to stimuli, training, and experience (hence the contrast with "empirical" views, according to which competence is a function of the input received).

Perhaps the most famous case of a proposed innate competence is Noam Chomsky's idea of a *Universal Grammar*.[57] The term "Universal Grammar" is alternately used to refer to either the fundamental and thus shared features of all human languages, or to the brain circuitry (in the human child) which allows them to pick up their first language. In any event, the central idea is that language acquisition and the basic structural similarities of unrelated languages make it plausible to assume there is a commonly shared, biologically innate faculty of language. As Steven Pinker puts it:

> Language is not a cultural artifact that we learn the way we learn to tell time or how the federal government works. Instead, it is a distinct piece of the biological makeup of our brains. Language is a complex, specialized skill, which develops in the child spontaneously, without conscious effort or formal instruction, is deployed without awareness of its underlying logic, is qualitatively the same in every individual, and is distinct from more general abilities to process information or behave intelligently.[58]

Here are some features of language and language-users that universal grammar accounts for quite well: that all human societies have language; that children are able to understand and use sentences they have never heard before; that languages around the world vary but show adherence to a limited set of basic shared rules; that children quickly learn to judge whether a new sentence is grammatically correct or not; that people know what is correct but are unaware of any grammatical rule justifying that assessment; that speaking our first language is effortless; that children surrounded by an artificial *pidgin* lacking consistent grammatical rules develop a *creole*, a natural language with consistent grammatical rules absent from the stimuli they've encountered; that deaf babies of deaf parents "babble" using hand gestures in the same rhythmic way hearing infants babble with their

[57] Over the years Chomsky has revised – some say, retracted – the hypothesis of universal grammar. His early statement of the view appears in Chomsky 1965. A more recent discussion appears in Chomsky et al 2002.

[58] Pinker 1994, p. 18.

voices; that deaf babies spontaneously develop a grammatically natural sign language.

Can this enormously powerful idea from linguistics – a shared basic competence leading to different but structurally related languages – fruitfully be brought to bear on the question of moral innateness? Is there, in morals, as in linguistics, a shared universal "grammar"? Chomsky himself suggested the idea, it was later favorably hinted at by Rawls, and it has more recently been developed meticulously by John Mikhail.[59] According to Mikhail, the human mind holds "a complex and possibly domain-specific set of rules, concepts and principles that generates and relates mental representations of various types. Among other things, this system enables individuals to determine the deontic status of an infinite variety of acts and omissions."[60] This feat is done, so the theory says, by internal, inaccessible, fast and automatic representations of action-types valencing the subject's assessment of what the right thing to do is. Mikhail mentions as an example *battery*, i.e. acts which are "purposefully or knowingly causing harmful or offensive contact with another individual or otherwise invading another individual's physical integrity without his or her consent".[61] The presence in human psychology of a tendency to keenly observe battery makes evolutionary sense and also accounts for the universal finding that all human populations surveyed rank the various versions of the trolley dilemmas the way they do.[62] (*Push* being a form of battery while *Switch* and *Loop* are not)

If this innate automatic categorization and moral valancing of various behavior exists, we should expect some of the following:

1. A competence on the part of children to acquire, accept, and use moral judgments that goes beyond explicit instruction.
2. A cross-cultural variety of moral norms, nonetheless understandable as surface phenomena compatible with a shared basic set of norms.
3. A competence to judge novel moral situations.
4. A competence to make moral judgments on situations and behaviors while being unable or hard-pressed to supply a justifying foundation for the judgment.

[59] Mikhail 2007, 2008, and 2011. Rawls, p. 46-7.

[60] Mikhail 2007, p. 144.

[61] Ibid, p. 145.

[62] Ahlenius & Tännsjö 2012. More recently, and with a larger data set: Awad et al. 2020.

5. A first-person phenomenology of moral judgments typically being obvious and effortless, akin to perception.
6. The presence in all natural languages of expressions to convey what is obligatory, permissible, and forbidden.

These predictions are very accurate.[63] For our purposes here, there is no need to definitively try and settle which ones of the different nativist systems is the correct one. It is also not totally clear if they are incompatible and, if so, what evidence could refute or vindicate any one of them over the others. What I do want to establish, is that we have an innate tendency to make moral judgments, and that this capacity is pre-structured. There are domains we are more likely to moralize than others, and there are dispositions or biases making us more favorable to some forms of moral judgments over others.

Since Haidt's view so explicitly grounds moral thinking and moral judgment in emotional responses, it may seem to be incompatible with, or at least quite different from, Mikhail's universal grammar view. Is not the appeal to emotions sufficient to account for moral convictions – do we really need this talk of a complex moral "grammar", hidden from even ourselves yet at work deep in our minds? But emotions alone cannot explain the data. We know that some situations give rise to more emotions than others, but *why*? This is where the action type, the "grammar", enters the picture. Some behaviors give rise to emotional responses because of the way they are represented to us.

8 Framing effects

There are additional ways in which the way we represent actions taps into innate triggers making us favor or reject options while having a hard time articulating just how they differ. Daniel Kahneman and Amos Tversky, and many others in their aftermath, famously studied human irrationality and our use of *heuristics*, i.e. tools or cognitive autpilots simplifying the process of arriving at a decisions. Here is a telling example of their research. Two groups of respondents are both given the following scenario:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

[63] See for example Haidt 2001, Hauser et al 2007, Mikhail 2011, Hamlin 2013.

At this point the two response groups are given different versions of the options. Participants of group 1 are given the following two response options, and asked to recommend one of them:

> If Program *A* is adopted, 200 people will be saved.
>
> If Program *B* is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

And participants of group 2 are given the following two options, and asked to recommend one of them:

> If Program *C* is adopted 400 people will die.
>
> If Program *D* is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

Among participants of the first group, 72 percent preferred A and 28 percent preferred option B. In the second group 22 percent of participants recommended C and 78 percent recommended D.[64] *But* A *and* C *are the same options, and* B *and* D *are the same options*. So it is not the case that people are consciously applying, say, a cost-benefit or a maximin strategy to solve matters like these. Rather the way we assess the options is biased by a certain representative scheme that makes us separate options that are logically equivalent. The framing is a framing precisely because the different descriptions tap into and activate deep-seated moral appraisal systems. If there were no such deep-seated appraisal systems, but simply an all-purpose reasoning ability to approach public policy or moral matters, we would not be framed.

It is a crucial feature of the moral grammar hypothesis that (as with language) we do not have conscious access to the normative rules it contains. If we could not be framed by the wording of these scenarios that would be evidence for *resistance* to automaticity based on an innate normative rule box, and speak for the notion that moral problems are approached in a cognitively transparent way by domain-general features of our rational capacities. But we can, and they are not.

9 The moral/conventional distinction

I think it is fair to say common sense recognizes a distinction between transgressions of norms of, on the one hand, convention and, on the other, morality. The primitive idea is that some of the norms we abide by are simply the result of agreement and habitude – and can be lifted

[64] Tversky & Kahneman 1981, p. 453; also recounted in Kahneman 2011, p. 368.

or changed by an act of will – whereas other norms command subservience in a manner not reducible to mere agreement or habit. These latter norms, then, would be the moral norms.

There is some debate as to the possibility of giving this intuitively plausible distinction any philosophical and psychological substance. As Nicholas Southwood notes, "We are therefore in the unfortunate but all too familiar philosophical position of being in possession of a distinction that makes a great deal of sense pre-theoretically, yet being in want of a compelling vindicating account of the distinction".[65]

If there is a universal moral grammar it seems we should expect children to make that distinction. And they do. A typical study will have children answer questions like "In school, no one is allowed to speak before raising his or her arm. Suppose the teacher said 'today anyone can just speak directly without first raising their arm', would it then be okay to speak directly without raising the arm?", and then compare their answers to questions such as "In school, no one is allowed to pull the hair of another kid. Suppose the teacher said 'today you can pull the hair of someone else if you want to', would it then be okay to pull someone else's hair?".

If an authority figure cancels the rule, children think it is okay to not raise the hand before speaking, but they do not think pulling someone's hair would be made okay by an authority figure's cancellation of that rule.[66] Or, in other words, it is not the kind of rule that derives its clout from an authority. Pulling someone's hair constitute a form of harmful battery, and is hence seen as violating a *moral* rule, the compliance to which is not optional (see Joyce's set of characteristics of moral judgments above).

10 The case against innateness

So there is a lot of psychological baggage which appears to make the nativist position well-supported. But the nature of the field is such that no position is uniquely plausible or uniquely compatible with the available evidence. In this final section I will examine Jesse Prinz' plea for the role of culture. Prinz is what is called an *empiricist* about the mind, and rejects innateness about psychological traits, language, emotions, thinking, and values. He does not think human moralizing is an adaptation and he does not believe there is a specific domain or

[65] Southwood 2011, p. 764

[66] An overview and critique is given in Kelly et al 2007. Kumar 2015 defends the distinction, arguing that moral judgments in a certain sense are a "natural kind".

content which we are naturally disposed to moralize, although as human life is organized it may often be that some domains rather than others are in fact the object of attention from the moral point of view. Prinz calls himself a "methodological anti-nativist" which amounts to a burden of proof-position in the present debate: "we should assume that a faculty is not innate until evidence leads us to say otherwise".[67] The idea is, we know we have some domain-general rational capacities and any theory postulating specific faculties or mental specialized routines *adds* to that, and we should accept these additions only insofar as doing so is prompted by the available empirical evidence.

Even if I were to accept this non-nativist default, it is clear from the preceding sections that I believe there indeed is evidence to support such specialized competencies as well as tendencies towards certain domains and contents. Speaking of contents, Prinz grants that there is evidence that people of different cultural backgrounds judge the trolley problems similarly, always holding *Switch* to be more permissible than *Loop*, which is deemed more permissible than *Push*.[68] The nativist take on this agreement is that human beings share innate mechanisms for representing and emotionally valancing action types. But, Prinz, claims, this cross-cultural agreement can just as well be explained in other ways:

> The fact that people in different cultures give similar responses might be explained by prototype effects. When people learn the concept murder, the paradigm cases involve direct intentional physical assault, not indirect harms. The reason for may have nothing to do with innateness. All cultures must have rules to stop people from directly and intentionally aggressing against each other, on pain of societal collapse. Rules against indirect harms, however, are less prevalent, because there are fewer circumstances within a society when indirect actions will result in someone's death, and a society that failed to have such rules might be relatively stable. The pushing scenario conforms most closely to the kind of actions that every society is likely to condemn. It is more clearly an instance of murder than the scenario in which a person is killed as the side-effect of diverting the trolley. In the "diversion" scenario, the death is also less salient and the cause of death for the one person is rendered comparable to the cause of death for the five, making the comparison between the two outcomes

[67] Prinz 2014, p. 105. Prinz complains nativism in the study of psychology is often implicit, while, interestingly, Laurence & Margolis 2013 complain empiricism is often assumed rather than argued for.

[68] Ibid., p. 106.

> vivid. So there need not be any unconscious rules at work here. People are taught that murder is wrong by means of prototypical cases, and they tolerate killing more readily when it departs from the prototype, lacks salience, or is rendered comparable to an alternative action that involves the same kind of killing but greater losses.[69]

Is Prinz right in suggesting that the data is just as well if not better explained by prototype effects and shared environmental constraints of human societal affairs? First, the notion that something catches on in our psyches because it is a *prototype* of something is quite close to the nativist notion of cognitive organization in advance of experience. Why is it that in all cultures murder is prototypically a form of *battery*, i.e. harmful or offensive contact with another? Prinz may respond that any kind of act can be used as a prototype, and then that act will be the standard people think of as the canonical token of that type of act. On the nativist view, some constructs are more likely to be seen as prototypical in part because the instruction we are all given exploits a preparedness in us, similar to how it is much easier to make a child fear snakes than electric sockets quite independent of how dangerous or common these things are in the environment the child grows up in.[70]

The picture on the empiricist or anti-nativist take is that emotions can be *directed at* customs and action types, thereby moralizing those customs and action types. Nativists (in particular adherents of the strong version) additionally hold that some customs and action types are prone to triggering emotions and associated judgments *in us*. On the anti-nativist view, we do not have emotions the target of which shape our moralizing. Rather, anything can be moralized, and emotions are then coopted to enforce that moralization.[71] But what is in fact moralized may vary wildly. This indeterminacy or openness, I claim, fits ill with the various sources of evidence covered so far in the chapter, and is in particular hard to explain given an observed continuity between us and related non-human animals. Chimpanzees do not *learn* to feel enraged by other's failure to reciprocate; they are enraged and display these emotions to prevent defections and punish those that have already occurred. And because we share this suite of emotions it is more plausible to suggest that their presence in us will guide what we moralize and how.

[69] Prinz 2014, p. 106.

[70] See Schmitt & Pilcher 2004 for references to the fear of snakes literature, but above all for the more general discussion of psychological adaptations.

[71] This approach is more fully developed in Prinz 2007.

The anti-nativist proposal further predicts that it would be equally easy to raise children by using indirect, nonviolent forms of killings as the prototype of wrongful killings, and the young members of this community would struggle to see that violently pushing someone in front of a train could be just as bad as switching a lever that would redirect the train onto them. More generally, the proposal would seem to imply that there are no innate psychological constraints on what human beings may be cultured to favor and disfavor in the moral domain. Training children to punish those who fulfill promises and reward those who break promises, and laude those who share and shun those who steal would all be equally effortful, since there would be no preprogrammed software making any behavior or principle seem to the children more or less attractive than any other. There is lots of evidence from research on infants and toddlers showing that they like and prefer those whom they have observed helping as opposed to hindering others. They also expect praise and blame to be directed at helpers and hinderers, respectively.[72] This is an expected observation on a nativist view, but seems harder to square with the empirical or non-nativist supposition. For sure, because of recurring facts of human life, norms of reciprocity are bound to appear in all cultures, but on the non-nativist view there is no expectation that toddlers come equipped with preferences or expectations for one or the other norm.

11 Conclusion

The whole moral innateness discussion is somewhat messy. The nature of the questions asked is not crystal clear, falling somewhere in a no-man's-land between psychology, evolutionary biology, cultural anthropology, and philosophy. The positions advanced tend to take the form of "models", and I think it is fair to say that all stances are, as of now, underdetermined by the available evidence. I believe the most fruitful way to think of nativism is as a developing research program rather than a certain position tied to the specifics of John Mikhail's universal moral grammar or Jonathan Haidt's moral foundations theory. Nativism in this broad sense provides, I have claimed, the best account of many observations, ranging from cross-cultural data to child development and people's justifications of their moral judgments. There are many details that remain to be filled in for any project aiming to pro-

[72] See Bloom 2013 for further references and discussion.

vide a theory of innate moral psychology. Haidt's and Mikhail's theories are just two examples, and there are others still of course.[73] There will be much overlap among any plausible accounts, and there will be remaining disagreements that may be possible to resolve given careful observation and testing as well as more theoretical considerations.

But, after all, maybe the disagreements here are overblown. Prinz speaks of plasticity but also accepts that,

> Perhaps some moral rules are easier to learn than others and some might even be impossible to sustain. Morality is no doubt constrained by our biological endowment. The emotions we have, our capacity to attribute mental states, and our care for kin all serve as building blocks that help shape the outcome of norm construction […] the scientific study of morality should not be limited to psychology, neuroscience, ethology and biological evolution. It should expand to include anthropology, history, sociology, and other fields that track sources of cultural variation. A complete science of morality will work at multiple levels.[74]

While placing our emphasis at different parts in the full story of human morality, we can all agree to the richer, non-reductive mode of inquiry Prinz pleas for. That mindset is especially fitting given a research area where the absence of decisive evidence is so manifest.

Remaining research and disagreements notwithstanding, there are two lessons to draw, that will turn out to be of interest in ensuing discussions: 1) We have an innate set of emotional responses and biases that make some moral judgments seem more plausible than others to us; and 2), we often lack conscious access to a set of justifying principles from which individual judgments are derived, making us vulnerable to self-deception or at least ignorance of why it is we hold certain moral views. Importantly, these conclusions can be accepted by many of the participants in the debate over moral innateness.

[73] See for instance Curry, Mullins, and Whitehouse 2019.

[74] Prinz 2014, p. 115-6.

2 Deontology: Reductio ad Amygdalam

It was truly a new thing when, in the fall of 1999, philosophy grad student Joshua Greene approached Jonathan Cohen of the neighboring Department of Psychology to suggest that they together put people pondering moral dilemmas into a brain scanner. The research that followed led Greene both to formulate an empirical theory of moral judgment and to challenge certain moral theories and ways of doing moral philosophy. His initiative came after a nagging suspicion that philosophers had exhausted the possible theoretical maneuvers by which to accommodate a commonly shared but irritatingly straggly set of intuitions regarding the so-called *Trolley Problems*. The fact that it seems hard for philosophers to formulate a coherent whole wherein these intuitions fit, is, after all, an interesting psychological phenomenon: how and why do individuals harbor closely related views that there is no obvious common support for, or that may even be incompatible with one another?

In this chapter I will present Joshua Greene's challenge to deontological moral philosophy which he bases in what he calls the dual processes theory of moral judgment. This theory is, he believes, the best way to make sense of a vast body of research from behavioral and neuroscience on cognition in general and moral thinking in particular. After going through the neuroscientific studies, I will present and critically assess the challenge the results are purported to face deontology with. My presentation here will follow the rebuttal offered by Selim Berker, whose critical assessment of Greene is the most thorough presented so far. In this part my discussion is mostly critical of Berker and so defends Greene, although I also side with Berker in the larger assessment that the victory over deontology was called prematurely. I finish by discussing post-trolleyology work by Frances Kamm and Judith Jarvis Thomson, pointing to avenues left to explore for a deontologist view.

1 The trolley problems

Philippa Foot first described a moral dilemma subsequently known as the trolley problem in her 1967 paper "The Problem of Abortion and the Doctrine of the Double Effect".[75] The story of the dilemma goes like

[75] Foot 1967.

this. A runaway trolley plunges toward five people strolling about further down the tracks. The only way to save this unfortunate group of people is to divert the trolley down another branch once it reaches a junction just ahead of the strollers. The problem, alas, is that further down that other branch one person is strolling about equally innocently. Should the bystander interfere, and divert the trolley down the less populated branch, so that five may live and one die rather than the other way around?

Now consider the *Footbridge Push dilemma*.[76] As before, a runaway trolley is heading towards five flaneurs further down the track. This time, however, there is no alternative branch for it to veer into. The bystander is now standing on a footbridge overarching the tracks between the rushing trolley and the five people. The only way to save the five individuals is for the bystander to push a very large person, standing next to her on the bridge, in front of the trolley, thereby thwarting its course. (Our protagonist bystander herself, we may assume, is not sufficiently heavy to stop the trolley.) Would it be morally okay for the bystander to push the large stranger, thereby killing one person so that five may live?

Surveys and classroom experience again and again confirm that people generally accept or favor switching the train in the first case but reject pushing the person in the second case.[77] Why is it that most people think it is permissible to redirect the trolley, thereby killing one instead of five in the first case, but impermissible to kill one in order to save five in the other? Perhaps it has to do with the notion of using someone simply as a means to other's ends? In the first case, the death of a single person is merely an unfortunate side effect of the bystander's diverting the trolley onto another branch. In the footbridge case, however, the large person is used as a means to stop the trolley: his death – or rather the fact that the trolley grinds to a stop when it hits him – is a prerequisite of the other's surviving.

This resort runs into trouble once we consider yet a variant, the *Loop case*.[78] As in the first case, the bystander can divert the train onto another branch. This branch, however, loops back to the same track, thereby only prolonging for seconds the lives of the five people further down. As it happens, a very large person wanders on the tracks of the

[76] Thomson 1985.

[77] Ahlenius & Tännsjö 2012, Greene 2013; 2016, Hauser 2006; Awad et al. 2020.

[78] Thomson, op. cit.

loop, and should the trolley be shifted onto the loop he will be sufficiently heavy to arrest the trolley (but not sufficiently heavy to survive the collision). The only way to save the five is to direct the trolley onto the loop. Most people deem it permissible to divert the trolley onto the loop, thereby killing one to save five.[79] Since his colliding with the trolley is a prerequisite for saving the others, it may seem unreasonable to call his being hit by the trolley an undesired side effect; it is, rather, part of the very plan to save the five. Nor can the reasoning behind people's different judgments be traced to the notion of using someone as a means, since in both the footbridge and the loop case someone *is* used as a means. Yet, most people are willing to accept diverting the trolley onto the loop, and most do not accept pushing the large person off of the footbridge.

What, then, does account for people's different judgments? After all, all three cases – the "standard" trolley switch problem, footbridge push dilemma, and the loop case – seem to raise the very same problem of whether or not it is permissible to sacrifice one in order to save five.[80] This is a problem for philosophers of a deontologist bent, and they go to great lengths in order to reach harmony amongst our different judgments on the ever more diabolic scenarios.[81]

2 The trolley problem from a psychological point of view

The suspicion that people respond in an incoherent way is not only a problem for philosophers interested in justifying our differing judgments, but in itself opens an attention-grabbing psychological issue: "How is it that nearly everyone manages to conclude that it is acceptable to sacrifice one life for five in the switch but not in the footbridge dilemma, in spite of the fact that a satisfying justification for distinguishing between these two cases is remarkably difficult to find?"[82] Perhaps the best explanation of people's differing opinions when confronted with the two cases is not that the scenarios differ in morally relevant ways, but rather in what psychological effects the descriptions of the situations have on us. This suspicion of Greene's led him to attack the topic from a different angle than philosophers so far had done.

[79] Greene 2013, Awad et al. 2020.

[80] For brevity, I will often refer to the different cases as *Switch*, *Push*, and *Loop*.

[81] For further discussion see Kamm 2007. A short version of her view can be found in Kamm 2000. Unger 1996 painstakingly dissects the dilemmas and our usual responses to them. A more recent very readable book-length treatment is Edmunds 2014.

[82] Greene et. al. 2001, p. 2106.

3 *An fMRI investigation of emotions and moral dilemmas*

Greene and his fellow researchers speculated that, from a psychological point of view, the difference between *Switch* and *Push* might lie "in the latter's tendency to engage people's emotions in a way that the former does not".[83] What is it, then, about certain moral dilemmas that elicit our emotions in a way others do not? They had an idea: situations where we are personally required to use force or otherwise harm a particular individual evoke our emotions to a significantly higher degree than situations where we bring about *the same result* in a less up close and personal way. Intent on testing this hypothesis, the team predicted that "brain areas associated with emotion would be more active during contemplation of dilemmas such as the footbridge dilemma as compared to during contemplation of dilemmas such as the trolley dilemma."[84] A further empirical consequence of their hypothesis was articulated: because normal people will have an emotional reaction that disposes them to discard as inappropriate personal violations in order to bring about overall better consequences, the minority of subjects who reach the opposite conclusion will take longer time to do so.[85]

In order to measure the presence of emotional processing, the team used a technique called Functional Magnetic Resonance Imaging, or *fMRI*.[86] By tracking changes in blood flow, the technique tells us which parts of the brain are activated. In a typical experiment, the subject will lie in an *fMRI* scanner and a particular form of stimulation will be set up, for instance text or images on screens in front of the person. As

[83] Ibid. They write, further, "The thought of pushing someone to his death is, we propose, more emotionally salient than the thought of hitting a switch that will cause a trolley to produce similar consequences, and it is this emotional response that accounts for people's tendency to treat these cases differently."

[84] Ibid.

[85] "In light of our proposal that people tend to have a salient, automatic emotional response to the footbridge dilemma that leads them to judge the action it proposes to be inappropriate, we would expect those (relatively rare) individuals who nevertheless judge this action to be appropriate to do so against a countervailing emotional response and to exhibit longer reaction times as a result of this emotional interference. More generally, we predicted longer reaction times for trials in which the participant's response is incongruent with the emotional response (e.g., saying "appropriate" to a dilemma such as the footbridge dilemma). We predicted the absence of such effects for dilemmas such as the trolley dilemma which, according to our theory, are less likely to elicit a strong emotional response." (Ibid.)

[86] The initial kind of empirical evidence Greene invoked was retrieved using *fMRI*, but the larger findings are not constrained to neuroscience specifically, nor to the merits of the *fMRI* technique. See Rachul & Zarzeczny for further discussion of *fMRI*.

stimuli and tasks are presented, MRI images of the subject's brain are taken. After the experiment has finished, the set of images is analyzed. Firstly, the raw input images from the MRI scanner require mathematical transformation (a kind of spatial "inversion") to reconstruct the images into "real space", so that the images look like brains. The final statistical image shows up bright in those parts of the brain that were activated by the experiment's stimulus. These activated areas are then shown as colored blobs on top of the original high-resolution brain image.

In a set of two consecutive experiments, subjects were undergoing *fMRI* brain scannings as they were asked to respond to a battery of 60 practical dilemmas. These dilemmas had been previously categorized as either moral or non-moral on the basis of the assessments of pilot participants. Typical non-moral dilemmas posed questions such as the best way to arrange a travel schedule given certain constraints, which of two coupons to use at a store, or what kind of nuts to use for your brownie given your tastes. Here is an example: [87]

> You are looking to buy a new computer. At the moment the computer that you want costs $ 1.000. A friend who knows the computer industry has told you that this computer's price will drop to $ 500 next month.
>
> If you wait until next month to buy your new computer you will have to use your old computer for a few weeks longer than you would like to. Nevertheless you will be able to do everything you need to do using your old computer during that time.
>
> Is it appropriate for you to use your old computer for a few more weeks in order to save $500 on the purchase of a new computer?[88]

The moral dilemmas, in turn, had been categorized as either "personal" or "impersonal". Typical impersonal moral dilemmas concerned what policy to vote for given certain differences in outcome, whether or not to return the lost wallet of an immensely rich person, the switch trolley dilemma etcetera. Personal dilemmas, on the other

[87] Subjects were positioned in front of a display where the dilemmas were presented as text through a series of three screens, the first two describing a scenario and the last posing a question about the appropriateness of an action one might perform in that scenario. Subjects responded by pressing one of two buttons (Appropriate" or "Inappropriate".)

[88] www.sciencemag.org/cgi/content/full/293/5537/2105/DC1. Ensuing dilemmas are all available there.

hand, involved materials such as a doctor deliberately killing one patient in order to transplant that individual's organs to five dying patients, the footbridge dilemma, pushing your despicable boss off a tall building etcetera. Over the years, Greene has tinkered a bit with what makes a dilemma "personal" or "impersonal", but in these early studies what he had in mind was the idea that personal violations are (1) likely to cause serious bodily harm (2) to a particular person (3) where the harm is not a result of deflecting an existing threat onto a different party. A moral dilemma is "impersonal" if it fails to meet these criteria (each necessary; jointly sufficient).[89] Here is how the standard trolley switch problem (categorized as "Moral-Impersonal") was presented:

> You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman.
>
> If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman.
>
> Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?[90]

And the footbridge push dilemma (categorized as "Moral-Personal"):

> A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large.
>
> The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved.
>
> Is it appropriate for you to push the stranger on to the tracks in order to save the five workmen?

Many other moral dilemmas were presented, such as this (categorized as "Moral-Personal"):

[89] Greene 2004, p. 345.

[90] The term "appropriate" is a bit peculiar, since it may be interpreted as "in accordance with societal standards" rather than, simply, "morally permissible". It was chosen as it was deemed the best term covering both moral and non-moral dilemmas in the battery of questions (Greene, personal communication).

> Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.
>
> Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
>
> Is it appropriate for you to smother your child in order to save yourself and the other townspeople?

The term "dilemma" was used in a wide sense indicating simply that there was a choice to be made, though it did not have to be a tough one in the eyes of most people:

> You have been dissatisfied with your marriage for several years. It is your distinct impression that your wife no longer appreciates you. You remember how she appreciated you years ago when you took care of her after she was mugged. You devise the following plan to regain your wife's affection.
>
> You will hire a man to break into your house while you are away. This man will tie up your wife and rape her. You, upon hearing the horrible news, will return swiftly to her side, to take care of her and comfort her, and she will once again appreciate you.
>
> Is it appropriate for you to hire a man to rape your wife so that she will appreciate you as you comfort her?

The scanning consistently showed a higher level of activation in emotion-related brain areas during contemplation of the personal moral dilemmas compared to the impersonal ones. Correspondingly, parts of the brain known to be involved in working memory and information processing were more activated by the impersonal than by the personal dilemmas.[91] In effect, from the point of view of what goes on in the brain, our processing of impersonal moral dilemmas have more

[91] More specifically in personal moral dilemmas, the medial frontal cortex, posterior cingulate cortex, and angular gyrus/superior temporal sulcus are active. In impersonal moral dilemmas there is increased activity in the dorsolateral prefrontal cortex and parietal lobe. See table on page 2106 of Greene et. al. 2001. It is beyond the scope here to assess the robustness of claims about one part of the brain being "associated with" or "responsible for" emotion rather than information processing or vice versa.

in common with other kinds of cognitive skills and information processing than with our way of dealing with (or, rather, responding to) personal moral dilemmas.

The prediction that emotionally incongruent answers would need longer reaction times was confirmed too. As expected, most subjects did judge throwing the large person off the footbridge, or killing one patient in order to save five, as "inappropriate". Subjects who nevertheless found it appropriate did so against their own emotional reaction. Hence, due to "emotional interference", subjects who judged that actively harming one for the sake of minimizing overall harm was appropriate in the personal settings needed more time to reach their conclusions.[92] Greene and his collaborators compare this interference or ambivalence to the so-called Stroop effect, a familiar case of subjects' needing a prolonged time to name the color green in response to the word "red" written in green ink.[93]

4 Accounting for the results

Greene sees the results as fitting in to, and supporting, a theory he calls *The dual process theory of moral cognition*. The dual process theory proposes that there are two distinct but interacting kinds of processes involved in assessing and producing moral judgment. One part of the system is automatic, quick, and introspectively opaque. The other part of the system is effortful, slow, and introspectively accessible. Readers familiar with Daniel Kahneman's work will call them System 1 and System 2 respectively.[94] Because we are not really aware of the workings of the automatic and quick processes, our introspective experience of moral thinking tends to mostly mirror the consciously engaged part. Hence, the dual process theory may be absent from or in tension with common sense.

Greene likens our moral psychology to a camera with automatic and manual modes. For most occasions, the user will get the best result by putting the camera in automatic mode, letting it quickly sense and adapt to typical situations such as portrait, landscape, sport etcetera. If

[92] Greene et. al. 2001, page 2107.

[93] Ibid., page 2106. In a later study, Greene et al. 2004, this link was more thoroughly investigated.

[94] Kahneman 2011. Dividing the workings of the brain and the mind into *two* distinct kinds of processes is obviously an oversimplification, but nonetheless a useful one. The more relevant distinction is that between automatic and deliberate processes. For some complications, see Melnikoff and Bargh 2018.

you want to be creative, or if you find yourself in a situation that the automatic settings seem inept at making the best of, you will need to switch to the camera's manual mode instead. This allows you to be as flexible as you like, but it will take more time.[95] The dual processes operation of the moral mind is thus, according to Greene, the solution to a type of problem or tradeoff for any organism facing choices: that between fast efficiency and time-consuming flexibility.

That moral decision-making, like many other mental tasks, has these features seems empirically not very outlandish and is philosophically fairly innocuous. It is when we combine the empirical theory with some philosophical content that we get to the more controversial features. According to Greene, there is as *Central tension*, such that

> Characteristically deontological judgments are preferentially supported by automatic emotional responses, while characteristically consequentialist judgments are preferentially supported by conscious reasoning and allied processes of cognitive control.[96]

This tenet of his position – the idea, roughly, that the division of our moral psyches into passion and reason is mirrored by a divided output in the form of deontological versus consequentialist moral judgments – is both empirically more controversial and, more importantly for present purposes, the core of the debate on the possible philosophical relevance of this kind of empirical work on moral psychology.

Describing the human mind in general, and moral psychology in particular, as a semi-stable amalgamation of passion and reason is hardly new. What is new is the possibility now, as opposed to during the times of Plato or Hume, to interconnect a variety of disparate evidence into one explanatory whole. Here are some findings, in addition to Greene's initial *fMRI* study that lend support to the dual process theory of moral judgment:

i. Moral dumbfounding: people find themselves accepting moral judgments that they continue to hold even after having their stated reasons for holding them cancelled.[97]
ii. Emotive priming induced under hypnosis affects moral judgment. People that were primed to experience disgust upon hearing the word "often" bizarrely condemned a stu-

[95] Greene 2014.
[96] Ibid., p. 699.
[97] Haidt 2010; 2012.

dent council representative in charge of scheduling academic discussions for *often* choosing topics that will be of interest to both students and professors.[98]

iii. Inducing disgust makes people judge moral transgressions more severely. For instance, a jury deliberating in a filthy room will arrive at more draconic verdicts than a jury deliberating in a tidy room.[99]

iv. Cognitive load decreases utilitarian judgments but does not affect deontological judgments.[100]

v. Empathy increases deontological judgment.[101]

vi. Mirth increases utilitarian judgment.[102]

vii. Psychopaths are more prone to making utilitarian judgments.[103]

viii. VMPFC patients are more utilitarian.[104]

ix. Activity in the Amygdala decreases utilitarian judgments.[105]

x. Alcohol dependence blunts emotions and increase utilitarian judgments.[106]

xi. Removing time-frames and encouraging deliberation increases utilitarian judgments.[107]

xii. Solving math problems with unintuitive solutions lead to people making more utilitarian judgments.[108]

xiii. Individuals favoring effortful thinking over intuitive more likely to make utilitarian judgments.[109]

[98] Wheatley and Haidt 2005.

[99] Ibid. See also Schnall et al. 2008. The evidence for a causal link between disgust and severity of moral judgment is in a state of flux. This study supports a link: J. L., Steckler, C. M., & Heltzel, G. 2019. Whereas others have found it hard to replicate, see for instance Ghelfi, et al. 2020.

[100] Bonnefon, De Neys, and Trémolière 2012. Greene et al. 2008. Also see Greene et al. 2009.

[101] Conway and Gawronski 2013.

[102] Valdesolo and DeSteno 2006 and Strohminger, Lewis, and Meyer 2011.

[103] Koenigs et al 2012.

[104] Koenigs et al. 2007. Also Thomas, Croft, and Tranel 2011.

[105] Shenhav & Greene 2014.

[106] Khemiri et al. 2012.

[107] Suter & Hertwig 2011.

[108] Paxton, Ungar, and Greene 2012.

[109] Bartels 2008.

xiv. People have a conscious awareness of what speaks in favor of a utilitarian judgment but often are unaware of why they accept a deontological judgment.[110]

xv. The more need for cognition a person has, the more utilitarian and less punitive will be their justification of criminal punishment.[111]

xvi. Negative emotions produce condemnation of harmless "wrongs"[112]

5 Philosophical relevance of the study: a problem for deontology?

The dual process theory, and its auxiliary central tension thesis, is an empirical theory about the causal mechanisms involved in processing and judging moral matters. In order to establish or refute the theory a smattering of moral philosophy might help, but ultimately the decisive evidence comes from neuroscience and psychology. How, then, might this empirical theory influence debates in philosophy? Here I will take on Greene's and Peter Singer's proposals that if the dual process theory is true or roughly so, we have reason to lower our credence in deontological normative theories, and a corresponding reason to strengthen our relative credence in utilitarianism.

But how, more precisely, can we understand this challenge to deontological ethics? Here I will follow Selim Berker's statement of it, and in turn I will offer some criticism of his rebuttals.[113] Berker goes through three possible candidate formulations of the challenge that he thinks are weak, but that Greene and Singer in some instances could be interpreted as relying on. After discussing these weak arguments, he moves on to what he takes to be the most promising way of stating the challenge to deontology. Before taking on his rejection of the more promising case, let us look first at the weaker ones. The three weak arguments are: 1) The "Emotions bad, reasoning good" argument; 2) The argument from heuristics; 3) The argument from evolutionary history. I will briefly comment on each of these weaker arguments, and focus my discussion on what Berker takes to be the best formulation of the challenge.

[110] Cushman, Young, and Hauser 2006; Hauser et al. 2007.

[111] Carlsmith, Darley, and Robinson 2002. Also see Sargent 2004. The link between "need for cognition" and utilitarian thinking is questioned in Kahane, et al. 2018.

[112] Haidt, Koller, and Dias 1993. Also see Wheatley and Haidt 2005.

[113] Berker 2009.

6 *"Emotions bad, reasoning good"*

Because deontological judgments are driven by emotion and consequentialist judgments involve abstract reasoning, the former carries no "genuine normative force".[114] This argument is weak, according to Berker, because it assumes rather than argues for the view that emotions are not a reliable way of discerning moral truths. Although Berker is right that there is a tradition that would happily embrace the view that emotions are guides to moral truths, I am not sure deontologists typically see themselves as belonging to that tradition. So, for *deontologists*, the finding that some characteristically deontological judgments correlate with brain activity having to do with disgust, fear etcetera may seem to them of some relevance.

As far as I am aware all existing deontological theories are constructed alongside some form of rationalist, non-sentimentalist metaethical view. Mark Timmons has suggested that this need not be so, and thus that these findings do not threaten deontological theory. He does not really develop such an account but writes,

> Although all of the versions of deontology that I know of have been embedded in a rationalist metaethic, I don't see why one cannot embrace sentimentalism (or expressivism) and go on to defend a deontological moral theory. Sentimentalism is a metaethical account about the nature of moral judgment; deontology is a normative theory about the right, the good, and their relation to one another. Although sentimentalism may seem to fit most comfortably with consequentialism, accepting the former metaethical view does not commit one to the latter normative moral theory. So again, I don't see how (without further elaboration) the empirical facts about emotion-laden, intuitive moral reactions pose a threat to deontology. Indeed, I would suggest that the way to develop a deontological moral theory is to do so within the framework of a broadly sentimentalist metaethic[115]

This is an interesting proposal, and one that would need to be spelled out in some more detail before it can be assessed. But are metaethical views on the nature of moral judgment really theories of the causal mechanisms producing evaluative talk? Though semantic views like expressivism and cognitivism purport to make the best sense of available data, our subjective experiences of accepting a moral view etcetera, they are something different than psycholinguistic theories. It may be a bit unclear just what they are more precisely, and what kind of evidence can be brought to bear in order to test them. But, at bottom,

[114] Berker's expression (page 316).

[115] Timmons 2008, p. 102

they are philosophical, not psychological.[116] The issue here is whether or not what triggers an emotional reaction in creatures such as ourselves is a reliable guide to what is morally pertinent. Saying that they are, and specifying all of the characteristics of actions and agents that should lead a moral person to experience such emotions, will mean working out a normative theory. Maybe this theory will say that the fact that a harm was brought about in an up-front and violent manner matters morally; presumably it must. This will not really be a change at the level of metaethics, but at the level of first-order moral theory.

7 "Deontological judgments are based in heuristics"

An elaboration on the Emotions bad argument, Berker claims, would be the notion that deontological judgments are unreliable because they take the form of heuristics, i.e. a form of standardized shortcuts that facilitate or make redundant costly deliberation.[117] The problem with this would be some well-known errors associated with employing heuristics in domains such as probability.[118] The major problem with using the heuristics charge to undermine deontology is that, in contrast to probability, in morals we cannot (as easily) compare the heuristics to truths and note which heuristics are right and which are wrong. Berker is right about that: we can assess if heuristics about a certain domain are largely trustworthy or not only if we have an independent way to assess the truth about claims made in that domain. And in the case of ethics, arguably we have no such thing. Still, considering what we know from the use of heuristics in areas like rationality and probability may make us somewhat hesitant to take them at face value in ethics.

8 "The argument from evolutionary history"

The third way of characterizing the challenge Berker calls "The argument from evolutionary history".[119] It says that our emotion-driven deontological judgments have their evolutionary roots in an ancient environment we no longer find ourselves in, and that, because of this, in contrast to consequentialist intuitions, they have no "genuine normative force". But, Berker says, this is no more than a just-so story, and "presumably consequentialist intuitions are just as much a product of

[116] See Joyce 2008.
[117] Berker 2009, p. 317.
[118] See Kahneman 2011, esp. chapters 10 thru 18.
[119] Berker, p. 319.

evolution –whether directly or indirectly – as deontological intuitions are, so an appeal to evolutionary history gives us no reason to privilege consequentialist intuitions over deontological ones." The possible contrast between tendencies in us that favor deontological ethics and utilitarian ethics and their link to evolution will be explored in the next chapter, so I will postpone commenting on that point here.

To sum up, I do somewhat disagree with Berker that these three versions of the argument leave deontology altogether unaffected. If we come to believe, based on studying neuroscience and other empirical data, that deontological judgments are often triggered by emotions, take the form of heuristics and are part of our evolved psychology from a time when moral issues could not arise beyond the horizon of the surrounding tribe, this may subtract some plausibility points from this way of thinking about ethics. Separating these three interpretations from the coming fourth most interesting one, to an extent understates the combined weight of the challenge made. The charge that deontological judgments (are prone to) respond to morally irrelevant factors is actually best explained by the hypothesis that they are based in emotions, take the form of heuristics, and are by-products of an ancient evolutionary past social world where harming someone could not take an impersonal form. So the best challenge is actually not unrelated to these, in Berker's assessment, lesser challenges. But let us now move on to discuss what he takes to be the best version of the challenge.

9 Deontological judgments as responding to irrelevant factors

Berker says the best version of Greene's challenge to deontology claims that deontological judgments respond to morally irrelevant factors. This is how he states the argument:

> P1 The emotional processing that gives rise to deontological intuitions responds to factors that make a dilemma personal rather than impersonal.
>
> P2 The factors that make a dilemma personal rather than impersonal are morally irrelevant.
>
> C1 So, the emotional processing that gives rise to deontological intuitions responds to factors that are morally irrelevant.
>
> C2 So, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force.[120]

[120] Ibid., p. 321.

Berker is skeptical of this argument and expresses his misgivings in terms of four worries.[121] His first worry is that P1 might not be true. His second worry is that if P1 is true, P2 will be seen as less plausible. The third worry is that C2 does not follow from C1, and finally, his forth and most pressing worry: neuroscience seems to be doing no job in the argument. Let us go over these worries in turn.

The first worry is that P1 might not be true, based on the ambiguity of a dilemma's being "personal":

> Since Greene et al.'s initial characterization of the personal versus impersonal dilemma distinction does not track the gives-rise-to-a-deontological-judgment versus gives-rise-to-a-consequentialist-judgment distinction, it is far from clear that premise P1 is true… [A]ny attempt to precisely characterize the features that give rise to distinctively deontological judgments reintroduces many of the intricacies of the original trolley problem: formulating a principle that distinguishes what separates cases-eliciting-a-deontological-judgment from cases-eliciting-a-consequentialist-judgment is likely to be as difficult as the old problem of formulating a principle that distinguishes the permissible options in trolleylike cases from the impermissible ones… So settling on a fully adequate account of the sorts of features to which deontological judgments are responding is likely to be an extremely difficult, if not impossible, task, and until that task has been completed, we cannot be sure whether P1 is true.[122]

It is true that the characterization of what makes a moral dilemma "personal" or "impersonal" in Greene's terminology has evolved over time. In the 2001 study, coders categorized dilemmas as personal if, and only if, they exhibited the following three features: a) the act foreseeably causes bodily harm to b) a specific, identifiable individual (or group thereof) and c) the harm is not the result of the agent deflecting an existing threat (sloganized as ME HURT YOU). In later studies Greene and his collaborators have tried to isolate the influence of the agent's intention, her spatial proximity to the individual harmed, the role of bodily force etcetera which led them to categorize dilemmas as personal only if the action considered involves the use of *physical force* (as distinct from mechanically operated switches and the like) in a way that causes harm as a means (as distinct from as a side-effect).

[121] Ibid., p. 322.
[122] Ibid., p. 322-4.

An illustration of how this updated version of the personalness factor plays out might be seen if we consider some variations of the footbridge case. In one experiment, different versions were given to different groups. One group was asked about the standard footbridge scenario where the subject can either approve or disapprove of the person's *pushing* the stranger. A second group was given a description were the action under consideration was not pushing the stranger, but *flipping a switch* next to him that would release a trap door underneath him. A third group was asked to consider a scenario were the stranger was *pushed with a pole*, and, finally, a fourth group considered a version were the agent was switching a *trap door from afar*.[123] Answers to the four scenarios cluster into a distinct pattern. Around 30 percent of respondents think it is permissible to push the stranger using either your hands or a pole. Around 70 percent, then, do not support such actions. Around 60 percent of respondents, however, approve of releasing a trap door, either from afar or when standing right next to the stranger. Releasing a trap door is creating a new threat but it does not involve the use of physical force. Using a pole, depending on its length, creates a distance between the agent and the stranger and so is less up close than the case where one stands next to the stranger and turns a switch, but it does involve the use of physical force. Though spatial proximity and personal force often hang together, it seems, then, it is the force factor distinctively that gives rise to a strong emotional reaction.

How does this matter for P1? If P1 is interpreted as saying that *all* deontological judgments derive from intuitions that are caused by emotional processes responding to the personalness factor, then P1 is false. But that is too strong a claim, one that is very unlikely to be true. A more plausible suggestion says that for the *subset* of moral dilemmas that are, in the stipulated sense, personal, deontological intuitions and corresponding judgments derive from emotional processes. These processes respond to features that make a dilemma personal, and not the features that, according deontological ethics, make a given act right or wrong. So, P1, thus understood, stands. Greene does not have to claim, absurdly, that it would not be a deontological judgment to deem as impermissible the release from afar of the footbridge trapdoor. What he does claim is that people will be more willing to make that admittedly deontological judgment than the deontological judgment that it is impermissible to push the stranger in the original footbridge case,

[123] Greene, et al. 2009, p. 365-6.

and that the difference is accounted for not in terms of morally relevant features as outlined by deontological ethics but in terms of emotional processes triggered by the use of personal force.

Let us now turn to, P2, which says "The factors that make a dilemma personal rather than impersonal are morally irrelevant". It is a normative principle, not an empirical prediction or explanation. My first reaction was that it is uncontroversial. As we have seen, philosophical agreement on this principle is what allows us to observe to what extent people in real life seem to place importance on it. But Berker is skeptical of P2 too. This is because that while it may be uncontroversial that the use of personal force itself is not of moral importance, it is "quite another thing to say that whether one has initiated a new threat that brings about serious bodily harm to another individual is a morally irrelevant factor."[124]

This overestimates the importance of the changes to the definition of personalness. In all of the tests people have had a stronger emotional reaction to some of the scenarios than to others. This tendency is a human constant and has nothing to do with how researchers define "personal dilemma". It is still totally kosher deontology to insist that it is impermissible to initiate new threats harming another individual. But, as we have seen, creating a threat is distinct from using force, so this does not really challenge P2.

So what is the disagreement really over? There is a certain nebulousness about the nature of the dispute here, which I think derives from Greene. In some of his writings he comes close to expressing the thought that deontological judgments are a sort of psychological natural kind, or in any event a phenomenon that science discovers for us.[125] But we need to make a division of labor here: what makes a judgment a deontological judgment or a consequentialist judgment is determined by the relation between the content of that judgment and the moral theories we call deontological and consequentialist respectively. Typically, people working in moral philosophy will be the most competent jurors here. What goes on in the brain, what stimuli trigger these judgments etcetera, these questions are, of course, outside the competence of philosophers to answer. Greene seems at places to propose that deontologist philosophers do not really understand what deontology is, just the way members of an isolated tropical community may fail to realize ice is a form of water. But we do not have to get into

[124] Berker, p. 324n.

[125] See Greene 2008.

complicated matters of reference and natural kinds here. The following is sufficient: the best explanation of the failure to separate at a principled level the permissibility of acting in cases like *Push* and *Loop* is that human beings have an innate emotional reaction to personal force, and deontological philosophy has *rationalized* these reactions into articulate principles. The principles have nothing to do with personal force, but it is with the presence or absence of that factor that the perceived plausibility of the principles will vary. This is controversial enough, but seems to me as something we have grounds for accepting.

According to Berker, there is a tension between P1 and P2 such that if the triggers of deontological judgments are characterized in simplistic terms (e.g. responding to degree of personalness), they are easily dismissible as morally irrelevant, but the more complex and nuanced the characterization of what triggers deontological judgments become, the less plausible it will be to dismiss that account as responding to morally irrelevant factors. But, as I said when discussing P1, there is no need to claim that *all* deontological moral judgments respond to morally irrelevant features, such as personalness. It is enough that a sufficiently large and important sub-group of deontological judgments do. And remember, it would be question-begging to assume at this point what features are morally relevant (number of survivors versus intentionally causing harm to an innocent person etcetera). All that is needed is that under none of the theories discussed is personalness per se considered morally relevant.

Berker's third worry is that, even if we grant P2 (that personalness is morally irrelevant), the argument is invalid since C2 does not follow from C1. We may come to conclude that deontology has been given a blow, but this alone does not justify us in assuming that a similar blow might not be in the cards for consequentialism too. For instance, Berker claims, it might be said that consequentialist judgments fail "to respond to morally relevant factors by ignoring the separateness of persons, or by treating people as vats of well-being, or by assuming all value is to-be-promoted, or by making morality incompatible with integrity" leaving us with the same old battle over what moral intuitions are the right ones.[126]

If separateness of persons (in some sense incompatible with consequentialism) is a moral desideratum and if utilitarians do not pay attention to that, then they are indeed mistaken. But the dialectic here is

[126] Berker, p 325.

different. For the suggestion is that we have found a feature, unlike separateness or whatever, that deontologists and consequentialist alike *agree* is morally insignificant. It is because of this shared commitment that Greene and Singer got a foothold in the debate: it is a non-question-begging challenge to the datum that, in certain scenarios, deontological judgments seem more plausible than consequentialist ones even in the absence of a principled way of making sense of the discrepancy. Berker may be right, of course, that an analogous undermining of consequentialism in fact is possible, although the specific routes he mentions would not seem to be plausible candidates, since they are not agreed by consequentialists and deontologists alike to be morally relevant.

In any event, it has to be admitted C2 does not follow from C1. We should also note that Greene does not claim it does either (this is after all Berker's reconstruction of what he takes Greene's challenge to be). The only way C2 would follow from C1 is under the implausible assumption that every factor speaking against deontology automatically entails a corresponding strengthening of consequentialism. Because Greene does not claim C2 follows from C1, and since it clearly does not, a more interesting interpretation, I claim, is to think of C1 as providing an undercutting defeater against certain ways of questioning consequentialism, i.e. ways which, upon empirical inspection, are revealed to rely on the defunct personalness factor. This is more in line with the shape of the debate, and with Greene's intentions. Still, there may be specific direct ways of undermining beliefs favoring utilitarianism too, as well as a more general skepticism towards all moral beliefs. These worries will be the topic of the next chapter.

10 What is added by the empirical work?

Let us now look at Berker's most pressing worry, viz. that the neuroscience does no work in the case against deontology.

We have, Berker notes, three distinctions under consideration:

1) dilemmas that engage emotion processing versus dilemmas that engage "cognitive" processing;
2) dilemmas that elicit deontological judgments versus dilemmas that elicit consequentialist judgments;
3) personal moral dilemmas versus impersonal moral dilemmas.[127]

[127] Ibid., p. 325.

And continues:

> Greene et al.'s dual-process hypothesis posits that the first of these distinctions matches up with the second. In order to experimentally assess this hypothesis, Greene and his colleagues identified the second distinction with the third one, and then directly tested whether the first distinction matches up with the third. But the argument from morally irrelevant factors only depends on Greene et al.'s identification of the second distinction with the third one. Thus the neuroscientific results are beside the point.[128]

Here is how I think this may be answered. We should not *identify* the second distinction with the third. Rather, dilemmas of type 3), i.e. personal dilemmas, *tend to cause* an emotional reaction in respondents, i.e. type 1) dilemmas. That is a scientific finding from, among other sources, *fMRI* experimentation, and so is not something which makes the neuroscience or other empirical findings "beside the point".[129] This emotional reaction in turn, especially when not countered by a response produced in brain structures known to be involved in "cognitive" or "information processing" tasks, will typically cause the respondent to make a deontological moral judgment about what would be the right thing to do. So the work done by neuroscience is that it provides us with a causal account of moral judgment, one that was not known to philosophers from the armchair and for which there is much empirical evidence in addition to neuroscience. The account in question, the dual process theory, says deontological judgments are often the result of features of the case at hand not themselves judged to be morally relevant on the deontological moral theory, such as the presence or absence of personal force. No such similar systematic mismatch between professed features of moral relevance and moral judgment seems to exist in the case of utilitarian judgments, which consistently track the feature judged to be of moral importance on that view: the maximization of value. So, neuroscience and other kinds of psychological (broadly defined) evidence indeed do some work in Greene's work.

11 Restating the challenge

Berker's central expression in the discussion has been if Greene has established if there is or is not any "genuine normative force" to deontological intuitions. This way of phrasing things, I believe, is a little

[128] Ibid, p. 325.
[129] Ibid, p. 325.

unclear. As I mentioned, I am critical of his reconstruction of Greene's view, and I suggest at this point to rephrase Greene's argument, so that it a) is more explicitly one with an *epistemic* conclusion b) allows for assessment in terms of *degree* or "plausibility points" and c) does not rely on an overly ambitious view about the psychology of all deontological judgments. Here is how I suggest we understand Greene's challenge:

P1′ A significant subset of deontological judgments are caused by emotional processes which respond to factors that make a dilemma personal rather than impersonal.

P2′ The factors that make a dilemma personal rather than impersonal are morally irrelevant on both deontological and consequentialist accounts.

C1′ A significant subset of deontological judgments respond to factors that are morally irrelevant on both deontological and consequentialist accounts.

C2′ When criticisms of consequentialist ethics rely on judgments that respond to morally irrelevant factors we may justifiably take that particular criticism as possessing no or minimal evidentiary value.

This, I suggest, is the challenge of the dual process view to deontological ethics. It clearly makes neuroscience (and other kinds of empirical data) do some work. C1 follows from P1 and P2, and these premises stand a good chance of being true. By "significant subset" here I do not necessarily have in mind a number or relative proportion, but rather a set of judgments we take an interest in because they constitute an important conflict line in ethical theory, viz. the debate between consequentialist and deontological ethics. Admittedly, even on this reconstruction, C1′ does not logically imply C2′ but rather puts epistemic pressure on some of the most stubborn traditional criticisms of consequentialism.

12 What are deontological and what are consequentialist judgments?

A lot has been said about the neural correlates of deontological and utilitarian judgments. We know what these are, right? Deontological judgments, in these contexts, are often judgments to the effect that a proposed line of action, while leading to a positive overall outcome in

some sense (more survivors, more wellbeing etcetera), is immoral because it violates some fundamental restriction. Like pushing someone in front of a trolley to save five lives. And utilitarian judgments are judgments that a proposed line of action, while violating some purported restriction, is never the less morally okay because it leads to overall better effects. Like pushing someone in front of a trolley to save five lives.

Frances Kamm and others have voiced the concern that the way the options are labeled in the tests and surveys used is often inaccurate in a way that may make utilitarian judgments seem more appealing than they would be if more accurately expressed.[130] In Greene's first study the question asked after describing a dilemma was if a given course of action (say pushing the stranger on the footbridge was "Appropriate" or "Inappropriate". This is not typically the vocabulary of moral philosophy, but was used to create a seamless transition between moral and nonmoral contexts. But more importantly, the term "appropriate" does not distinguish "permissible" from "obligatory", and the most plausible interpretation is probably that the suggested action is permissible but not obligatory. And, it may be argued, on utilitarianism, refraining from pushing the stranger is *not* permissible, and hence pushing him is obligatory: one does the wrong thing if one refrains. Take *The crying baby dilemma* mentioned above for instance. A group of people are hiding in a basement while enemy soldiers search the neighborhood. If you are found, you will be killed just as all the others already rounded up and instantly killed. Your baby cries, and the sound will soon be noticeable to the enemy soldiers. If nothing else helps, would it be "appropriate" of you to suffocate your baby so that the group will not be detected and killed? Greene suggests saying yes to this is a characteristically utilitarian judgment. And saying no is a characteristically deontological judgment, since although the outcome will be less bad, killing your baby is just wrong.

Here is what Kamm says about this case. The baby will die no matter what you do, and hence is not made worse off by being killed. Also, since it is the baby's cries which would reveal the group's whereabouts, the baby is a so-called innocent threat. Given these strange and appalling circumstances, it is unclear what deontological principle forbids killing this doomed baby.[131] If the child's life depended on your actions or if it was a bystander and not an innocent threat, things

[130] Kamm 2009, p. 339-40.

[131] Ibid, p. 337-8.

would be different according to deontological ethics, but not according to utilitarianism. On utilitarianism, killing a bystander child who is not a threat and who would live if it were not for your action may be morally *obligatory*. If "appropriate" means "merely permissible", that is an incorrect way of describing the utilitarian verdict on the crying baby dilemma. While, according to the deontological ethics proposed by Kamm, killing the child is permissible. So how can an answer, that killing the baby is permissible but not obligatory, which is actually false according to utilitarianism but true according to deontological ethics, be classified as "characteristically" utilitarian?

These are good questions. If the only response options are "appropriate" and "inappropriate", a reasonable interpretation is that the former includes both the prerogative and the obligation to act. But Greene had probably assumed that the crying baby case was one where the deontological judgment was definitely to call the proposed action inappropriate. And if Kamm says he is wrong about that, who can disagree? This kind of inexactness in gauging what it is people are doing when they provide their assessments of what may be done should make us feel less secure in the conclusions drawn by Greene and others. It may be that the net they threw at these matters was, here and there, a bit too coarse. On the other hand it may be argued that the resolution is high enough, for what we are trying to get at are not the precise statements of philosophers but psychological tendencies and how these tendencies give rise to approval or disapproval of major alternatives in the ethical debate. Some parts of our psychology, Greene would say, make us favor a utilitarian way of thinking where action-types are not seen as right or wrong in themselves but right or wrong given the alternatives and consequences, and a corresponding deontological flipside where some action-types are seen as right or wrong in themselves quite distinct from their overall consequences. I think it is fair to say we have seen a measure of that, some lack of finesse notwithstanding. But it is also fair to say that there are many ways of developing a deontological ethics, and not all of them will conserve people's intuitions about drastic cases such as the crying baby. As Kamm phrased it, "nonconsequentialists are not squeamish", and are indeed willing to up close personally use violent force, only under different conditions than the utilitarian.[132] It seems not even Kamm is a preservationist.

[132] Ibid., p. 335.

Foot's, and later on Thomson's, formulation of the trolley problems, and other related dilemmas of causing death and saving lives, caught our attention because of the inconsistency or tension occurring when we say yes in *Switch* and no in *Push*, and then yes again in *Loop*. The attempts to come up with a principled way of accounting for this led us down a route of philosophical contortionism. They could have just said no at the very first junction. Greene suggests people fail to see what needs to be done in *Push* because emotions cloud our minds. But a possible reply to this is that we fail to see the impermissibility of turning the trolley in the firsts case because we are not emotionally triggered. The absence of an emotional response leads people to falsely adopt a utilitarian mindset, or some version of deontology where this choice is permissible. Thomson herself now accepts this view. Her reasoning is as follows. In the *Switch* case, let us introduce a third option: in addition to the track containing the group of five and the side track with the one single person, we now also have a third track – where *you* are at. You said it was permissible to kill one to save five. Would you turn the trolley on yourself – *must* you, morally, do it? Thomson thinks you have no obligation to be that altruistic. And you are not permitted to impose that same cost on someone else instead. The man who thinks he does not have to pay with his own life to save five cannot "decently regard himself as entitled to make someone else pay it."[133]

13 Conclusion

The upshot is that, for decades, even deontologist philosophers managed to forget that "negative duties really are weightier than positive duties".[134] But how could that be? Thomson does not even want to mention Greene, she has instead herself concluded that,

> what seems to vary is at heart this: how drastic an assault on the one the agent has to make in order to bring about thereby that the five live. The more drastic the means the more strikingly abhorrent the agent's proceeding. That I suspect may be due to the fact that the more drastic the means the more striking it is that the agent who proceeds infringes a negative duty to the one.[135]

[133] Thomson 2008, p. 366. She acknowledges the work by MIT grad student Alexander Friedman as instrumental in triggering this change.

[134] Ibid. 353.

[135] Ibid. 374.

I will not more fully detail and assess Thomson's trolley turnaround here.[136] The reason I bring up her new, highly interesting, take on these matters – and the same is true for my brief discussion of Frances Kamm – is, rather, to point to the many options available for us in continuing a discussion from the point of view of moral philosophy even after all of the psychological, anthropological, and neuroscientific facts are acknowledged. Greene's criticism of deontological ethics is based on the idea that many of its verdicts are rationalizations of emotional responses triggered by features of actions and situations which are, even on the deontological view itself, morally irrelevant. This charge is inapplicable in cases where the position defended rejects the more intuitive response (switching the trolley) or goes against judgments congruent with a strong emotional response (killing the baby). This suggest that there are many ways of working out a deontological ethics, and that this project seems feasible enough even if we come to suspect that many judgments often thought of as characteristically deontological originate in emotional responses to morally irrelevant factors. The struggle continues.

[136] For some assessments, see FitzPatrick 2009 and Graham 2017.

3 Coping with Debunking: Ethical Truths of Reason

People have been wary about the possible moral implications of evolutionary theory ever since Darwin published *On the Origin of Species* in 1859.[137] Many of us know the (regrettably apocryphal) story of how the wife of the Bishop of Worcester is supposed to have said, upon learning of the book's claims, "Descended from the apes! My dear, let us hope that it is not true, but if it is, let us pray that it will not become generally known". The initial worry was that the non-divine origin of humans did not sit well with the teachings of Christian ethics. In the last decades, however, the biological challenge to morality has often been thought to consist in the fear that evolution has produced our capacity for moral judgment in a way that is adaptive rather than accurate.[138] The fashionable term for this is that evolutionary considerations might *debunk* our moral claims.

Peter Singer was the first philosopher to write about Joshua Greene's neuroscientific research, and he took it to support consequentialist modes of thinking in ethics over deontological ones.[139] At that time, it seemed to him that neuroscientific and evolutionary considerations were particularly compromising for deontological moral judgments, leaving consequentialist judgments intact or at least relatively less affected (see chapter 3). Since then a more general onslaught on ethics based in evolutionary psychological considerations wider than Greene's dual process theory has gained prominence in the field, notably through the works of Sharon Street and Richard Joyce. Singer never intended to throw out the baby with the bathwater, so he had to return, with co-author Katarzyna de Lazari-Radek, for some damage-control, seeking again to examine the notion that bringing evolutionary psychology into the picture will undermine some but not all ethical

[137] For an overview, see Ruse & Richards (eds.) 2017.

[138] An early but still contemporary example is Ruse 1985. More recently Joyce 2006 and Street 2006 are probably the most prominent exponents of this line of challenge (more on Street shortly). There are, of course, also suggestions that evolutionary theory in some way or other *vindicates* morality, providing it with some external ground or even first order moral principles. See eg. Sterelny & Fraser 2017.

[139] Singer 2005.

views.[140] The aim of this chapter is to examine whether or not that rescue mission succeeded. It was a very ambitious project, and so a degree of skepticism towards it is to be expected.

1 Evolutionary debunking

The starting assumption is the idea that the causal origin of a belief may affect how justified a person is in holding the belief. If I open my fridge and carefully inspect the interior and observe that it is completely empty save for an old ketchup bottle, I am justified in claiming "We have no milk", whereas if I come to believe the same thing based, not on looking in the fridge, but on the testimony of my old demented uncle, clearly my justification is weaker. When I look outside and notice the sun is shining, this gives me grounds for accepting that the sun is in fact shining. Barring extravagant circumstances, we take such observations to be reliable indicators of what the facts are. But suppose I now learn that someone is actually blocking the visibility right outside my window and is instead projecting a perfect image of a shining sun. Absent other information on the matter, this new evidence undermines, i.e. debunks, any belief I have about what the weather is like outside. Since I realize that the projection would make me form the belief that the sun is shining even if it was not, I should suspend belief about the current weather. It might be, of course, that behind the projection of a shining sun, the sun is actually shining. But it should also be clear that I would have no grounds for that belief, since I would hold it regardless of the actual weather. The (causal) link we typically require between the external world and my observation has been severed. So a debunking account is not intended to establish the falsity of some claim, but instead removes the positive grounds for accepting said claim.

Debunking explanations such as these are pretty unremarkable. How do they relate to matters of ethics? Let us turn to our brains. The human brain is a physical organ and its structure and function are the result of millions of years of natural selection. Our capacity to "moralize", that is our propensity to evaluate behavior, characters, situations and states of affairs from a moral point of view, is made possible by the brain. Why this capacity?[141] Presumably because it helped our ancestors survive and procreate. In order for the capacity to have this

[140] Lazari-Radek & Singer 2012, and the 2014 book by the same authors.

[141] See chapter 1.

function, do we also need to suppose it gives us an accurate representation of a mind-independent moral reality? The answer is not obvious, since we have no grounds for assuming that a capacity to make moral judgment is adaptive only insofar as it is truthful. Consider, by analogy, our vision. This capacity has evolved because, again, it has helped our ancestors to survive and procreate. In order for the capacity to have this function, do we also need to suppose it gives us an accurate representation of a mind-independent world? This time, the answer is obvious. The visual stimuli we receive are useful for survival and procreation *because* by-and-large they accurately transfer information about surrounding dangers, food-sources, partners etcetera. For many of our visual experiences the best explanation of their occurrence postulates the object of the experience as a mind-independent fact. The challenge is: can something similar be said for moral judgments? If no positive answer is available, the reliability of moral judgments is threatened.

2 Sharon Street's challenge

In her 2006 paper "A Darwinian Dilemma for Realist Theories of Value" Sharon Street stated the challenge in the following way:

> Evolutionary forces have played a tremendous role in shaping the content of human evaluative attitudes. The challenge for realist theories of value is to explain the relation between these evolutionary influences on our evaluative attitudes, on the one hand, and the independent evaluative truths that realism posits, on the other. Realism, I argue, can give no satisfactory account of this relation. On the one hand, the realist may claim that there is *no* relation between evolutionary influences on our evaluative attitudes and independent evaluative truths. But this claim leads to the implausible skeptical result that most of our evaluative judgments are off track due to the distorting pressure of Darwinian forces. The realist's other option is to claim that there *is* a relation between evolutionary influences and independent evaluative truths, namely that natural selection favored ancestors who were able to grasp those truths. But this account, I argue, is unacceptable on scientific grounds.[142]

So what are evolved social apes like us likely to think ethics demand, according to Street? Here are two examples: that a) We have greater obligations to help our own children than we do to help complete strangers, and b) The fact that someone has treated one well is a reason

[142] Street 2006, p. 109.

to treat that person well in return.[143] It is obvious why, on evolutionary grounds, we would come to believe we have special duties toward our children. Not being prone to think that, or being prone to think the opposite, would produce fewer surviving offspring. Or rather, the mechanism is this: because we and other mammals are naturally disposed to care for our offspring and place a primacy on their interests, once morality evolves we will be disposed to accept moral judgments conducive to these tendencies. So the crux here is that it is the very natural appeal of these positions that spell the dismantling of our confidence in them. For, as with that sunny projection right outside my window, we would hold those beliefs even if they were not true. Russ Shafer-Landau has usefully labeled the notion that, due to evolutionary forces, our moral faculties are more disposed to accept beliefs with certain propositional contents rather than others "*doxastically discriminating*".[144] This bias undermines, or so one might fear, our grounds for thinking our beliefs are responsive to how the purported moral facts are, which in turn should lower our confidence in them.[145]

3 Accounting for miracles

If our moral beliefs have been shaped independently of any moral facts it would be something of a miracle if any one of them were true. Like, in Street's words, arriving at the shores of Bermuda after setting out for the islands while "letting the course of your boat be determined by the wind and tides".[146] Is there a better way to navigate this? In response to Street's challenge, a variety of coping strategies have been formulated. In the remainder of this chapter, I would like to pay special attention to a suggestion made by Katarzyna de Lazari-Radek and Peter Singer.

In the course of writing a book on Henry Sidgwick's philosophy and how his thinking relates to themes in contemporary debates in

[143] Ibid. In the terms presented in chapter 1, Street (like myself) is a strong nativist, i.e. she holds our natural tendency to make moral judgments is skewed towards judgments of a certain sort, namely those that would favor behavior that by and large are conducive to inclusive fitness.

[144] Shafer-Landau 2012, p. 4.

[145] The debunking thesis is variously expressed as the charge that our moral judgments are *insensitive* to the moral facts, or that we can *explain* our judgments without references to them being true, or that any successful correspondence between moral judgment and moral facts is just a *coincidence* or that our moral judgments are *unreliable*. For discussions, see e.g. Enoch 2010 and Srinivasan 2015.

[146] Street, op.cit., page 121.

moral philosophy, Lazari-Radek and Singer came up with a nifty response to Street which they claim salvages the objectivity of ethics in a way that is not vulnerable to evolutionary debunking.[147] Lazari-Radek and Singer first take us through another debate, which much troubled Sidgwick: the relation between the demands of ethics (on his view, universal benevolence) and the demands of rational egoism (individual prudence). Sidgwick famously ended his *The Methods of Ethics* in a tone of despair, noting there is an "ultimate and fundamental contradiction in our apparent intuitions of what is Reasonable in conduct," concerning the individual's reasons to promote the good of oneself and the good of others.[148] Having spent hundreds of pages establishing universal benevolence as the correct ethical position, Sidgwick still thought there exists no decisive reason to hold ethical demands as more rationally compelling than self-interested demands. But, Lazari-Radek and Singer argue, Sidgwick was overly pessimistic about the irreconcilability of the conflict between universal benevolence and prudence. Coming to see this then gives us, so they claim, ammunition that can be used to fight back on Street's evolutionarily inspired challenge to moral realism.

Lazari-Radek and Singer grant, and claim Sidgwick would too, that Street is right in claiming that many (perhaps most) of our moral judgments are unjustified.[149] They also grant that it is hard to defend the notion that evolution might have equipped us with some psychological capacity that would make us identify true moral judgments as such. But, they argue, evolution has equipped us with a more general reasoning ability, which was not itself selected for because it helped us identify moral truths, but is possible to put into such use once in place:

> A plausible explanation of the existence of these capacities is that the ability to reason comes as a package that could not be economically divided by evolutionary pressures. Either we have a capacity to reason that includes the capacity to do advanced physics and mathematics and to grasp objective moral truths, or we have a much more limited capacity to reason that lacks not only these abilities but others that confer an overriding evolutionary advantage. If reason is a unity

[147] Lazari-Radek & Singer 2014. I concentrate in this chapter mostly on their 2012 paper.

[148] Sidgwick 1907, p. 508.

[149] Lazari-Radek & Singer 2012, p. 13.

> of this kind, having the package would have been more conducive to survival and reproduction than not having it.[150]

It is easy to see why there would be selective pressures causing our ancestors to develop mental skills that helped them plan and to form complex cooperative hunting schemes, anticipate the behavior of partners, prey and predators, perform basic arithmetic tasks, build shelters, gossip, assess threats, build weapons and so forth. Although we may have some modular machinery for specific tasks like that, it seems parsimonious to postulate some degree of general, domain-neutral emergent capacity for rationality as well. This domain-general rational capacity, Lazari-Radek and Singer believe, may help us identify moral truths that are not vulnerable to debunking, statements which are true and evidently so.

4 Self-evidence in ethics

There is a long tradition, of which Sidgwick is a part, in moral philosophy of thinking that some ethical statements are *self-evident*. How are we to understand this attribute? John Locke, though not himself a proponent of the view that there are self-evident ethical statements, beautifully characterized a self-evident proposition as one that "carries its own light and evidence with it, and needs no other proof: he that understands the terms, assents to it for its own sake"[151] Eighteenth century philosopher Richard Price said in a more psychological vein that "There are undoubtedly a variety of moral principles and maxims, which, to gain assent, need only to be understood."[152] W.D. Ross gave a more epistemological emphasis and defined a self-evident statement as being "evident without any need of proof, or of evidence beyond itself".[153] C.D. Broad said that there are some statements "such that a rational being of sufficient insight and intelligence could see it to be true by merely inspecting it and reflecting on its terms and their mode of combination".[154]

[150] Ibid., p. 17.

[151] Locke 1690, p. 32. As I understand Locke he rejected the notion that ethical principles are self-evident, claiming "there cannot any one moral rule be proposed whereof a man may not justly demand a reason" (ibid). For a concise presentation of Locke's views, see Schneewind 1994.

[152] Price 1787, p. 284.

[153] Ross 1930, p. 29.

[154] Broad 1936, p. 102-3.

A more recent statement comes from Robert Audi, who says that self-evident propositions are "truths such that (a) adequately understanding them is sufficient justification for believing them … and (b) believing them on the basis of adequately understanding them entails knowing them"[155] Some intuitionists state the view in terms of a permissible *prima facie* credence to what seems plausible, as when Michael Huemer writes, "If it seems to *S* that *p*, then, in the absence of defeaters, *S* thereby has at least some degree of justification for believing *p*", and that beliefs thus formed are justified "unless countervailing evidence should arise that is strong enough to defeat the initial presumption in their favor". [156]

All of these varying statements belong in a tradition called ethical intuitionism. This position comprises both a metaphysical thesis and an epistemological. The metaphysical thesis is a non-naturalist form of moral realism according to which there exists moral properties and these properties are not identical, nor reducible, to any natural property or properties. The epistemological thesis is the above stated notion that some basic moral propositions are self-evident and need no further support. Many, of course, are skeptical of the metaphysical thesis and so defend either moral nihilism or various forms of naturalist realism.[157] In this context my primary concern is with the *epistemological* thesis, i.e. the view that belief in a proposition based on properly understanding and reflecting on it entails, with some reservations to be spelled out shortly, knowledge. There are, then, on this view, some justified, true ethical claims which are able to withstand attacks of the sort offered by Sharon Street.

5 Debunkproofing moral principles

Sidgwick formulated a four-step test by which we could separate the only seemingly self-evident statements from the genuinely self-evident. Propositions with a claim to self-evidence, Sidgwick suggested, must meet the following conditions with "complete fulfillment":

a. The proposition is *clear and precise*.
b. The self-evidence is ascertained by *careful reflection*
c. The propositions accepted as self-evident must be *mutually consistent*
d. There is *general acceptance* of the propositions.[158]

[155] Audi 2008, p. 478.
[156] Huemer 2007, p. 30, and Huemer 2005, p. 105, respectively.
[157] Mackie 1977, Olson 2014; Brink 1989, Foot 2001.
[158] Sidgwick 1907, pp. 338-42. My italics.

Lazari-Radek and Singer adopt parts of Sidgwick's litmus test, and adds a non-debunking clause to it. On their suggestion, a moral intuition enjoys "the highest possible degree of reliability" when

1. careful reflection leads to a conviction of self-evidence;
2. there is independent agreement of other careful thinkers; and
3. there is no plausible explanation of the intuition as the outcome of an evolutionary or other non-truth-tracking process.[159]

We have a lot of seemingly plausible moral intuitions, and the job of moral philosophy is to put them through this updated Sidgwick test. As utilitarians, Lazari-Radek and Singer are of course quite happy to have statements defending family partiality and justice as reciprocity debunked on account of them being plausibly explained as the outcome of an evolutionary or other non-truth-tracking process. But could there be other moral principles, with no link to an evolutionary benefit, that survive Street's challenge? Lazari-Radek and Singer argue that there is still room for some moral statements to remain standing after the philosophical worms of evolutionary debunking have done their job.

In addition to Sidgwick, Lazari-Radek and Singer are inspired by Derek Parfit and his turn in later works to a form of non-naturalist moral realism conjoined with the epistemological thesis that some statements are self-evident. According to Parfit an example of a self-evident proposition on a non-moral matter would be:

W: No statement can be both wholly true and wholly false.[160]

And, as an example of self-evident moral statement he provides:

T: Torturing children merely for fun is morally wrong.[161]

For Lazari-Radek and Singer, the statement at the center of their attention is Sidgwick's point-of-view-of-the-universe idea, which says:

U: The good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other.[162]

[159] Lazari-Radek & Singer 2012, p. 26.
[160] Parfit 2011b, p. 544.
[161] Ibid.
[162] Sidgwick 1907, p. 382.

6 *Agreement of other careful thinkers*

Lazari-Radek and Singer actually say very little on what self-evidence is or how to tell when we face a case of it. This may not be a particular weakness of their position, since they stand on the shoulders of Sidgwick and others after him who have developed views on that particular part. Instead, they focus on how the idea that some ethical claims are self-evident interacts with anthropological, biological and other empirical data. According to Lazari-Radek and Singer there is a tendency for the most widely adopted religions to espouse an ethical commitment nearly identical to U. We can see this thought expressed, they claim, in the Jewish and Christian versions of the Golden Rule, and similar ideas are present in the Confucian, Hindu, and Buddhist traditions.[163] That the originators of these various traditions, in most cases isolated from one another, would converge over accepting something like U is best explained, Lazari-Radek and Singer suggest, by its being a truth of reason: "Like our ability to do higher mathematics, it can most plausibly be explained as the outcome of our capacity to reason."[164]

I think it is safe to say, my limited scholarly reach notwithstanding, that all these traditions are actually not in agreement, and that much of the purported agreement is at least partly sustained simply because the target claims are very imprecise. One could find elements of more or less any moral position in those ancient schools. We could grant that all ethical traditions broadly speaking are about the proper restraint of self-interest and the promotion of a wider good. But that is just not distinct enough. We need more precisely formulated positions before we can say we have an agreement or not.

Though we probably overestimate the occurrence and depth of genuine and fundamental moral disagreement in the world, if we turn our focus to present-day moral philosophers ("other careful thinkers") it seems clear the disagreements in normative ethics cannot entirely be explained by differences in empirical assumptions, cultural background or cognitive shortcomings on the part of one side of a divide in opinion. In fact, the agreement criterion is more suitably directed *against* the notion that there are some moral intuitions that are truths of reason. For if there were, we would expect to see reasonable people

[163] Lazari-Radek & Singer 2012, p. 26.

[164] Ibid. In addition to these tendencies toward impartiality and inclusion, anthropological data would also tell a story of tribalism and differential moral status, a tendency in ethics not even the most careful thinkers have been immune to.

converge, but they do not. Parfit worried a lot over peer disagreement and more or less wrote three thick volumes to finally make everyone agree, because

> If we had strong reasons to believe that, even in ideal conditions, we and others would have deeply conflicting normative beliefs, it would be hard to defend the view that we have the intuitive ability to recognize some normative truths. We would have to believe that, when we disagree with others, it is only we who can recognize such truths. But if many other people, even in ideal conditions, could not recognize such truths, we could not rationally believe that we have this ability. How could *we* be so special? And if none of us could recognize such normative truths, we could not rationally believe that there *are* any such truths.[165]

In some matters, disagreement is a tolerable thing and does not threaten the notion that we are dealing with a robust domain of discourse. Some participants may simply be in error due to biases or insufficient familiarity with the evidence. Even disagreement among experts is to be expected in areas where data are either very sparse or complex. But disagreement among experts over statements purporting to be *self-evident* is really a warning flag that the statements in question are not self-evident. One may respond that if lack of widespread disagreement is crucial, it would appear this will (over)generalize to other philosophical areas, outside of moral matters. And there may even be a sort of self-undermining quality to this requirement: is it not the case that qualified philosophers disagree what the implications of disagreement are for a given domain of discourse? I do not think these are absurd implications. One may concede that, yes, maybe widespread disagreement among philosophers constitutes to an extent evidence that the domain of discourse in question is not one where self-evidence is to be found. Additionally, all statements, including the notion that the presence of peer disagreement is *prima facie* evidence that we are not dealing with self-evident claims, are to be assessed on a coherentist fashion. Lazari-Radek and Singer, to my mind, pay too little attention to these matters, especially as it is listed as one among three necessary conditions on a maximally reliable ethical intuition. It is probably because their main focus is the specific counter to debunking strategies that they neglect this part. And given that this thesis is on the impact on moral philosophy of research on human behavior I too will turn to that part now, setting aside criticisms of the criteria for self-evidence,

[165] Parfit 2011b, p. 546.

focusing instead on whether their candidate claims plausibly satisfy those criteria, in particular that about evolutionary debunking.

7 No evolutionary explanation

Here is where Lazari-Radek and Singer put down most of their work. They strive to show that the norm of universal benevolence is not vulnerable to evolutionary debunking. We have already seen that moral statements expressing special concern for our offspring or for the value of survival are easy to explain from an evolutionary point of view. That is, given the kind of creatures we are, it is a given we would find moral propositions like that appealing (perhaps even self-evident). But because of this, we can see that our grounds for thinking that they, in addition to being appealing, correspond to any purported moral facts, are weakened. They might be, of course, but the fact that we hold them gives no support to that further assumption, since we would hold them anyway. We may therefore come to suspect that we are not epistemically justified in believing them, or at least that an account of a reliable link between our acceptance of these moral judgments and the alleged moral facts they imply the existence of should be provided.

Lazari-Radek and Singer think there are, though, some ethical claims which can meet this challenge; in particular Sidgwick's principle which states that the good of one individual is of no more importance, from the point of view of the universe, than the good of another individual. This principle, Lazari-Radek and Singer think, is not fobbed off upon us by evolutionary forces, but is insulated from such undermining processes by being instead the product of reason. Lazari-Radek and Singer continue:

> Street correctly points out that a specific capacity for recognizing moral truths would not increase our reproductive success. But a capacity to reason would tend to increase our reproductive success. It may be that having a capacity to reason involves more than an ability to make valid inferences from premises to conclusions. It may include the ability to recognize and reject capricious or arbitrary grounds for drawing distinctions and to understand self-evident moral truths—what Sidgwick referred to as "rational intuition." In other words, we might have become reasoning beings because that enabled us to solve a variety of problems that would otherwise have hampered our survival, but once we are capable of reasoning, we may be unable to avoid recognizing and discovering some truths that do not aid our survival. That can be said about some complicated truths of mathematics or physics. [...] Either we have a capacity to reason that includes the capacity to do advanced physics and mathematics and to

> grasp objective moral truths, or we have a much more limited capacity to reason that lacks not only these abilities but others that confer an overriding evolutionary advantage. If reason is a unity of this kind, having the package would have been more conducive to survival and reproduction than not having it.[166]

It seems very plausible that evolution has provided us with a capacity to reason which can be employed in various domains and for many different kinds of task, and also that this general capacity can by adaptive even if it also may allow the holders of it to reach conclusions which at the individual level are not conducive to their fitness. This alone, of course, does not show that there are ethical truths of reason, only that we are the kind of thinkers who may come to discover them.

I think there are several problems with Lazari-Radek and Singer's acceptance of Sidgwick's principle as a non-debunkable truth of reason. Look again at the statement Lazari-Radek and Singer want to claim is self-evident and non-debunkable:

> U: The good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other.

How is this principle to be understood? The most straightforward interpretation is that, seen from an impartial perspective, the significance of something's being the good of this rather than that individual vanishes. Alternatively, if the world is made better by me being happier (other things being equal), then it would also be made better by you being happier (other things being equal). So, from an impartial point of view, this principle tells us, my well-being is equally important as your well-being. Could we need any evidence for that? Could anyone seriously believe, not just that his happiness is more important for him than the happiness of others is for him, but that his happiness is more important *simpliciter*; that others, in not realizing this and confusedly pursue their own good, are guilty of an error of judgment? Perhaps, then, this idea is about as self-evident as they come in ethics.

I grant that the view is not trivial, and it is evaluative, but, I submit, it is unclear if the principle holds any action-guidance. In short it is unclear if it is an ethical view at all since the deontic implications of the view that the good of one compared to another is equally im-

[166] Lazari-Radek and Singer 2012, p. 16-17.

portant, when viewed impartially, are indeterminate. Most philosophers who defend agent-relative reasons would probably agree that from the point of view of the *universe* the good of one is of equal importance as the good of another. Problem is, these philosophers think it makes all the difference that human beings typically do not occupy the point of view of the universe. And when one looks a little closer at Lazari-Radek and Singer's discussion it is actually clear that they are laboring with three interconnected theses, not just the point-of-view-of-the universe idea. In addition to U, we also have

R: "As a rational being I am bound to aim at good generally – so far as it is attainable by my efforts – not merely at a particular part of it".[167]

As well as

E: "Each one is morally bound to regard the good of any other individual as much as his own, except in so far as he judges it to be less, when impartially viewed, or less certainly knowable or attainable by him"[168]

The relationship between U, R, and E is not altogether clear. A plausible suggestion is that R, properly understood, implies E. It is also noteworthy that R and E, in contrast to U, are deontic views, not (merely) evaluative. A link between these principles is provided by Sidgwick's treatment of rational self-interest. In that discussion R serves to justify why we should not discount the value of our future selves merely because they are distant in *time*, something that peculiarly would suggest some moments of our own existence have greater significance than others.

Aiming at the good generally is a helpful way of seeing that different moments of our lives have, and should be given, the same value. And just as it is a sign of irrationality to give intrinsic weight to *when* a good state occurs, it might be thought that there is something irrational with assigning weight to to *whom* the good accrues. But there is an obvious answer to this latter challenge, which is different from what can be said to the discounter of his future self. When someone acts or rea-

[167] Quoted from Sidgwick, Lazari-Radek and Singer 2012, p. 24.
[168] Lazari-Radek and Singer 2012, p. 17.

sons in a way implying different moments of their life are given different weight it is a plausible move to say "remember, it will still be happening *to you*". Obviously, that prod is unavailable when the good in question has to be allocated to one individual rather than another, not to one and the same individual but at different moments of their life. Whether this special concern for ourselves is rationally defensible or not, and how this requirement relates to other normative concerns is well-trodden terrain in moral philosophy. And while utilitarians are attracted to the ideal that both the when and the who are irrelevant, it might not be self-evident that "Each one is morally bound to regard the good of any other individual as much as his own" just because prudential time-partiality is irrational.

When Lazari-Radek and Singer assess "universal benevolence", they have the point-of-the-universe-thought in mind and, as careful thinkers, arrive at a conviction of self-evidence. Then, when moving to the criterion of not being debunkable by evolutionary considerations, they seem rather to have E or R in mind. But the interesting question is if this *package* of views – evaluative *and* deontic – is both self-evident *and* not a plausible candidate for evolutionary debunking on the grounds that acceptance of it was adaptive. My contention is that the more other-regarding we make it, the less vulnerable to evolutionary debunking it becomes (a good thing), while at the same time seeming to careful thinkers less and less self-evident (the not so good flipside).

We can see the dialectic more clearly by transforming the package of Sidgwick's three rather lofty statements into the following two:

CG: *Constancy of goodness*: if two individuals have the same well-being goods, then their goods have the same impartial value.

PG: *Promotion of goodness*: each agent has a moral obligation to promote good universally, adopting the outsider's perspective on his or her own good relative to the good of others.

CG and PG are separate claims, and CG does not imply PG, and so one can be self-evident without the other being it too. In fact, PG is judged by many as implausible, and even Parfit agreed with Sidgwick and disagreed with Lazari-Radek and Singer in claiming that when there is a conflict of rational self-interest and impartial demands, while the more impartial option is permissible it is not rationally required.[169]

[169] Parfit 2011a, p. 143.

8 *Is what survives empty?*

Lazari-Radek and Singer argue we cannot explain the seeming attractiveness of universal benevolence as the product of evolution. However, in order for benevolence to be a substantive moral view, a view of moral goodness as benevolence need to be spelled out. Singer nowadays adheres to the classic hedonist account of wellbeing. Hedonism and the desire-fulfillment view remain the major candidates in the debate over what makes someone's life go best, what wellbeing ultimately is, among philosophers combining consequentialism with an axiology to form utilitarianism.

Guy Kahane has pointed out that our attraction to these views can be given an evolutionary explanation too.[170] Presumably humans and other sentient animals experience pain in response to certain stimuli, because beings in the past who were so equipped stood a better chance at avoiding certain harmful stimuli, and thus stood a better chance at surviving and reproducing. It is no wonder we would come to think of pain as *bad* and pleasure as *good*, and the avoidance of pain as good etcetera. So axiological beliefs about what has value seem to be open to evolutionary debunking just the way more general theories in ethics: "is this a judgment we make because we have a biological disposition to make it, or is there an independent advantage to this view such that any rational being would make that judgment?" For all we know, the truth about prudential value may be very far from what evolution brought us to *think* is prudentially valuable. Perhaps, Kahane suggests, what is ultimately valuable is "ascetic contemplation of deep philosophical truths" or why not a "Nietzschean perfectionist aestheticism (which might even revel in pain)".[171] How can Lazari-Radek and Singer prefer hedonism to such views, while maintaining any normative belief we can explain as the result of an evolutionary process is to be distrusted?

So, what idea of "good" might be non-debunkable and still recognizably utilitarian? If no candidate is found, the victory for universal benevolence seems pyrrhic and impotent. Lazari-Radek and Singer respond that "if no theory of well-being or intrinsic value were immune to a debunking explanation, this would show only that no theory could be preferred over others on the ground that it alone cannot be debunked. It could not show that no theory of well-being is true."[172]

[170] Kahane 2011 and Kahane 2014.

[171] Kahane 2014, p. 334.

[172] Lazari-Radek and Singer 2012, p. 28.

While I am more sympathetic than Kahane to this notion of mutual destruction as simply levelling the playing field, this of course is not a positive basis for accepting hedonism or any other view of prudential value. Still, establishing that wellbeing – whatever it amounts to more precisely – is something morality demands we promote universally is no small thing.[173] But what about the consequentialist – the to-be-promoted – part, is it insulated from debunking?

9 Can universal benevolence be debunked too?

When Lazari-Radek and Singer attempt to identify ethical claims which can be given an evolutionary explanation, they have in mind positions that promote or accept egoism or partiality, reciprocity, or loyalty to a limited group. I agree it is plausible to assume we are equipped with emotional and other mental dispositions that would make us attracted to such views. Our acceptance of them, then, is vulnerable to debunking since we have no way of telling if their appeal comes from, as it were, within us, or from the accuracy of the view as such. But Lazari-Radek and Singer are pretty confident their impartial principle cannot plausibly be explained in a similar fashion, since acceptance of it would appear to make individuals behave in ways that are detrimental to their evolutionary success.

But this assumes, though, a rather tight match between overt behavior and acceptance of norms. What does seem impossible to expect is a mechanism which causes humans to *behave* in a universally benevolent manner, paying neither more nor less attention to the wellbeing of themselves or their offspring compared to any other sentient being. But a mechanism which causes humans to *believe* in universal benevolence, or a mechanism which causes humans to *signal a belief* in universal benevolence, do not seem to be unlikely evolutionary products of a highly social, intelligent animal who in a lifetime encounters a couple of hundreds of individuals and is often asked to justify its behavior towards others.

There is a complex story to be told here. On the one hand utilitarian or impartialist ways to think about ethics may be seen as a form of virtue signaling: being a moral, generous, good person. And we may believe that that in turn conferred an advantage to the individual in

[173] Jaquet 2018 defends Lazari-Radek and Singer against Kahane's point, arguing that adopting a subjectivist account of wellbeing, such as the desire view, would provide immunity to evolutionary debunking. MacAskill, Mogensen, and Ord argue that, in practical deliberation, utilitarianism is again less susceptible to debunking.

our ancestral past (as today).[174] And if we also take on board the idea that this may even involve a degree of self-deception such that the person sees himself as acting on impartialist concerns but in fact is not, that person can have the social cake and eat it too.[175] These aspects would be able to explain, from an evolutionary point of view why tendencies to think impartially might have been adaptive even though that seemed counterintuitive at first glance. If that is so, aspects that make utilitarianism appealing may not be that it rests on self-evident ideas but on tendencies in us which are the result of evolutionary processes. Such a debunking story will necessarily be speculative, but that is hardly a contrast to the already existing stories.

On the other hand, recent research suggests people judge those who employ a utilitarian way of thinking as *less* trustworthy and *less* moral than people who employ a deontological way of thinking.[176] This suggests Lazari-Radek and Singer are correct in claiming utilitarianism is relatively more sheltered from evolutionary debunking, since a tendency to employ a utilitarian mode of thinking on moral matters would be costly not only in the sense that you would act against your own self-interest, but also in the sense that that would not even be considered laudable. If we believe that an important function of ethics was to enhance cooperation and we are informed people trust utilitarians less than they trust deontologists, it seems being a utilitarian is not an advantage in forming cooperative alliances. Again, the winner in this game is whichever position we find it the most *difficult* to come up with a story of how it would be that we came to like it. In any event, this shows that any argument attempting to expose or shelter a moral view from evolutionary debunking cannot just rely on the view considered as a semantic content but on the emotional and behavioral tendencies associated with accepting it – or professing acceptance of it. When we say that a certain type of moral judgment or moral theory is possible to explain or not as the result of an evolutionary process, we need to take onboard all of these factors.

[174] Miller 2001. A shorter treatment of the view is developed in Miller 2007. For philosophical discussions of a related issue, see Tosi and Warmke 2020.

[175] See Trivers 2011.

[176] Everett et al. 2018. Also see Everett, Pizarro, and Crockett 2016 and Montealegre et al.

10 Concluding remarks

One of the difficulties is this debate seems to be that it is hard to tell what can and what cannot easily be given an evolutionary explanation. It is pretty easy to come up with such an account for almost anything. Why do people have tribal views, leading to xenophobia and focus on cooperation within one's group? Well, we evolved from hunter-gatherers living in groups of between 50-150 individuals, all dependent on each other, often with genetic family ties to one another, often in conflict over territory and food with other hunter-gatherers. Why did the Christian faith spread throughout the Mediterranean region in the centuries following the death of Jesus? Well it was the first system of belief which emphasized the equal value of all human beings, making it very attractive to those living under oppression, providing a meaningful and coherent worldview which was easy to proselytize.

There will always be a more or less plausible story of why, psychologically, we are prone to accepts certain views or employ certain ways of thinking or feeling. The question is if, in ethics, these explanations are always debunking, or if there are, as it were, *vindicating* accounts of the genesis of a moral judgment or capacity for moral judgments.[177] In this chapter I have interrogated one suggestion to the effect that there is, and that this account favors a utilitarian position in ethics. I concede that there are some facets of utilitarianism which makes it hard to debunk by offering an evolutionary explanation of an acceptance of it, namely its impartial demandingness and the experimental findings that people are less likely to trust and esteem utilitarians. These features suggest that being prone to accept utilitarianism, or displaying behavior dictated by it, would not have increased the inclusive fitness of individuals so disposed. But there are other features of it where we may suspect that there is an evolutionary account of why we would find at least certain aspects of it appealing. Thinking about ethics in an impersonal way may have started at a point in human history where "impersonal" in practice did only extend to a limited set of individuals. Today we think of "universal benevolence" and utilitarianism as very demanding moral outlooks, but this demandingness, which is reflective of a certain incompatibility with our psychology, is a recent feature of the view, present only because of modern technology. The gist of the outlook became attractive to us at a time

[177] Cf Tersman 2008 and Tersman 2017.

when the horizon was much closer. Obviously there cannot be selective pressure to directly accept or reject a moral *principle* but rather on psychological features that lead us to think and feel in certain ways and ultimately to act in certain ways. Because of this cut between psychological dispositions and general moral principles, there will always be a degree of uncertainty in any inquiry into whether or not the moral principle is debunkable. For there will always be a multitude of psychological features that go into making a principle appealing or repellent to us.

I am skeptical of the notion of self-evidence in ethics. I think it piggybacks on the more familiar and epistemically much secure way it is used in logic and mathematics. I also think it is inevitable that we assess ethical statements by considering how well they fit together with other statements we believe there is good reason to accept, as well as considering the plausibility of what it entails in cases, real and imagined. That means I reject the foundationalism which goes hand in hand with Lazari-Radek and Singer's appeals to self-evident ethical intuitions. Ironically, the one statement I find it the hardest to doubt, as close to self-evident as they come, is Quine's dictum that no single statement is immune from revision in the light of new evidence.[178]

A recurring theme in Singer's writings is that there is a place for reason in ethics, and also that employing reason leads to less tribal, moral inclusive, more universal moral views.[179] I mentioned above that it is only recently that impartialist modes of thinking about ethics became so revolutionarily demanding. The moral community always seemed so small. But what made it expand? According to Singer, the expansion is the result of stepping on the "escalator of reason". Suppose you do not care for the neighboring tribe, and experience only satisfaction or indifference in response to their plight. At the same time, you care a lot about your kids and the other member of *your* tribe. How things fare for them is a deep concern of yours. But what is so special about your people, morally separating them from the others? You realize of course, for members of the other tribe, matters are felt just the inverse way: they care a lot about what happens to them and could not care less about what happens to you. But as soon as we pose the question, "What are the morally relevant differences between *X*

[178] Quine 1951, p. 40.

[179] Singer was an early participant to the discussion on the relevance for moral philosophy of evolutionary considerations on human psychology. See his 1981 book *The Expanding Circle: Ethics and Sociobiology*.

and *Y*, making it proper to assign greater weight to *X* than to *Y*?", we start a process of reason-governed reflection on what ultimately gives someone moral status, and what grounds for exclusion are arbitrary. As is well known, Singer has employed this tool to argue for an expansion of moral concern, to not only other tribes, but to eradicate the moral significance of distance, ethnicity, sex, race, and species.

To some degree this vindicates Lazari-Radek and Singer's position: there is no direct evolutionary debunking explanation of utilitarianism. All that was needed for us to arrive at it was some initial sympathy and a domain-general capacity to reason. As Singer has often returned to in his work, we can employ a domain-general reasoning capacity to assess the merits of suggested ideas about what gives a being or entity moral status or consider the merits of a distinction drawn. Even if we would grant that such an assessment itself is not biased by features of our psychology whose deployment here would not be conducive to finding moral truths, there are still two important caveats. First, the fact that a system of beliefs is coherent and contains no internal inconsistencies is not sufficient to show that the beliefs therein are also *true*. Second, the kind of equality or impartiality which these steps lead to does not separate utilitarianism from other competing moral theories.

Kahane's challenge about what notion of wellbeing is really invoked brings to the fore a familiar conundrum about how to think about ethical objectivity. If ethical properties and ethical theories about them are truly mind-independent, they can be just about anything, implying there is no pre-theoretical higher probability of happiness having positive intrinsic value than pain or ascetic contemplation. On the other hand, we come to ethical inquiry with a set of implicit constraints on what the field is about and what kinds of positions we take seriously; that it concerns ideas about how humans can live together, what we owe one another in terms of respecting or promoting our putative rights or interests etcetera. If we take these brackets away, anything is possible. You could say these constraints, which are the arbitrary results of our evolved psychology, debunk the field. But without them, there is no field.

4 Virtue Ethics, Schmirtue Ethics?

Although it has been with us since ancient Greece, virtue ethics was for a long while pretty dormant and made a comeback to "Modern Moral Philosophy" in 1958 with Elizabeth Anscombe's influential essay of that name.[180] Since then, virtue ethics is seen as a major, theoretically independent, option in ethical theory, alongside the consequentialist and deontological traditions.[181] In the last two decades, however, this ongoing renaissance has been subjected to attacks from philosophers citing experimental data from social psychology to question the virtue ethical approach to moral philosophy. Crudely put, the charge is that the evidence for the common-sense notion that people have, and differ in, character traits is meager or non-existent. Instead, the most powerful explanation of any given individual's behavior, contrary to folk wisdom and our everyday notions of ourselves and people around us, is said to be the situational factors within which we act. If individual differences in character traits play no role, or only a subordinate role, in explaining human behavior, it may seem misguided to place moral assessments' most basic emphasis on them. To the extent that virtue ethics rely on there being character traits, and it seems plausible to think that it does, these result would thus appear to pose a threat to virtue ethics.

In the present chapter I state this so-called situationist challenge to virtue ethics, going through some of the psychological evidence invoked to cast doubt on the theories' descriptive contents. I will describe and assess some of the ways virtue ethics has been defended against this line of attack. In the spirit of this thesis' general biological approach to human psychology, I will then counter the situationist charge, bringing substantial causal and explanatory force back to the person side in this person vs situation debate. For a while there it seemed as though virtue ethics was down for the count, but my contention is that the situationist challenge was too hastily conceived and that the research program on which it was based has all but collapsed in social psychology's ongoing replication crisis.

[180] Anscombe 1958.

[181] On the history, decline, and return of virtue ethics, see Frede 2013 and Chappell 2013.

1 Situating the debate

That human behavior is greatly influenced by the circumstances has been known (or suspected) long before there was a science of psychology. Francis Bacon remarked in a letter to the Earl of Essex that "opportunity makes a thief".[182] And way before him, Plato had Glaucon and Socrates discuss the Ring of Gyges, which makes its bearer invisible. Glaucon says:

> Suppose now that there were two such magic rings, and the just put on one of them and the unjust the other; no man can be imagined to be of such an iron nature that he would stand fast in justice. No man would keep his hands off what was not his own when he could safely take what he liked out of the market, or go into houses and lie with any one at his pleasure, or kill or release from prison whom he would, and in all respects be like a god among men. Then the actions of the just would be as the actions of the unjust; they would both come at last to the same point.[183]

Moving to the modern era (1920s), Hugh Hartshorne and Mark May followed schoolchildren over several years, documenting their behavior with a focus on honesty and deception. To the surprise of many, their studies suggested belief in some children as honest and others as dishonest was largely unwarranted. Finding that a boy does not cheat on an exam even when he could, or that another tells a lie, they observed, were of almost no predictive value when trying to figure out what these individuals would do in similar situations where honesty was a relevant factor. In a phrase that could be the credo of what was to become the "situationist" strand in social psychology, Hartshorne and May concluded that honesty is not an "inner entity" but "a function of the situation".[184]

For the general audience, the power of the situation really sank in as Stanley Milgram's obedience experiments became widely known and discussed, and, soon thereafter, Philip Zimbardo's Stanford Prison Experiment.[185] Oddly, the increasing support for situationism, and the

[182] Bacon 1598, p. 99.

[183] Plato, *The Republic*, Book II, 360b–d. The thought experiment of the ring making the bearer invisible could also be interpreted as arguing for a certain other psychological thesis, viz. egoism.

[184] Hartshorne & May 1928, p. 385.

[185] Milgram 1963 and Milgram 1974. Philip Zimbardo details his Stanford Prison Experiment in Zimbardo 2007. Recently, The Stanford Prison Experiment has come under massive critique, and my assessment must be that we should disregard it as a sham. See Texier 2019.

corresponding skepticism of character traits, in psychology, was contemporaneous with a growing popularity of virtue ethics in the field of philosophy. It was not until the early 1990s that philosopher Owen Flanagan noted that the "entire enterprise of virtue ethics depends on there being individual traits of character which are causally effective in the production of behavior across situations of a kind".[186] Philosophical attention to psychological research of a more critical, indeed destructive, kind came some years later with John Doris' "Persons, Situations, and Virtue Ethics" and Gilbert Harman's "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error", published the following year.[187] From then on, the so-called situationist challenge is one of the standard criticisms of virtue ethics, and one that every proponent of virtue ethics is expected to have something to say about.

One of virtue ethics' main relative advantages has always been its presumably more true-to-life and complex view of human moral psychology: it captures the way people actually think about moral issues. Julia Annas says "a large part of its appeal is the thought that, unlike its competitors, Kantianism and consequentialism, it can give a realistic account of our ethical life."[188] Whereas consequentialism and deontology are preoccupied with formulating criteria of what makes acts morally right or wrong, it is often said that virtue ethics seeks a return to that ancient Athenian question, "What kind of person should I be?"[189] In answering this question, virtue ethics prescribe that we develop certain virtuous character traits, such as generosity, courage or temperance. In so doing it assumes that we may *have* character traits, and that they are evolvable under our control to some degree. Virtue ethics also assumes that human behavior is explicable in terms of individuals' character traits (as opposed to external factors not under the person's control). If this were not so, the prescription to develop certain traits of character would be unrelated to our behavior towards one another and so would not really be intelligible as a *moral* ideal. The question then is: just what assumptions about human psychology are implicit in virtue ethics, and is there reason to believe the truth of these assumptions have been threatened by social psychological research?

[186] Flanagan 1991, p. 282.

[187] Doris 1998 and Harman 1999. Doris' book length treatment is Doris 2002.

[188] Annas 2003, p. 21.

[189] A virtue ethics position may also be about formulating theoretical criteria for morally right action. The most notable version of such a view is Hursthouse 1999.

2 *The problem*

Here are the two questions discussed in this chapter:

1. What empirical claims regarding human psychology is virtue ethics committed to?
2. Has psychological research shown these claims to be false?

Aristotle, in the *Nicomachean Ethics*, describes ethical or moral virtues as the character traits that permit a person to properly adjudicate between two desires or courses of action at opposite extremes (vices of excess or deficiency, respectively).[190] Thus the virtue of courage lies between the vices of cowardliness and rashness, generosity between stinginess and extravagance and so forth. Since character traits relate to our behavior and psychology, studying them seems prima facie an empirical, rather than philosophical, endeavor. If virtues are character traits, then virtue ethics branches into psychology in a way that makes it interesting to find out whether or not that part of the theory holds up to scrutiny. That virtue ethics does make empirical claims is acknowledged within the tradition itself. Here is Alasdair MacIntyre:

> To identify certain actions as manifesting or failing to manifest a virtue or virtues is never only to evaluate; it is also to take the first step towards explaining why those actions rather than some others were performed[191]

Rosalind Hursthouse says something similar:

> Suppose someone were described as having the virtue of honesty. What would we expect them to be like? (...) Most obviously we expect a *reliability* in their actions; they do not lie or cheat or plagiarize or casually pocket other people's possessions. You can rely on them to tell the truth, to give sincere references, to own up to their mistakes, not to pretend to be more knowledgeable than they are; you can buy a used car from them or ask for their opinion with confidence.[192]

[190] *Nicomachean Ethics* $1105^{b}28$-$1106^{a}9$. There are other notions of virtue, e.g. Maria Merritt's Humean (Merritt 2000) and Michael Slote's sentimentalist versions (Slote 2005 and Slote 2007), but a roughly Aristotelian conception is the most common, and this standard view is what is the target of the situationist critique. I should also note that my primary concern are the *moral* virtues, not, say epistemic or intellectual ones.

[191] MacIntyre 1984, p. 199.

[192] Hursthouse 1999, p. 11. Emphasis added. She goes on to say virtue is much more than just this reliable regularity of behavior, and connects with doing the right thing for the right reasons etcetera. We shall return to that later on.

In order to make virtues accessible to empirical study we must link them to *overt behavior*. But, as any virtue ethicist will tell you, virtues are more than just behavior. The virtues are connected to how a person feels, reasons, her perceptions of social situations and the way she deliberates. I will return to these complications, and the issue of whether or not they affect social psychological criticisms directed at virtue ethics. Meanwhile, let us look at how John Doris, the leading voice of the social psychological critique of virtue ethics, portrays the virtue ethical notion of character traits.

In Doris' formulation, "If a person possesses a trait, that person will engage in trait-relevant behaviors in trait-relevant eliciting conditions with markedly above chance probability p".[193] So you cannot tell whether or not I'm a brave person by looking at how I make up my dishes, but there are situations where you could make assumptions like "If he's a brave guy he's going to do X".[194] According to Doris, virtue ethics is committed to a certain view of human psychology, and a view on character traits to be more specific, which he calls *globalism*. Globalism is the object of his attack, and he states it in the following three theses:

a) *Consistency*. Character and personality traits are reliably manifested in trait-relevant behavior across a diversity of trait-relevant eliciting conditions that may vary widely in their conduciveness to the manifestation of the trait in question.
b) *Stability*. Character and personality traits are reliably manifested in trait-relevant behaviors over iterated trials of similar trait-relevant eliciting conditions.
c) *Evaluative integration*. In a given character or personality the occurrence of a trait with a particular evaluative valence is probabilistically related to the occurrence of other traits with similar evaluative valences.[195]

[193] Doris 2002, p. 19. "Character trait" is often abbreviated to "trait". Some such traits might be considered virtues or vices whereas others are thought to be neutral with respect to such categorization. So when I talk of people having or not having a particular virtue, what I mean is a character trait of the appropriate type. Exactly which character traits, or other features of a person's psychology, are virtues need not concern us here.

[194] We might want to allow for exceptions (no one is perfect), adding a probability or "acting in character" clause.

[195] Doris 2002, p. 22.

Globalism, to wit, describes our personality as an "evaluatively integrated association of robust traits".[196] Most of the situationist attacks have been targeting the consistency thesis, which says that if an individual acts honestly or courageously say in a situation where these traits are relevantly at play we should expect that individual to at honestly or courageously at some other kind of situation where again the traits are at play. The stability thesis, albeit in a very narrow or local sense, seems to be accepted by both Doris and Harman. That is, they accept the reality of stable traits within a constrained domain, like being talkative in class or being helpful to family members. The thesis of evaluative integration, which states that the possession in a person of one virtue makes it more likely that she also possesses evaluatively closely related virtues too, is given even less attention. So the center of the situationist challenge is really the consistency thesis. According to Doris and Harman, systematic observation has failed to confirm the behavioral patterns expected by globalism, and in particular the cross-situational consistency part. The upshot is that the psychological facts simply do not match the virtue ethical theory, centered as it is on developing and acting from virtuous character traits. Let us now attend to some of the evidence invoked for this conclusion.

3 The case against robust character traits ("globalism")

Underpinning Doris' challenge to virtue ethics is a large body of experimental work in social psychology. Common to all the studies is the conclusion that people's behavior seems overwhelmingly influenced, if not determined, by variables *outside of the person*. In a nutshell, it is the *situation* the individual acts within, not the alleged *character* he or she possesses, that explains their behavior. This view – the antithesis to globalism – has become known as *situationism*.

Let us now look at some of the evidence in support of that second premise. I will focus on four landmark studies, though there are many others with seemingly similar implications.[197] My account will be purely descriptive so readers familiar with the research may skip to section 4.

[196] Ibid., p. 23.

[197] Consult Doris 2002 or Ross & Nisbett 2011 for accounts of further such experiments. Other overviews and discussions include Alfano 2013 as well as three books by Christian Miller: Miller 2013, Miller 2014, and Miller 2017. Also volume 5 in the *Moral Psychology* series edited by Walter Sinnott-Armstrong (this one on virtue and character was co-edited with Miller). A concise recent overview can be found in Miller 2020.

3.1 Help for a dime

People exiting a phone booth witness a woman who accidentally drops a folder of papers in front of them. Who stops to help collect the scattered papers and who rushes on? For one group of callers, a dime was placed in the coin return slot of the phone; for the other group, there was no bonus dime. Here are the results:

	Helped	*Did not help*[198]
Dime	14	2
No dime	1	24

3.2 Obedience to authority

Stanley Milgram drew together people from different walks of life to participate in an experiment studying the effects of punishment on learning and memory. Arriving in pairs to the laboratory and Yale University, a rigged lottery assigned one to be "learner" and the other to be "teacher". The learner is in fact, unbeknownst to the "teacher", a confederate of the experimenter. The assigned teacher witnessed as the learner, a man in his fifties wearing a white shirt and tie, was seated in an "electric chair", his hands strapped to the armrests and wires attaching his fingers. The teacher is then taken to an adjacent room where he can hear the learner on an intercom radio. The teacher is instructed to read groups of words, and the learner is to repeat them to the best of his recollection. If the learner's answer is incorrect, the teacher is to administer an electric shock to the learner. In front of the teacher is a "shock generator". It has an instrument panel with thirty horizontally placed switches, each of which is clearly labeled with a voltage designation ranging from 14 to 450 volts. In groups of four from left to right, the switches are arranged in the following categories: "Slight Shock", "Moderate Shock", "Strong Shock", "Very Strong Shock", "Intense Shock", "Extreme Intensity Shock", "Danger: Severe Shock". The two switches after this last designation were simply marked "XXX".

As the learner made mistakes, the experimenter instructed the teacher to give ever stronger shocks to the learner. At 75 volts, the learner started moaning and objected to the pains inflicted. At 150 volts, he stated that he could no longer endure the pain, demanded to be released and complained that his heart was bothering him. Beyond 200 volts he screamed in pain and reiterated his request to discontinue

[198] Isen & Levin 1972, p. 387.

the experiment ("Let me out! Let me out!") The screams got more and more hysterical until reaching 345 volts, after which there was only silence. According to the experimenter, the teacher was to continue administering shocks since no answer is a wrong answer. The learner in fact didn't receive any shocks at all, and his sounds of pain and protests were tape recordings prepared in advance.

A number of alterations to the basic experimental idea were tested to control for the relative influence of different variables. For instance, obedience was even higher when no verbal feedback of pain was heard from the learner. It was lesser when the teacher had to physically force the learner to hold his hand to a metal plate in order to receive the ever-stronger shocks (still, in this condition, 30 percent of subjects were fully obedient, i.e. administered the maximum shock of 450 volts). To test the alternative hypothesis that the results were not necessarily obedience to authority but rather an example of the aggression (these) people carry inside, in one variation of the experiment the teachers were able to choose the voltage themselves. Very few administered shocks stronger than 75 volts, the point where the learner for the first time indicated that the shocks started to bother him.[199]

Here are some further statistics from Milgram's studies. In the very first version the learner is not visible for the teacher, and there is no voice-feedback. However, at 300 volts the walls vibrate as the learner pounds in protest. In this version of the experiment, *all participants administered what they thought to be 345 volts of electric shock to an innocent person strapped down in a neighboring room, and 65 percent of them kept giving the shock until the maximum of 450 volts was reached*. When the learner's vocal protests were vividly heard the percentage of subjects fully obedient to the experimenter dropped to 62.5 percent. When the subject was sitting next to the learner, who protests, screams and pleads from 150 volts and upwards, full obedience dropped to 40 percent, with 25 percent of the subject dropping out at 150 volts. When subjects had to physically press the protesting subject's hand to a shock plate, 60 percent of subjects administered 180 volts or more, and 30 percent were fully obedient, forcing the desperately screaming subject's hand to the shock plate 22 times from the first protests at 150 volts. In subsequent versions, women turned out to be no less obedient than men, though they often showed more distress during the experiment.

[199] Milgram 1974 pp. 71-3.

3.3 Clerics in a hurry

40 Divinity students at Princeton Theological Seminary were asked to participate in an study concerned with their capacity to quickly improvise a public speech. Experimenters told them to walk to another building to give a lecture. Half of the subjects were instructed to talk about job opportunities for divinity students after graduation, and the other half were instructed to discuss the parable of the Good Samaritan, which as you recall highlights the moral responsibility to help people in urgent need even if one has other commitments. In addition to this variable, subjects were told that they were either already late and had to hurry, that they had just enough time or that they had some extra minutes.

On their way to the other building, the subjects (walking alone) passed a man who slumped over against a wall while coughing and groaning in distress. The question was: how many would stop to ask if he needed help, and what were the influence of the variables content of speech and degree of hurry relative to that behavior? Here are the results:

Degree of hurry	Low	Medium	High
Percentage offering help	63	45	10

Degree of hurry turned out to be the only factor correlated to helping behavior; subjects preparing to talk about the Good Samaritan acted no differently from subjects preparing to talk about career opportunities.[200]

3.4 Bystander effect on helping behavior ("Lady in distress")

People sitting in a room suddenly hear a loud bang from bookshelves collapsing followed by the sounds of a woman screaming in pain in an adjoining room. If the subject was sitting alone, they stopped what they were doing and tried to help in 70 percent of cases. If the subject was sitting with a confederate of the experimenter who did nothing, subjects initiated helping behavior in just 7 percent of cases.[201]

[200] Darley & Batson 1973, results on p. 104-5. Subjects were also interviewed on the nature of their religious beliefs and on their motives for becoming a minister. None of the differences between subjects in those regards mattered.

[201] Latané & Rodin 1969. Also see Latané & Darley 1970.

4 Implications of the psychological data

After administering 150 volts, the point at which the learner first states that he wishes the experiment to end, the subject turns to the experimenter and tells him he will not go any further. The experimenter retorts that the learner's protests are to be disregarded, and the following dialogue ensues:

EXPERIMENTER:	It's absolutely essential to the experiment that we continue.
SUBJECT:	I understand that statement, but I don't understand why the experiment is placed above this person's life.
EXPERIMENTER:	There is no permanent tissue damage.
SUBJECT:	Well, that's your opinion. If he doesn't want to continue, I'm taking orders from him.
EXPERIMENTER:	You have no other choice, sir, you must go on.
SUBJECT:	If this were Russia maybe, but not in America.[202]

This is the kind of reaction we would expect from any decent person in light of what the participants were asked to do. In reality, though, this subject was almost unique. The studies we have reviewed all seem to involve, at least, and to varying degrees, virtues such as compassion, benevolence and fortitude. We would surely predict a compassionate person of some fortitude not to obey the instructions to punish a perfectly innocent individual who desperately pleads to be released. Likewise, a benevolent person would offer help in the other three studies, particularly, I think, in the lady in distress study where excuses such as risks or time constraints do not apply. However, what is most striking about the results is not that that acts of compassion, benevolence or fortitude are rare but that they do not seem to have much to do with what *kind of person* is studied, but very much to do with the *particulars of the situation* any given person is acting in. If the character variable did most of the work there would seem to be no difference in the degree of obedience depending on the proximity of the victim, for instance, since a compassionate person would terminate the experiment at the point where the subject wishes it to be terminated; and the presence of a passive, non-helpful person would not have much of an influence on a consistently helpful person's decision to offer help. Let us

[202] Milgram 1974, p. 48-9.

now state the situationist challenge as a *modus tollens* argument along the following lines:

i. If human behavior were largely organized by robust traits, rigorous observation of human behavior would find general behavioral consistency.
ii. Rigorous observation has not found general behavioral consistency.
iii. Human behavior is *not* largely organized by robust traits.

5 Virtue ethical responses

There are some ways of defending virtue ethics in a tweaked form, a form not in collision with the situationist challenge. For instance, instead of speaking about virtuous *people*, one may speak about virtuous *acts*. Judith Thomson and Thomas Hurka have developed such accounts.[203] Or one may settle for the stable but very narrow traits accepted even by Doris, such as *helpful to fellow church-goers* or *courageous in the face of weather-based threats*. My focus here will be on attempts to defend virtue ethics in a more headstrong and ambitious way. The old-fashioned way if you will.

5.1 Virtues do not allow for that kind of testing

Julia Annas and Rosalind Hursthouse are probably the two most prominent virtue ethical philosophers alive. They have both been unimpressed with the situationist challenge, which they take to miss the notion of virtue relevant for virtue ethics.

As you remember, Doris characterizes character traits as robust dispositions to act in certain ways given certain conditions (with above chance probability p). Annas points out that this notion only sees character traits "from the point of view of a scientific observer".[204] But, she continues, "a virtue, unlike a mere habit, is a disposition to act which is exercised in and through the agent's practical reasoning".[205] This aspect of virtue allows for, indeed, demands, a responsiveness to the particular situation at hand:

> Practical expertise, including the understanding of the virtuous person, is highly situation-sensitive. […] A virtue is not an entity in me

[203] Thomson 1997 and Hurka 2006.
[204] Annas 2003, p. 22.
[205] Annas 2003, p. 24.

> determining my behavior; it is the way I am, my disposition to decide. And a virtue is a disposition to respond to situations in an intelligent and flexible way, not a stubborn habit that is indifferent to circumstances.[206]

For instance, being virtuous involves the competence to recognize if, in the given situation, being honest is more appropriate than being kind, as well as the inner emotional and deliberative processes involved in coming to a decision about how to act. Because of such nuances, Annas believes, the experiments and observations done in social psychology are too blunt to justify skepticism about virtues.

While it is valuable to point out that virtue involves aspects which are not easily attainable by scientific methods, the situationist can still contend that the present challenge concerns a necessary but not sufficient part of being virtuous: having a set of behavioral dispositions that remain stable over time and consistent across situations. The situationist critique does not assume that virtue is *solely* "an entity in me determining my behavior", but it does seem fair to maintain virtue is *partly* that, i.e. a disposition shaping my behavior. Even if, according to Annas, virtues are *intelligent* dispositions, not some mechanical regularity or disposition, this does not really help. For, what in the experiments cited is supposed to reasonably influence the intelligent situation-sensitive virtuous agent? The upshot of the situationist challenge is that tweaks of the situation make us act differently, *for no reason whatsoever.* Anyone can accept that the particular circumstances may warrant different behavioral responses. But the situationist challenge takes it as given that some variations are morally irrelevant. So Annas probably dismisses the challenge too quickly by assuming too simplistic a notion of virtue is being investigated.

Annas' points about virtue not being mere habit but critical reflection does not rise to a defense of virtue ethics against situationism if she does not also provide the particular situation-sensitive "excuses" in the cases under discussion. Alternatively, she may conceive of virtue primarily as an inner activity. But that is an unattractive route for other reasons. What good is the kind of sensitivity and practical reasoning she emphasizes, after all, if they do not help us do the right thing (or, if you will, that which a virtuous agent would actually *do*)? In cases where, because of the complexity of the situation, it is genuinely uncertain what the virtuous person would do, we would not be warranted in drawing conclusions about an agent's virtuousness

[206] Ibid., p. 27.

based on her behavior in those cases. But Annas has not argued that the Milgram cases are like that, and especially not that the different variations of them give rise to morally significant changes.

Rosalind Hursthouse too is quite dismissive of the psychological research being brought to bear on virtue ethics. According to her, because virtues are not just habits of behavior, such as telling the truth or helping, but "is concerned with many other actions as well, with emotions and emotional reactions, choices, values, desires, perceptions, attitudes, interests, expectations and sensibilities",[207] she thinks "the social psychologists' studies are irrelevant to the multi-track disposition ... that a virtue is supposed to be".[208] But this dismissal is also unsatisfactory. Is she saying that it is (always) impossible to find out whether or not a particular action was virtuous or not, or whether a particular person possesses a particular virtue or not? Situationist psychology denies that we have broad-based character traits that consistently generate trait-relevant behavior across different situations. Of course, Hursthouse denies situationism. On what grounds? Presumably because she thinks people do have, and differ in, character traits that are either virtues or vices, and that these traits play an important part in bringing about behavior. But if it is an empirical possibility to find out, as Hursthouse thinks she has, that people have character traits, it seems perfectly natural to suppose it equally much of an empirical possibility to find out they do not have.

If, on the other hand, virtues as multi-track dispositions are elusive to any test, asserting them is no less epistemically risky than being skeptical of their role in helping us understand human behavior. Quite the contrary. Hursthouse offers no criticism of the specific design and interpretation of individual studies invoked by situationists to undermine character attributions. Perhaps they do fail to track the multi-tracked virtues, and one wants to know in what ways. One way to do this would be to question the assumption that it is clear in the respective experiments what the virtuous thing to do is.

The type of critique Doris and Harman direct against globalism, and indirectly against virtue ethics, does not presuppose that virtues are *simple* character traits, or that they reflect unfailing, rigid habits. What it does presuppose, though, is that there is such a thing as a trait-relevant behavior given certain trait-relevant eliciting conditions. This comes down to the claim that there are descriptions of situations such

[207] Hursthouse 2013.

[208] Ibid.

that we can say what, e.g., a brave or a generous person would do in that situation. Perhaps such descriptions would have to be either very exhaustive or simplistically typified for them to make us feel confident in making a judgment as to what behavior in terms of a certain virtue is called for. To take an example, as familiar as mother's milk to any philosophy undergraduate, if I pass a shallow pond on my way to lecture and notice a child about to drown, it would be monstrously ignoble of me to keep on walking on the sole ground that I do not want to spoil my trousers.[209] *Beneficence requires that I save the child*, since the cost to me is trivial. We make this judgment instantly because, given the description, it is clear that the child's drowning and my being able to save it at only a trivial cost is a *beneficence-eliciting condition*. If, on the other hand, the pond is not shallow but deep and I do not know how to swim, or if my saving this child puts my own child's life in peril, it is no longer evident that this is a situation where beneficence is required, or what the beneficent thing to do would be.

For at least some particular occasions, virtue ethics must be able to inform us what the virtuous thing to do is. It will then be possible to examine empirically how common it is for people to act virtuously at those occasions. This in itself should not be controversial. Is the situation where, in Milgram's studies, the absent experimenter instructs the subject over the phone to administer 400 volts to a person sitting in an adjacent room such an occasion, i.e. one where virtue ethics gives us an answer as regards the virtuous thing to do? I can think of no possible reason why a virtue ethicist would doubt the least what the subject ought to do. If you think of virtues such as compassion or beneficence, this is a trait-eliciting condition, i.e. one where compassion or beneficence is appropriate or called for. Consider next the situation where the same instruction is given by the experimenter now present in the same room. Are the conditions rendering these same virtues appropriate or called for still present to the same degree or not? Surely they are, but people do not respond to them. The trait-relevant eliciting conditions are the same, but the trait-relevant behavior differs. That is why we may come to suspect people lack character and that what determines behavior is manipulation of the situational variables. It may be possible that the two conditions above really differ in morally relevant ways. But what, on virtue ethics' own account, could those differences be? Only if Hursthouse got into such details could she substantiate the

[209] Singer 1972.

claim that social psychology is irrelevant in studying "the multi-track disposition ... that a virtue is supposed to be".[210]

5.2 No surprise here, virtues are rare

Since virtues, on Annas' view, are like complex skills they take time to perfect and we cannot expect the perfection of these skills to be widespread. She did not conclude that virtue must be rare after reading Milgram and other social psychologists. Rather, virtue must be rare because it is a difficult skill. If virtue is very rare, finding it seldom is to be expected.

I think this is a pretty good response – virtue *is* rare. And this is not some *ad hoc* maneuver; no, Aristotle too held the same view. Mark Alfano, however, thinks this view, "though plausible for Plato, Aristotle, and Nietzsche, rubs our democratic ethos the wrong way" and that it goes against an egalitarian notion "that almost anyone can be brought reliably to do what the virtuous person would do".[211] But these are not incompatible beliefs. One may hold both that virtue is in fact very rare and that more or less anyone could be brought to act in *accordance* with virtue (n.b., not the same thing as acting virtuously). The rarity thesis is not threatened by the relatively weak egalitarian condition Alfano holds to be a core tenet of virtue ethics.

If virtues are rare, then what is the fuss? The more radical suggestion here must be that, given the situationist findings, not only is virtue rare but human behavior in general is not caused by underlying psychological character traits (of which virtues are an instance) at all. How may that be? Because having a character trait in the form of a virtue is having some specific set of values, norms, beliefs, desires, emotions etcetera. These things are intercorrelated in the person such that what a generous person desires is related to how she feels and what norms she are inclined to accept and how she is inclined to act. There is a causal unity which determines how each one of us reasons, feels, acts etcetera. The situationist challenge says there is no such unity; you can

[210] Kamtekar 2004 p. 460 similarly argues that the situationist challenge is based on an impoverished notion of character, the unrobustness of which poses no threat to virtue ethics: "Traditional virtue ethics offers a conception of character far superior to the one under attack by situationism... Briefly, the so-called character traits that the situationist experiments test for are independently functioning dispositions to behave in stereotypical ways, dispositions that are isolated from how people reason."

[211] Alfano 2013, p. 32-3.

change the mood and nothing else need to follow. Edouard Machery formulates this stance:

> Remember that the notions of character and of kind of person are meant to explain why behaviors differ (because characters differ) and how to change people's behavior (change their character). But now suppose that the mental states and dispositions that constitute our character and the kind of person we are are not unified. Then, one would not explain why behaviors differ by referring to people's character; rather, one would refer to their emotions, or to their values, or to their moods—viz. to specific psychological causes. Similarly, one would not propose to change people's behavior by changing their character; rather, one would propose to intervene on their moods, emotions, values, second-order desires, and so on. It would then seem that people have no character. […] Much of the recent research in psychology suggests that behavior is the product of numerous causes that are not correlated with one another.[212]

This very deep-going fragmentation of human agency sounds troublesome for virtue ethics. It seems to imply not only that virtue is rare but that it is more or less unattainable. Annas, of course, does not accept this view of human psychology in general, but even if she granted the lion's part of it she could say that in the virtuous person there is such a unity, and it is upheld even when disrupted by situational cues.

Another way of answering this lack of character interpretation is to dispute that it is the only, or even the best, account of the available evidence from studies such as the four recapitulated above. Situationist philosophers take the evidence to show that, because behavior is so sensitive to situational manipulation, there are no character traits. But another interpretation is that there are many character traits, and that situational manipulation may bring them into conflict. Being loyal and being compassionate, respecting personal integrity and being helpful etcetera. Additionally, many of the studies may not show lack of character (and hence lack of virtue) but rather the presence of some vices, such as vanity, egoism, deference to authority, greed, cowardice, laziness etcetera.[213] This view aligns with the commonsensical observation that most of us are, as Christian Miller likes to say, a mixed bag.[214]

[212] Machery 2010, p. 226-6.

[213] The idea that proponents of the situationist challenge have given insufficient attention to so-called non-malicious vices as cowardice and laziness is developed in Bates & Kleingeld 2018.

[214] Miller 2017.

6 *The science of individual differences*

I turn now to present and partially resuscitate the other side of the person-situation debate, namely the person. Our everyday, folk-psychological, view tells us people have and differ in important character traits, some of which we think of as virtues and vices. But the situationist challenge insists this is something of a chimera. Our intuitions in this area, Doris and Harman tell us, are like pre-scientific intuitions on many other issues – misguided. Both Harman and Doris emphasize the need to replace the appealing and seemingly self-evident views on character, and psychology more generally, with the best possible scientifically supported view, be it appealing or not. Social psychology has certainly demonstrated the power of the situation, and that there is such a thing as the fundamental attribution error, i.e. a systematic failure to account for the role of situational pressures in explaining human behavior.

But we might not have to resign ourselves to folk-psychology in order to back up the existence of substantial non-situationist influence on character and behavior. In the following, I want to sketch some of the considerations that justify us in not abandoning the belief that individuals differ systematically, and that individuals exert an influence over both what situations they end up in and how they act once in them. Replacing folk-psychology with our best-supported scientific ideas about the mechanics of human behavior will not necessarily mean we become character-denying situationists.; quite the contrary.

From the perspective of evolutionary psychology, the situationist stance is *prima facie* implausible. If different individuals behaved similarly when in similar situations, evolution could not occur based on differences in how successful their behavior was. But evolution has occurred precisely because of differential reproductive success. Accordingly, we can expect there to be individual differences in how different individuals will behave when faced with the same situation. This simple observation shows just how extreme and *a priori* unlikely a strong form of situationism is, and we should therefore demand extraordinary evidence before accepting it. Humans have successfully bred dogs based on *their* character traits, giving rise to breeds that vary in assertiveness, energy, aggression, protectiveness, playfulness, kindness etcetera. Other non-human animals too, including our closest relative, the Chimpanzees, vary in personality traits, and these differences affect how they behave in similar situations, allowing observers

to make reliable predictions of their future behavior.[215] Birds, too, have different personalities, with some being more agreeable, aggressive, curious or cooperative than others.[216] It would be a mystery if a feature necessary for evolution and present throughout the whole of the animal kingdom, including our closest relatives – relatively stable individual differences in how an organism responds to the same stimuli – were absent in our particular species. Let us look at some of what we know of these individual differences.

6.1 Vicious biology

For any psychological trait, people *vary*: some having more of it and others less. This is true of traits like intelligence, aggression, empathy, impulsivity, and many more. There is overwhelming evidence that individual differences – whether it would be the ones just mentioned or political orientation, depressive episodes, phobias or emotional stability – to a large part are due to differences in genes or other biological factors such as pre-natal conditions.[217] These biologically based traits, in turn, themselves are, or partly constitute, traits of character we would call virtues or vices. Here are a couple of examples.

For a variety of reasons, anti-social behavior is more well-studied than pro-social behavior. The concordance rate for juvenile delinquency for identical twins is about twice that of fraternal twins, suggesting a strong genetic component in whatever traits (e.g. aggression, impulse control, and intelligence) account for the difference in behavior.[218] Looking at the influence of biological factors on personality as inferred from studies of adoption paints a similar picture. Criminal convictions on the part of one or both of the biological parents is a strong predictor of criminal convictions of adoptees raised in a family not involved in criminal behavior. One study of adoptees suffering from an aggression disorder found that 30 percent of their biological parents had an antisocial personality disorder (but none of the adopted parents, who, naturally had been screened for precisely such disorders). If you think adoption and foster parenting may themselves be

[215] See Pederson, King, & Landau 2005, and Weiss et al. 2017. A situationist chimp psychologist would likely also remind us that "chimpanzees immigrating to a new group abandon their superior nut-cracking technology in favor of the inferior local one, just in order to blend in" (Rolf Degen's phrase); Luncz et al. 2018.

[216] Naguib & van Oers 2013. Also, Schuett, Dall, and Royle 2011.

[217] Bouchard 2004.

[218] Christiansen 1970; Raine 2013.

the root of the problem, consider the following study of adoptees whose biological mothers were felons and adoptees whose biological mothers were not. Members of the former cohort were more likely to have been arrested (15 percent versus 2 percent), convicted (13 percent versus 1 percent) and incarcerated (10 percent versus none), as well as to have been diagnosed with antisocial personality disorder. Other evidence of the biological influence on individual differences and differences in behavior includes studies showing that low concentrations of the neurotransmitter serotonin accompanies violent and other forms of anti-social behavior in humans. In one study ranging over a period of two years, aggressive children were followed. Those with the lowest levels of serotonin at the beginning of the study were most likely to end up in serious trouble by the end of the study. In another study, newly released convicts of manslaughter or arson were followed over a period of three years. Individuals with the lowest levels of serotonin upon release were most likely to commit another violent crime during the period of the study.[219] The level of testosterone is known to influence behavior, mood and character in many ways. Testosterone is involved in the development of the male sexual organs, the deepening of the voice and the occurrence of facial and pubic hair at puberty; it helps determine muscle size and strength, bone growth and sex drive and the production of sperm. Men produce around 25 times more testosterone than women do and it is often referred to as the major male sex hormone. Testosterone may be a key factor in ambition and positive forms of aggression, but is also implicated in violent behavior. Manslaughter, murder and assault are predominately committed by male perpetrators, and most commonly so by men in the ages 17 to 24, the age span at which testosterone levels peak, whereupon the curves for violent behavior and that for testosterone levels parallel one another in decline.[220] Among female convicts, those charged with offensive violence on average have higher levels of testosterone than those charged with defensive violence.[221]

[219] See Anderson 2006; also Barnes et al. 2014.

[220] Male killers outnumber females by almost 10 to 1 (the proportion is the same among our closest relatives the Chimpanzees). 79 percent of homicide victims are male (again, roughly the same proportion among Chimps). When a woman does kill, her victim is often an abusive male partner. See The United Nations Office on Drugs and Crime's *Global Study on Homicide: Trends, Contexts, Data*, and, for the Chimps, Wilson et al. 2014.

[221] Ferguson 2010; Raine 2013.

6.2 The Five Factor Model

The individual differences that exists in propensity to anti-social behavior are an instance of the more general finding that people have personality traits that exert an influence over behavior. These traits, of course, cannot be thought of as the sole determinants of behavior – situationist psychology tells us as much – but as habitual patterns of emotion, thought, and behavior. The traits may have low predictive value for how a given individual will behave in a single given situation, but they do help explain behavior over time, and not simply because individuals face the same type of situation over and over again.[222]

Which are these traits? Different accounts have been developed over the years, and there is no uncontroversial and undisputed single view or theory on this, but the now most widely accepted alternative is called the Five factor model. According to this, the characteristics that together make up an individual's personality can be mapped on a continuum of five basic personality dimensions:

a. Openness to experience
b. Conscientiousness
c. Extraversion
d. Agreeableness
e. Neuroticism

Or, for short, *OCEAN*. These terms are only partly self-explanatory, so it is useful to dwell on them a little bit. Openness has to do with an individual's preference for novelty versus routine, both concerning ideas and habits. Conscientiousness describes how dutiful and organized a person is. It also captures a somewhat different quality, viz. industriousness. Extraversion is about the individual's need for social stimuli, talkativeness, whether the company of others is a boost or drain on energy. Agreeableness measures the individual's tendency to be friendly, compassionate, and conflict-avoidant versus more adver-

[222] Also, what situations an individual seeks out or is prone to find themselves in is itself not independent of the individual's personality traits. Paragliding accidents and gun wounds are not distributed randomly across the population.

sarial, suspicious, competitive. Neuroticism, finally, measures emotional stability and the presence of negative emotions such as anxiety and nervousness versus stability and calm. [223]

I already mentioned that a trait such as conscientiousness captures what are really quite distinct sub-traits like orderliness and industriousness. Each of the five factors constitute a cluster of traits forming a temperamental family together constituting the personality considered as a whole. This also means that there is no special magic to the number five, though for now most researchers think that set of personality dimensions captures the various sub-traits we find. (Interestingly, from the point of view of moral psychology, a sixth personality dimension is sometimes invoked, referred to as "honesty/humility", which is meant to capture aspects such as sincerity, modesty, fairness, and greed-avoidance.[224])

I mentioned the effects of testosterone on human behavior (including its role in violent offending and as part of important sex differences) as well as personality differences in dogs, apes, and other animals. All this is strong evidence that there is an important biological component to personality traits. And this has been borne out in studies taking into account the relative causal contributions of genes and environment by studying siblings reared together or apart, identical twins reared together or reared apart, fraternal twins reared together or reared apart, non-related adopted siblings reared together etcetera. The result of this body of research, where both the environmental and genetic contributions are controlled for, is rather striking. Adopted individuals growing up together as siblings in the same family, receiving the same upbringing and belonging to the same socio-economic strata, are no more similar in their personality traits than people selected randomly on the street. Fraternal twins are less similar than monozygotic twins, who in turn are concordant for most traits whether reared apart or together. These findings is strong evidence genes account for a good deal of a person's personality. But more surprisingly, "the environment" as usually understood seems to have very little lasting effects on an individual's personality type. Non-shared environmental factors, that is, factors which are unique to every individual within a household, such as their fetal development, having an infection, a concussion, being exposed to some chemical or having a certain group of

[223] For a comprehensive presentation and discussion of personality psychology, including the genetics of personality, see Larsen & Buss 2017.

[224] Ashton & Lee 2005.

friends, however, do have an effects. So identical twins reared together and identical twins growing up separated in different families are neither more nor less similar. The correlations you see between parenting and a given outcome are almost entirely explained by shared genes. That is, the propensity of a parent to exhibit a certain behavior will be passed on by their genes, not their behavior, to the child as well.[225] So next time you see headlines like "Helicopter-parenting creates anxious adolescents" or "Reading to your children increases their intelligence" ask if the studies invoked may be genetically confounded.

7 Between a rock and a hard place

Although the biological and trait psychological considerations I have mentioned do not refute situationism (which does not speak of life outcomes and the like) they do bring some relative weight back to the person and his or her character as a decisive factor in how a life is conducted viewed as a whole, or at least in larger chunks than situations. Is this good enough for virtue ethics? From one perspective, neither story gives a very hopeful picture of what is left to work with. According to situationist psychology, we are fragmented individuals, whose behavior is overwhelmingly shaped by the situational whims of the surroundings. The trait psychological story moderates this conclusion, but the complementing picture it offers seems threatening in another way: our personality traits are challenging to influence, they do not predict individual behavior very well, and they do not seem to line up in any obvious way with what the virtuous character traits are.

If you have ever listened to a motivational speaker, chances are you have come across the notion that the Chinese word for "crisis" also means "opportunity". Or, as Homer Simpson economically dubbed it, *crisitunity*.[226] Let us see if we can adopt this perspective for a while, and explore our mangled virtue situation as a crisitunity. In the following, I will offer some reasons why optimism might be warranted, but also suggestions for reformation and revision for those interested in developing a virtue ethics for our time.

[225] Plomin & Daniels 2011. For two recent overviews of the nature and nurture of personality, see Mitchell 2018 and Plomin 2018. For an accessible book critical of the biological perspective, see Fine 2017.

[226] *The Simpsons*, 1994. The trope that the Chinese word for crisis could also mean opportunity has been around since the 1930s, but reached a wider audience when John F. Kennedy used it in some of his speeches. In accepting the Nobel peace prize, Al Gore also made use of it in his speech.

Stuck between the Scylla of situationism and the Charybdis of genetically and otherwise bestowed stable traits, what is the virtue ethicist to do? First, the approach should assimilate that situations exert a much more powerful influence on our behavior than has been granted by common sense. In addition to stress, conformity to the group and obedience to authority, changes in ambient smells and sounds influence us, as do seemingly trivial changes in mood. Becoming virtuous involves familiarizing oneself with these issues, and developing strategies to counteract them.[227] One way to counteract them may be, paradoxically, to give up. In part because it seems prudent not to trust one's situational self, but also because there are simply too many situations to handle. Being helpful all the time is not going to be very, well, helpful. In the modern world, there is no end to the number of occasions. Instead, being virtuous means organizing life in such a manner that one's helpfulness is funneled in ways that seem to make overall sense and be morally defensible. Virtue ethics, as well as any school of thinking about ethics in a modern context, has to take into account that our fragile hominid moral psychology has been outstripped by the ethical demands of a globalized world. Virtue ethics should engage both in *harm reduction* by incorporating how situationally fragile we are, and *prevention* by making our overall ethical behavior less dependent on how we handle sudden situations. The rationale for the first strategy, trying to make us see when situational factors that should not matter nevertheless influence us, is a lesson from psychology, but the rationale for making sure situations come to be less important is also based on the wider realization that much of ethical relevance is not about things we encounter or people we interact with.

Virtue ethics is an ancient account of the ethical life. Even though the ancient Greek civilization may have been as international as was imaginable at the time, human interaction then was much more limited in time and space. No one could imagine the moral issues evoked by faraway famines, mass migration, or climate change. Because the moral landscape is so different for us today, virtue ethics was due for a tune-up nonetheless. The call for such a self-assessment is not because virtue talk is obsolete, false or misguided, but because we need to rethink its application in a context of so many new, less concrete and more far-reaching ethical issues.

[227] Cf. Miller 2017, where he suggests several explorable routes to virtue.

The recommendation – or retreat, depending on how you think of it – that we allow ourselves to think of virtue not as situational instantiation but as aggregate outcomes or life-strategies, is in a way a narrowing of the ambitions of virtue ethics, but in another a widening of its application. Doris thinks it would be a considerable narrowing of our ambitions to settle for an aggregationist approach. First of all, he says, we really are interested in predicting and understanding behavior on particular instances. Questions such as "how could he do such a thing?"[228] presuppose the meaningfulness of explaining individual instances of behaviors, not just general trends. More importantly, perhaps, Doris says, virtue ethicist themselves typically do not think of virtues as general trends or aggregative outcomes, but as robust traits: "Describe a situation, even one where the situational pressures toward moral failure are high, and one can confidently predict what the virtuous person will not do".[229]

But because of the confounding effects of situational variables – group effects, mood effects, and the like – aggregative outcomes might be the best thing virtue ethics could hope for. Consider Josephine, who decides to auto-debit a percentage of her income every month to the Against Malaria Fund. Presumably her generosity or helping behavior towards strangers is just as subjected to subtle situational manipulations as anyone else's. Still, it *is* generous and compassionate of her to give away money in order to alleviate the harsh plight of others, even if her generosity varies according to situational variables in many other instances. That our moral lives in this way become less myopic also makes situations less important. The most salient moral choices may not be daily interactions, but choices about what we consume, what we do for a living. These are things we can exercise control over. So virtue ethics may have to settle for a type of life-plan perspective and give up some confidence in the act-particular perspective. Since, as I have said, there are good grounds for focusing less on situations, even absent social psychology, this recommendation is partly making a virtue of necessity and partly a positive suggestion in itself.

A possible worry here is that what I am suggesting is that virtue ethics is euthanized and resurrected as consequentialism. But the worry is unwarranted, for two reasons. First, I am not suggesting that any moral theory is correct but only that whoever finds virtue ethics attractive but also vulnerable to some psychological research may

[228] Doris 2002, p. 73.
[229] Ibid., p. 74.

want to reformulate the theory while preserving what in it was attractive in the first place. Secondly, my suggestion leaves intact the idea that aretaic considerations are more basic than considerations of valuable outcomes or obligatory action-types rather than the other way around.[230] So what is aggregated is not utility but virtuous actions, actions made from a motive of concern, generosity, courage, helpfulness, honesty and so on.

Consider Bruno Batta and Fred Prozi, whose participation in the obedience studies is described in the appendix. From the miserable point of view of the learner their behavior is identical. But whom would you trust (if forced to choose between no others) to care for your elderly mother? Most would settle for Prozi, I would guess. Why? Because he is the *more compassionate*. This assumption is further strengthened by the fact that Batta was in the "touch-proximity" variant and Prozi heard the pre-recorded vocal feedbacks from the learner, a setting less conducive to virtuous behavior. He was the more sensitive, it even showed in his behavior, but the situational factors made it too difficult for him to go all the way towards disobeying the instruction to inflict meaningless and undeserved pain on another. The tale of their respective behavior, though identical in perhaps the most important regard, nevertheless reveals a difference in their states of character. Their behaviors are not indicative of lack of character, but of flaws in character as well as of differences in character. We need to take these nuances into account.

Prozi had a greater sensitivity for what the situation required – he clearly wanted to stop the experiment and let the poor subject go – but he lacked the courage to do so. Now conjoining this description with the biological bases for individual differences that I have only briefly sketched in section 6, and the aggregative effects of these differences, we get what I think virtue ethics could most ambitiously hope for given situationist psychology: a higher probability for an overall virtuous life – not relentless reliability in doing the right thing.

"'His ethical perceptions were unfailingly admirable, although he behaved only averagely'", Doris writes in a mock epitaph meant to discredit a notion of virtue which emphasizes sensitivity rather than actual behavior.[231] Separating the two altogether would indeed perverse virtue ethics, but there remains a good possibility that there is a

[230] For more on the relationship between virtue ethics and consequentialist considerations, see Driver 2001 and Hartman & Bronson, forthcoming.

[231] Doris 2002, p. 17.

connection between Prozi-type sensitivity and actual behavior, *viewed aggregatively*. Granted, just about anyone could falter in the swamping noise of the present situation, but some character traits will nevertheless result in overall more virtuous behavior in a life. So the sensitivity, you can call it intellectual or not, which virtue ethicists speak of – *is* a disposition, but its effect is noticeable only over a series of acts or perhaps viewed over a life-time.

So much for coping with Scylla. What about the Charybdis of stable personality traits? From the point of view of virtue ethics, there are two problems with assimilating this body of research, which I will call normativity and alignment respectively. Virtue ethics, typically, is committed to the idea that one's character traits are malleable, possible to shape and perfect. But according to personality research, our traits are very stable and not easily changed. The other problem is that there is no clear understanding of the relationship between these five basic personality traits and the vocabulary of virtues.

These two problems, I suggest, might dissolve one another. The normativity charge starts with the commonplace that virtues are the sort of thing where moralization is appropriate, i.e. we blame and praise people based on their display of virtuous or vicious behavior, we admire and believe we ought to mimic virtuous people. But if we have no freedom over what traits we have, and little or no freedom in changing or developing them once acquired, how can there be room for normativity? The alignment problem is about our understanding of the virtues as behavioral dispositions, given a toolbox of just five basic personality traits where there seems to be little apparent overlap. Whence the virtues?

But given the fact that our personality is not easily changed, it is a good thing, from the point of view of virtue ethics, that these traits are not identical to a complete list of the virtues. Instead – and here is the crisitunity again – the fact that virtue and personality come apart opens up the possibility that a person may develop and perfect her virtues without going through the miracle of changing her personality. And this possibility, assuming it is one, is also the solution to the normativity charge. We may think of our five basic traits as the ingredients each one of us has to work with. There is a degree of freedom to recruit from your dealt hand of traits and behave generously, courageously, compassionately, and make these behaviors habitual. Scoring high on conscientiousness is equally likely to lead you to be like Eichman as Mandela. Expressed another way, looking at a personality test

result for a given individual will not allow you conclude with any certainty that they are just, brave, modest etcetera. Some virtues may be easier to compose given certain traits than others, but the mixture of traits underdetermines the mixture of virtues. The basic building blocks of our character are the result of a natural lottery, something Aristotle knew, but there is still room for growth and development given these conditions. Of course, given your hand, some virtues might be more easily attained than others, and the reverse will be true for others. We must probably also accept that through no fault of their own, some people will have a harder time attaining virtue than others. This element of luck and natural virtue and vice is hardly welcome, but we need not think it nullifies the approach.[232]

One observation I make having surveyed quite a bit of the social psychological literature on person and situation, and importing to this debate some of the thinking and experimentation on moral judgment, is that we can understand the situationist findings as analogous to the use of heuristics concerning judgment. Just like our judgments are guided by quickly available but only vaguely conscious heuristics such as "don't kill" "don't be violent" etcetera, our behavior is guided by similar heuristics, such as "do as you've agreed to", "don't make a scene", "stay out of trouble" and so forth. But just as the existence of heuristics shaping moral judgment, while a potent distorting factor, does not rule out a capacity to critically reflect on moral theories and these heuristics themselves, the existence of situationally triggered behavioral heuristics, while potent, does not rule out our having personality traits that are critical in shaping behavior. Sometimes these traits get swamped by the power of the situation, but sometimes they shine through. Over the course of many situations, indeed a lifetime, individual differences in traits will show. According to the situationist critique, the correlation between an individual's being helpful in situation A and situation B is simply too low to warrant any ascriptions of helpfulness to her. But a clever and illuminating piece of reasoning by psychologist Robert Abelson might provide some grounds for not dismissing low correlations as inconsequential. Abelson asks us to consider baseball player Ted Williams, who ended his career with a record .344 batting average. As his nickname "The Greatest Hitter Who Ever Lived" suggests, he is considered one of the best players of the game. Abelson compared this top average to one of the league's lowest, that

[232] See discussions in Athanassoulis 2005, and Hartman 2017. Also, Church & Hartman (eds.) 2019.

of Bob Uecker (.200). The difference in skill between Williams and Uecker explains less than 1 percent of the variance for a given single hit being successful or unsuccessful.[233] Still, no one would propose a team would be better off sending out Uecker rather than Williams to make that hit. Comparing character traits to baseball skills might not spell the vindication of virtue ethics, but the comparison does show, I think, as Sabini and Silver put it, that "believing in globalism is not entirely a matter of succumbing to an illusion".[234] Some hit better than others. More generally, the upshot is that "Personality psychologists have lost hope of predicting 'all of the people all of the time' and focus on predicting 'some of the people some of the time'."[235]

8 Situationism – in psychology, and in philosophy

For many, the appeal of virtue ethics is its distinctive way of formulating the questions of ethics: How am I to live? What kind of person should I be? The other ethical theories are usually taken to offer answers to the question What acts are right, and what makes them so? For me, this difference makes virtue ethics less, not more, interesting. For it is somewhat elusive in what sense virtue ethics is then a competitor to, indeed incompatible with, say, utilitarianism if its basic tenets are answers to a different set of questions than the rest of the bunch.[236] A utilitarian or Kantian, too, may be interested in what kind of person to be, and it is not clear answers at that level would be incompatible. For virtue ethics to be a theoretically independent alternative to consequentialism and other ethical theories it needs to be *incompatible* with them. It must be more than just psychological advice.

Here are some of the separate claims hovering around the designation 'virtue ethics':

a) Human behavior is aptly described as flowing from character traits of the individual.
b) The idea that leading a worthwhile life is inseparable from leading a moral life; you cannot be a flourishing psychopath.

[233] Abelson 1985.

[234] Sabini & Silver, 2005, p. 541n.

[235] Larsen & Buss 2017, p. 95.

[236] For developments, see Svensson 2010; Svensson 2011, and Svensson & Johansson 2018. Also, Hursthouse 1999.

c) The moral worth of human action is not exhausted by the consequences of overt behavior, nor to be assessed as tokens of action-types, but rather by reference to the motives, emotions and reasons of the agent; an action is right if, and only if, it would be performed by a fully virtuous agent acting in character.

The debate on virtue ethics and situationism has focused mainly on a) and to what extent it is true or false and to what extent its truth or falsity affects the plausibility of virtue ethics. But if b) were true, situationism may be a threat not only to virtue *ethics* but to prudential value. If we have no characters we cannot, on this view, have fully flourishing lives; without moral excellence, the quality of our lives are diminished.[237] On hedonist or preferentialist accounts, lack of virtue is unrelated to wellbeing, but for the development of some sort of perfectionist or objective list view, there is a relation such that leading *a good life* has to do both with prudence and morality.

I have tended to believe c) is pretty insulated from experimental results. How can a criterion of rightness be affected by empirical data? The criterion is perfectly intelligible of course. We can imagine such a thing as a fully virtuous agent and can understand the idea that her motives, emotions, and reasons are what ground our moral obligations. I do not find the criterion particularly attractive, partly because it seems incomplete in an important sense.[238] But this is an old discussion in ethical theory and there is nothing about possible moral implications of psychological research which would settle *that* debate.

8.1 The ongoing reappraisal in social psychology

True, there is a tension in the field of psychology between those who emphasize the power of the situation and those who emphasize the stability of traits, but it is also important to note that this quarrel is about where within the middle ground to settle. Both Doris and Harman are motivated by a desire to confront moral philosophy with the

[237] Appiah 2008 brings up this implication.

[238] If the virtuous is explanatorily prior to the good and the right, then, as Julia Driver said, "it seems natural to ask the further question, 'Why would the virtuous agent advise me to do A?' If the answer is simply that what the virtuous agent advises determines right action, independent of any other reasons or considerations, then the account seems quite capricious; if, on the other hand, there are independent reasons, then aren't those the right-making features – and then isn't what the virtuous agent advises superfluous?" (Driver 2006, p. 118.). As you can see, there is an analogue here to the Euthyphro challenge to Divine command theory.

most up to date and accurate psychological science. They view our notions of personality as a folk-psychological hodgepodge, and they take social psychology experimentation to provide a sobering scientific corrective.[239] Ironically, once philosophers started to notice the situationist trend in psychology, psychologists themselves had already started to moderate the more extreme conclusions, and the pendulum had swung back towards the person part of the person-situation research.

A more general observation is that it has become apparent that much of psychological research fails to replicate, i.e. when a different research group attempts to follow the same protocol, they fail to demonstrate the previously published result.[240] So, when you read of an effect or proposed phenomenon of our psychology, chances are that the claim is false or unsubstantiated. In fact, of all the subdisciplines of psychology, *social* psychology is particularly frail; studies in this field successfully replicated in just 25 per cent of cases. That is about half the success rate compared to cognitive psychology.[241] Of the studies cited by situationist critics of virtue ethics, Milgram's is the most tightly controlled, and yet it did not sample a randomized selection of the population but people who volunteered to help the experimenter in a psychological study. After listening to the audiotapes still kept at Yale University, and after interviewing participants and their friends, families and relatives, psychologist Gina Perry became convinced that participants, as in the Stanford Prison Experiment, realized what was expected of them. She was also able to establish that the experimenter's cues were not employed in a consistent manner.[242] Both the notoriety of the basic study design and the stricter ethical restrictions on what unwitting participants may be exposed to have made exact replications difficult, but what has been done seems to confirm the initial basic findings, though there is ongoing discussion on just how to interpret the results. Milgram took himself to study *obedience*, but only one of the four prods where actually an order, namely the fourth, "You have no other choice, you *must* go on". And, in attempts to replicate

[239] "Personality psychology studies the ways ordinary people think about personality and character traits, which is to be distinguished from studying the truth about personality and character traits." Harman 2009, p. 236. Deleted from the published version but still available online is the assessment, totally erroneous in my view, that "personality psychology is in pretty bad institutional shape as a scientific discipline".

[240] Nosek et. al. 2015.

[241] Ibid., p. 5.

[242] Perry 2012. Also see Griggs 2017.

the study, *no one* continued upon hearing *it*. As mentioned, there was a bit of improvisation going on around the putatively fixed prods, but it appears the more explicitly order-like the prod was, the less likely participants were to keep administering shocks. In other words, the orders of an authority figure led to disobedience, not obedience. Prods emphasizing that the experiment was harmless or that the shocks were required for scientific purposes, however, where less likely to trigger disobedience. Instead of obedience, we may think of the subjects' behavior as identification with the scientific goals of the study.[243] Milgram was much more rigorous than Zimbardo; still, looking at all of the ancillary evidence, as well as his own accounts (see the this chapter's appendix, pp. 124-7), one becomes troubled by the plain observation that he seems so eager to *tell a story*.

Many of the other studies invoked by situationists are problematically underpowered. The fact that there are many of them may suggest an underlying trend nonetheless, but the 25 per cent replication rate remains our best sober assessment of the strength of research in this field. Social psychology research of the coins-in-a-phone-booth type sometimes look like middle school social science projects compared to the emerging science of behavioral genetics, which employs rigorous quantitative methodology, carefully controlling for possible confounding factors, and utilizes data from vast numbers of individuals, often whole populations.[244]

What accounts for the poor replicability rate in social psychology? For a long time, there has been strong incentives to build one's career by coming up with new theories and findings, rather than accumulating findings into large bodies of knowledge interconnected by a plausible theoretical framework. Psychology therefore runneth over with theories and effects, the one more silly and specific than the other. We have the "Google effect" (the tendency to forget things that are easy to find with an internet search) or the "Rashomon effect" (the tendency that people will describe one and the same series of events differently based on their previous experience and biases). These effects are obviously specific (and well-known) instances of broader psychological mechanisms, but you cannot publish a paper claiming just that. The field set itself up for this backlash by perpetuating a norm that

[243] See Griggs for more discussion and references. Also see Hollander & Turowetz 2017.

[244] For more on behavior genetics and how it relates to psychological research, see Barbaro & Penke 2020.

gifted scholars have their own theories and effects and only the bores would care about effect sizes and reproducibility. This phenomenon, too, of course, has a name: the toothbrush effect. What is true of toothbrushes, the saying goes, is true of theories: no self-respecting psychologists would use anyone else's.[245] Another explanation for the low replicability is the bias in favor of publishing *positive* results. We will never know how many times a team of social psychologist baked lovely croissants and did *not* get people to help strangers more than without the pleasant scent. For this reason, there is now a movement which aims to convince scholars to preregister interesting research questions, and then commit to publishing the result whatever it is. Sounds like a good idea.[246]

Another factor is a form of anti-biological bias. In a recent survey, the attitudes and beliefs of leading social psychologists on the relevance of evolutionary thinking for psychological research was studied. While almost every prominent researcher in social psychology affirmed that all life on Earth, including humans, are the product of evolution, only half assented to the proposition that evolution has had an effect on our *minds* and influences social attitudes and preferences.[247] In fact, we know that *all* psychological traits have a genetic component. We also know that no psychological trait is a hundred percent determined by genetics.[248] Studying humans without taking onboard these basic facts is futile.

9 Conclusion and moving forward

As is often the case when two people quarrel, there is actually more agreement than meets the eye. Annas criticizes Doris for thinking of virtue as "uncritical and rigid habit" (while simultaneously something "developed independently of activity") which is "radically unintellectual" and boils down to "cloddish habit-following".[249] She emphasizes that virtue instead is a "disposition to *act on reasons*" and that "the

[245] The term, if not the observation, was coined by Watkins 1984. Later discussed in Mischel 2009.

[246] For more on the turn to preregistered studies, see Nosek et al. 2018. Ritchie 2020 offers a good discussion, not just relevant for psychology.

[247] Buss and von Hippel 2018, p. 6.

[248] Plomin et al. 2016.

[249] Annas 2005, p. 637, 639.

more virtuous you are, the more complex and dynamic your character".[250] For these reasons, Annas believes, virtue ethics remains unharmed by Doris's clumsy cross-examinations.

If Annas found too little talk of dynamic adaptation and intelligent reasoning in Doris' interrogation of virtue ethics, it is because his whole project is premised on the idea that we look at differences in behavior in response to situations between which there *are no morally significant differences*, i.e. comparisons where there simply are no reasons to respond differently. Intelligence, nuance, and dynamic reasoning are all fine – in the end, Annas must agree with Doris these morally important qualities would not lead its possessor to conclude that the scent of freshly baked croissants spells an increased obligation to help bypassers. But her charge is in any event unfair since Doris frequently emphasized that being virtuous is not mechanically acting "the same" but is about deliberation, emotions, sensitivity to changes in the situation etcetera. Early on in the book, he writes, in what could equally well be from one of Annas' works on the subject, that "virtues are not *mere* dispositions but *intelligent* dispositions, characterized by distinctive patterns of emotional response, deliberation, and decision as well as by more overt behavior."[251] As for the "rigid" part, Doris states that "to attribute a virtue is not to say that a person can be counted on to reliably do the same thing but to say that they can be counted on to reliably do whatever is appropriate to that virtue."[252] Finally, before people brought up social psychology experiments as ways of criticizing virtue ethics, Annas expressed herself in ways that are similar to what she now blasts: virtue implies, she wrote, "a firm tendency to act and decide in one way rather than the other."[253]

In addition to creating disagreement where none exists, it seems as if Annas makes virtue ethics a moving target so as to prevent any criticism from sticking. This is a general bug of the debate: the difficulty of fixating what it is virtue ethics needs to be true of human psychology. Critics of virtue ethics, such as Doris, have been very clear on what they take to be the empirical presuppositions of virtue ethics, and how they think these presuppositions square with available evidence. It has been less easy to pinpoint what defenders of virtue ethics believe that their theory needs to be true of human psychology.

[250] Ibid. p. 637

[251] Doris 2002, p 17.

[252] Ibid., p. 176. Also see Doris' response to Annas in the *PPR* exchange.

[253] Annas 1993, p. 51. Cited in Doris 2005.

It is time to sum up what lessons can be learnt from exposing virtue ethics to the situationist critique. First, there is such a thing as attribution error, that is we tend to overemphasize the importance of the individual's trait and discount the significance of the situation. Folk psychology is on the side of personality psychology, and it is a good thing situationists have offered a sobering corrective. Aiming for virtue means coming to understand the mechanics of behavior as a person-situation interaction, and develop cognitive and emotional tools to recognize and offset those kinds of situational cues that are likely to make us act in ways we would not endorse on reflection.

Even if the situationist critique took virtue ethics by surprise, leading to a somewhat dazed, defensive response, the theory, as far as I can tell, has not been dealt a lethal blow. There is still hope for virtue ethics, and improved versions will certainly be developed. We should all admit virtuous development must include knowledge of, and strategies to overcome, situational variables. Also, while not becoming too elitist, virtue ethicists should not be ashamed to tell the world that developing into a virtuous person takes time and skill and is not easy, and hence we cannot expect it to be common. More controversially, the view will have to live with a degree of moral luck, in the sense that some of the ingredients of virtue are personality traits that the individual has only little control over.

How disappointing situationist conclusions from social psychology are for virtue ethics depends on what your expectations and ambitions were to start with. The relative advantage virtue ethics long has been thought to possess compared to, say, utilitarianism or Kantianism, has been its linkage to an appealing moral psychology, one that inspires us to moral development and enlightens moral education. Letting go of much of the psychological baggage, what remains is no longer a "philosophy of life", but simply one ethical theory among many others. That may not be too bad, either. As I stated above, I believe for independent reasons that a criterion of rightness along the lines, "An action is right if and only if it is what a virtuous agent would characteristically do in the circumstances" is unsatisfactory, but that is an argument we may not be able ground in experimental work.

Another lesson, perhaps not as salient for the ancients, is that moral concerns today are so much more far-reaching, both in time, space and in the number of individuals affected. There is no end to the number of occasions I can be helpful, and we therefore need to design our lives in ways that make us, on the whole, helpful. That may not be identical

to offering help to every one asking us for it in our daily lives. Even if I described this as a conversion from a "classic" view to an aggregationist view, the change should not be overstated. When we think of moral exemplars we think of people who committed their lives to fighting injustice, helping others, standing up to illegitimate power etcetera. They had a life-plan approach to being a good person, and that is still possible.

Appendix: Same but different

Reports from the Milgram experiments

Example 1

Bruno Batta is a thirty-seven-year-old welder who took part in the variation requiring the use of force. He was born in New Haven, his parents in Italy. He has a rough-hewn face that conveys a conspicuous lack of alertness. He has some difficulty in mastering the experimental procedure and needs to be corrected by the experimenter several times. He shows appreciation for the help and willingness to do what is required. After the 150 volt level, Batta has to force the learner's hand down on the shock plate, since the learner himself refuses to touch it.

When the learner first complains, Mr. Batta pays no attention to him. His face remains impassive, as if to dissociate himself from the learner's disruptive behavior. When the experimenter instructs him to force the learner's hand down, he adopts a rigid, mechanical procedure. He tests the generator switch. When it fails to function, he immediately forces the learner's hand onto the shock plate. All the while he maintains the same rigid mask. The learner, seated alongside him, begs him to stop, but with robotic impassivity he continues the procedure.

What is extraordinary is his apparent total indifference to the learner; he hardly takes cognizance of him as a human being. Meanwhile, he relates to the experimenter in a submissive and courteous fashion.

At the 330 volt level, the learner refuses not only to touch the shock plate but also to provide any answers. Annoyed, Batta turns to him, and chastises him: "You better answer and get it over with. We can't stay here all night." These are the only words he directs to the learner in the course of an hour. Never again *does he* speak to him. The scene is brutal and depressing, his hard, impassive face showing total indifference as he subdues the screaming learner and gives him shocks. He seems to derive no pleasure from the act itself, only quiet satisfaction at doing his job properly.

When he administers 450 volts, he turns to the experimenter and asks, "Where do we go from here, Professor?" His tone is deferential and expresses his willingness to be a cooperative subject, in contrast to the learner's obstinacy.

Example 2

Fred Prozi's reactions, if more dramatic than most, illuminate the conflicts experienced by others in less visible form. About fifty years old and unemployed at the time of the experiment, he has a good-natured. if slightly dissolute, appearance, and he strikes people as a rather ordinary fellow. He begins the session calmly but becomes tense as it proceeds. After delivering the 180-volt shock, he pivots around in the chair and, shaking his head, addresses the experimenter in agitated tones:

Prozi: I can't stand it. I'm not going to kill that man in there. You hear him hollering?

Experimenter: As I told you before, the shocks may be painful, but . . .

Prozi: But he's hollering. He can't stand it. What's going to happen to him?

Experimenter (his voice is patient, matter-of-fact): The experiment requires that you continue, Teacher.

Prozi: Aah, but, unh, I'm not going to get that man sick in there -- know what I mean?

Experimenter: Whether the learner likes it or not, we must go on, through all the word pairs.

Prozi: I refuse to take the responsibility. He's in there hollering

Experimenter: It's absolutely essential that you continue, Prozi.

Prozi (indicating the unused questions): There's too many left here, I mean, Jeez, if he gets them wrong, there's too many of them left. I mean, who's going to take the responsibility if anything happens to that gentleman?

Experimenter: I'm responsible for anything that happens to him. Continue, please.

Prozi: All right. (Consults list of words.) The next one's "Slow -- walk, truck, dance, music" Answer please. (A buzzing sound indicates the learner has signaled his answer.) Wrong. A hundred and ninety-five volts. "Dance." (Zzumph!)

Learner (yelling): Let me out of here! My heart's bothering me (Teacher looks at experimenter.)

Experimenter: Continue, please.

Learner (screaming): Let me out of here! You have no right to keep me here! Let me out of here, my hearts bothering me, let me out!

Prozi (shakes head, pats the table nervously): You see he's hollering. Hear that? Gee, I don't know.

Experimenter: The experiment requires . . .

Prozi (interrupting): I know it does sir, but I mean -- hunh! He don't know what he's getting in for. He's up to 195 volts! (Experiment continues, through 210 volts, 225 volts, 240 volts, 255 volts, 270 volts, at which point Prozi, with evident relief, runs out of word-pair questions.)

Experimenter: You'll have to go back to the beginning of that page and go through them again until he's learned them all correctly.

Prozi: Aw, no. I'm not going to kill that man. You mean I've got to keep going up with the scale? No sir. He's hollering in there. I'm not going to give him 450 volts.

Experimenter: The experiment requires that you go on.

Learner: Ohhh. I absolutely refuse to answer anymore. (Shouting urgently now.) Let me out of here! You can't hold me here Get me out. Get-me-out-of-here!

Experimenter: Continue. The next word is "green," please.

Prozi: "Greengrass, hat, ink, apple." (Nothing happens. No answering buzz. Just gloomy silence.)

Prozi: I don't think he is going to answer.

Experimenter: If the learner doesn't answer in a reasonable time, about four or five seconds, consider the answer wrong. And follow the same procedures you have been doing for wrong answers. Say, "Wrong," tell him the number of volts, give him the punishment, read him the correct answer. Continue, please.

Prozi: "Green -- grass, hat, ink, apple." Answer, please. (More silence. Subject just sits there.)

Experimenter: Go on with the procedure, please.

Prozi: Three hundred and fifteen volts The answer is "ink." (Zzumph!)

Learner: Ohh. Ooohh

Prozi: (relieved to hear response). Next one . . .

Learner: I absolutely refuse to answer.

Prozi: You got to. You get a shock if you don't answer. (Experiment continues with learner refusing to answer, and, finally failing even to grunt or groan. At 375 volts, the subject stands up.)

Prozi: I think something's happened to that fellow in there. I don't get no answer. He was hollering at less voltage. Can't you check in and see if he's all right, please?

Experimenter (with same detached calm): Not once we've started. Please continue, Teacher.

Prozi (sits down, sighs deeply): "Cool day, shade, water, paint." Answer please. Are you all right in there? Are you all right?

Experimenter: Please continue, Teacher. Continue, please. (Subject pushes lever. Zzumph!)

Prozi (swiveling around in his chair): Something's happened to that man in there. (Swiveling back.) Next one. "Low -- dollar, necklace, moon, paint." (Turning around again.) Something's happened to that man in there. You better check in on him, sir. He won't answer or nothing.

Experimenter: Continue. Go on, please.

Prozi: You accept all responsibility?

Experimenter: The responsibility is mine. Correct. Please go on. (Subject returns to his list, starts running through words as rapidly as he can read them, works through to 450 volts.)

Prozi: That's that.

Experimenter: Continue using the 450 switch for each wrong answer. Continue, please.

Prozi: But I don't get anything!

Experimenter: Please continue. The next word is "white."

Prozi: Don't you think you should look in on him, please?

Experimenter: Not once we've started the experiment.

Prozi: What if he's dead in there? (Gestures toward the room with the electric chair.) l mean, he told me he can't stand the shock, sir. I don't mean to be rude, but I think you should look In on him. All you have to do is look in on him. All you have to do is look in the door. I don't get no answer, no noise. Something might have happened to the gentleman in there, sir.

Experimenter: We must continue. Go on, please.

Prozi: You mean keep giving that? Four-hundred-fifty volts, what he's got now?

Experimenter: That's correct. Continue. The next word is "white."

Prozi (now at a furious pace): "White -- cloud, horse, rock, house." Answer, please. The answer is "horse." Four hundred and fifty volts. (Zzumph!) Next words, "Bag -- paint, music. clown, girl." The next answer is "paint." Four hundred and fifty volts. (Zzumph!) Next word is "Short -- sentence, movie . . ."

Experimenter: Excuse me, Teacher. We'll have to discontinue the experiment.

5 Concluding Remarks

Of all the beautiful and memorable quotes photographer Steve Pyke asked his portrayed philosophers to provide, I always liked Kit Fine's the most:

> Philosophy is the strangest of subjects: it aims at rigor and yet is unable to establish any results; it attempts to deal with the most profound questions and yet constantly finds itself preoccupied with the trivialities of language; and it claims to be of great relevance to rational enquiry and the conduct of our life and yet is almost completely ignored. But perhaps what is strangest of all is the passion and intensity with which it is pursued by those who have fallen in its grip.

This thesis project was spurred by what one may call a naturalist urge, i.e. a will to have science and philosophy interact and hopefully see those stubborn yet elusive philosophical problems either solved or dissolved. A lesson from my attempts to confront moral philosophy with science is an appreciation of just how resilient the field is. And not just philosophy as such, but individual philosophical theories and positions. Originally I had a hope that moral philosophy's confrontation with psychology would lead to an undermining of deontological and virtue ethical views, leaving consequentialist ones intact and relatively strengthened. And to an extent this hope or prediction materialized. But I also concede that all of the major options in ethical theory are still alive and kicking. As Folke Tersman put it, "so many moves are open to a clever philosopher, that people will soon figure out ways to accommodate the new data within just about any philosophical theory."[254]

And it is not just about being clever. Positions are actually revised and refined in light of these kinds of confrontations. There is an enormous discussion in philosophy about the evidentiary role of intuitions, inspired by work in psychology, neuroscience, cross-cultural anthropology and evolutionary theory. Psychologists and physical anthropologists are formulating grand views on the "nature of morality" and how these foundational models in some sense are innate or not. The competence of moral philosophers is badly needed in such endeavors, but it is equally true that philosophers should get up to speed on the developments in these neighboring fields.

[254] Tersman 2008, p 390.

The renewed interest in the role of intuitions has not just led to a more mature sense of self awareness in the field of philosophical methodology and to some great works being written, but also meant the beginning of a discussion on how the possibility that intuitions vary by sex, class, and ethnicity has an impact on who feels at home in the world of academic philosophy and who feels like the odd one out.[255]

Since its first inception there has been a more or less continuous interest from philosophers in evolutionary theory. This view of the world has made a mark on epistemology, metaphysics, philosophy of mind, and, of course, moral philosophy. The relationship between moral philosophy and evolutionary thinking is especially interesting, with responses ranging from dismissal to "the time has come for ethics to be removed temporarily from the hands of the philosophers and biologicized".[256] There was a wave of writings on evolution and ethics in the aftermath of Edward O. Wilson's controversial 1975 book *Sociobiology* (from which the quote above is taken), with philosophers like Michael Ruse arguing for skeptical and or anti-realist conclusions, and some, like Peter Singer and James Rachels employing evolutionary thinking to undermine appeals to partiality, thus having it make a positive contribution to normative ethics rather than a skeptical one.[257]

A second wave of interest in the implications for moral philosophy of evolutionary theory was incited by Sharon Street's and Richard Joyce's skeptical work. Theirs is a much more sophisticated and well-worked out challenge to moral realism than the one prompted by the earlier ones formulated by Ruse and Wilson. In response, moral realists have refined their view and offered ingenious replies and developments. [258] The relevance of evolution is also a cornerstone in metaethical work defending anti-realist or so-called quasi realist expressivist views on ethics by thinkers like Alan Gibbard, Simon Blackburn, and Mark Schroeder.[259]

What are the next steps in the field of empirically informed ethics? It is difficult to make predictions, especially about the future. On the one hand we may see a degree of saturation stemming from the reali-

[255] Fricker 2007; "Buckwalter & Stich 2014; Drożdżowicz 2018, Machery, et al. 2017. Demarest et al 2016; Figdor & Drabek 2016.

[256] Wilson 1975, p. 562.

[257] Ruse 1985; Singer 1981; Rachels 1990.

[258] Enoch 2013. Justin Clarke-Doane 2016.

[259] Gibbard 1990, Blackburn 1998, Schroeder 2010.

zation that empirical findings rarely amount to a knock-down refutation of any philosophical view of some complexity. On the other, we are bound to encounter ever more refined empirical measures to study moral judgments at the neural level, giving philosophers new stuff to think about. And the more philosophers themselves participate in designing such studies, the more likely it is that they will be of high quality and philosophical relevance.

The study of the purported innateness of morality involves a broad spectrum of research disciplines and is continuing to evolve. Researchers from comparative and biological anthropology will systematize the uniformity and variation of human morality, and will merge that kind of research with *fMRI* data on moral thinking at the neural level, providing us with both a wider and a more fine-grained analysis. My impression is that the researchers involved in these kinds of explorations often think of them as vindicating certain normative outlooks, while the perspective of philosophers is often to think of the findings as something potentially serving as biases in our moral thinking. The earlier in the process philosophers are involved, the less likely that the work produced contains philosophical mistakes, and the more likely, too, that the best available psychological, biological, and anthropological data trickle down to philosophical discussions.

What has been the relevance of the data and theories surveyed in this thesis has varied with the issue at hand. In the case of virtue ethics and situationist psychology, the way forward has mostly been in terms of setting the scientific record straight, realizing how much of social psychology has failed to replicate and is more and more complemented with (if not replaced by) findings anchored in behavioral genetics and personality psychology. As for the dual process theory in general, and *fMRI* studies of moral judgment in particular, here I have mostly taken the available data as given. But we should not trust the empirical parts here to be settled, and the picture may change with new developments and techniques. The finding that an important subset of deontological moral judgment responds to morally irrelevant factors is troubling for the view and does have an undermining effect on their normative force. But as the end of the chapter showed, just what deontologists should say about various cases remains unsettled, and so there is obvious room for development, including ones which bring onboard these findings to make sure the ensuing theory is not a rationalization of responses to morally irrelevant factors.

So moral philosophy has proved a very adaptive sounding board to advancements in evolutionary thinking, and far from being dissolved it has flourished. That the discussion on the innateness of morality is premised on an evolutionary understanding of human psychology is so obvious it hardly needs to be mentioned. The debate over the neuroscience and psychology of moral intuitions is also solidly anchored in evolutionary considerations. Even my rejoinder to the situationist critique of virtue ethics is premised on the impact of evolutionary considerations on individual differences in personality traits, and though not all virtue ethicists will welcome help of that sort it is their best bet. Daniel Dennett famously said that Darwin's idea of evolution by natural selection was "the single best idea anyone has ever had".[260] It is evident to see how its sway permeates everything human.

Many years ago, in an intro psychology class, I read about how pilots need to learn about and take into account several biologically grounded visual cues that in some cases would lead them to miscalculate the aircraft's altitude. Coming in to a landing field that starts after an ascending slope will give the pilot the impression that the angle is too steep when in fact it is good. Changing the plane's angle based on the pilot's visual impressions may cause the plane to crash. The landscapes surrounding the runway, and how lights from nearby cities or constructions present themselves, will affect what the human mind will take to be reliable information on the trajectory of the plane. Sometimes these input are accurate, sometimes not. A skilled pilot will have to learn to sometimes recruit our spontaneous visual cues, but to sometimes disregard them and instead trust what seems a counterintuitive but test-proven corrective. I like to think of moral philosophy as these pilots navigating hazardous landscapes. By learning more about ourselves, philosophers too may come to be able to every now and then successfully land a plane.

[260] Dennett 1995, p. 21.

6 Svensk sammanfattning

Avhandlingen undersöker olika sätt på vilka forskningsfynd i psykologi kan tänkas vara av betydelse för vårt ställningstagande i filosofiska debatter kring etisk teori. ”Psykologi” förstås här i vid mening och inkluderar bland annat socialpsykologi, kognitiv psykologi, utvecklingspsykologi och delar av hjärnforskning liksom evolutionsteoretiska resonemang och landvinningar relevanta för mänskligt beteende. Etiska teorier, å sin sida, är generella uppfattningar om vad det är som gör vissa handlingar och beslut moraliskt riktiga och andra oriktiga (jag använder här termerna ”etik” och ”moral” som synonymer). Det är i grova drag tre olika sådana etiska teorier som i avhandlingen på olika sätt konfronteras med psykologiska rön: utilitarism, deontologisk etik samt dygdetik.

Utilitarismen säger att en handling är moraliskt riktig om, och endast om, den leder till minst lika bra konsekvenser som varje alternativ handling som agenten kunde utföra i situationen med avseende på det totala välbefinnandet. Lite mer informellt säger utilitarismen att vi handlar moraliskt rätt när vi på ett opartiskt sätt i så stor utsträckning som det är möjligt handlar så att kännande varelsers välbefinnande främjas.

Vad kunde vara fel med det? Jo ett mindre tilltalande sätt att beskriva saken är att ändamålet enligt utilitarismen helgar medlen. Vilken *sorts* handling som maximerar välbefinnandet ges enligt utilitarismen ingen självständig vikt, vilket kan tyckas orimligt. En viktig etisk tradition – *deontologisk etik* – uppfattar istället riktighet som knuten till *handlingstyper*. Vilka handlingstyper som är i och för sig själva rätta eller felaktiga kan det finnas olika uppfattningar om, men några förslag kunde vara ”att avsiktligen döda en oskyldig människa”, ”att använda en individ enbart som ett medel för att gynna andra” eller ”att kränka någons grundläggande mänskliga rättigheter för att uppnå ett önskvärt utfall för flertalet”. Deontologisk etik förnekar också utilitarismens moraliska likställande av *handlingar* och *underlåtelser*. Vi har, säger deontologiska etiska teorier, ett strängare moraliskt ansvar för de handlingar som vi aktivt utför än för utfall som kan sägas vara resultatet av att vi har förblivit passiva. Enligt utilitarismen handlar vi ju fel närhelst vi misslyckas med att maximera välbefinnandet, även om vår ”handling” inte är annat än att sitta med armarna i kors.

En tredje etisk teoribildning, som faktiskt är den äldsta rent idéhistoriskt, fokuserar istället på våra *karaktärsdrag*, dvs individens stabila

emotionellt och kognitivt grundade beteendedispositioner. Vissa sådana karaktärsdrag är särskilt intressanta från moralisk synpunkt, till exempel ärlighet, hjälpsamhet, välvilja, generositet, mod och så vidare. *Dygder* som dessa – och motsvarande *laster* – är centrala för hur vi bedömer både oss själva och andra. Ett fokus på dygder i moralfilosofisk kontext kan anta olika former, men en central utgångspunkt som är särskilt relevant i just detta sammanhang är antagandet att människor *har* karaktärsdrag, att vi kan utveckla dygder, och att vårt handlande till stor del är förutsägbart och förklarbart med utgångspunkt i våra respektive individuella karaktärsdrag. *Han delade med sig av pajen eftersom han är* generös*; Hon sprang in i den brinnande byggnaden eftersom hon är* modig*; Eftersom han är en* ärlig *person kommer han att avvisa erbjudandet om att fuska på provet* och så vidare.

Vetenskap och normativ etik

Vetenskapen söker beskriva och förklara världen sådan den *är*; moral handlar om att ta ställning till vad som är *bra* och *dåligt* med världen, hur den *borde* förändras och så vidare. Det är tydligt att detta är två helt olika infallsvinklar på saker och ting – att studera dem och att bedöma dem. Efter den skotske 1700-talsfilosofen David Hume talar man om den så kallade *Humes lag* enligt vilken en värderande slutsats inte kan följa logiskt från en uppsättning premisser försåvitt inte någon av premisserna är en värdering. Denna insikt tycks innebära att inga fynd i psykologi eller andra empiriska vetenskaper kan ha några direkta implikationer för vad som är moraliskt rätt eller fel. I avhandlingen bestrider jag heller inte denna princip. Den relevans som psykologiska forskningsrön kan ha för moralfilosofiska frågor är därför mer *indirekt*.

Ett vanligt sätt att åberopa psykologisk forskning har att göra med *tillförlitligheten* i vår bedömning av olika etiska utsagor, vare sig dessa är generella teorier eller enskilda omdömen. Detta kan ses som en specialinstans av en mer allmän observation, nämligen den att kunskap om en trosföreställnings tillkomsthistoria kan komma att påverka hur vi ser på trosföreställningens grad av rättfärdigande. Betrakta följande två olika upprinnelser till en trosföreställning:

A: Ingrid kom fram till att det är 43 studenter närvarande genom att dra en lott från en urna med siffror från noll till 100.
B: Britt-Marie kom fram till att det är 24 studenter närvarande genom att noggrant räkna alla i seminarierummet.[261]

[261] Exemplet kommer från Sober 2018, sidan 211.

Det är tydligt att Ingrid kommit fram till sin övertygelse på ett sätt som är otillförlitligt. Det finns inga skäl att tro att sanningen om hur många som befinner sig i seminarierummet har någon relation till vilken siffra som dras från urnan. Det skulle kunna vara så att man råkar dra rätt siffra, men denna metod för att avgöra antalet närvarande leder inte till rättfärdigade övertygelser, inte ens när de råkar vara sanna.

Någonting liknande detta har kommit att riktas mot olika etiska övertygelser, med verktyg från psykologisk vetenskap. Till förslagen hör att vissa eller alla etiska övertygelser är otillförlitliga för att de härrör från regioner av hjärnan som producerar *emotionella responser*, eller för att det har haft ett *överlevnadsvärde* att vara benägen att hysa sådana etiska övertygelser, eller för att de liknar det slags *förenklade* och ibland felaktiga *tumregler* som vi omedvetet applicerar för att förstå oss på sannolikheter och rationalitet och så vidare. I den moralfilosofiska diskussionen har dessa så kallade *undermineringsstrategier* (debunking strategies) blivit ett populärt och hett omdiskuterat delområde. I avhandlingen undersöker jag flera sådana strategier relaterade till oenigheten mellan utilitarister och försvarare av deontologisk etik.

Den konfrontation med psykologisk forskning som dygdetiken utsatts för är delvis av annat slag. Utmaningen här är inte främst att stödet för dygdetik skulle härröra ur psykologiska processer som leder till otillförlitliga övertygelser, även om detta är en tillkommande aspekt hos debatten. Den centrala utmaningen mot dygdetiken är snarare att den bygger på vissa antaganden om mänskligt beteende som har kommit att ifrågasättas av socialpsykologisk forskning. Denna utmaning formuleras och diskuteras i avhandlingens kapitel 4.

Kapitel 1

Har människan en medfödd moral? Som alltid börjar svaret med konstaterandet ”det beror på vad man menar”. Ibland tolkas frågan som om vi av naturen är snälla och samarbetsinriktade snarare än våldsamma och egoistiska. Men i kapitlet argumenterar jag för att frågan bör preciseras till: är vår benägenhet att tänka i moraliska termer och fälla moraliska omdömen en evolutionärt skapad *anpassning* eller är det en sidoeffekt av att vi kan tänka, känna, tala osv? Det ligger i sakens natur att frågan inte kan ges ett definitivt svar, men jag argumenterar för att vår moraliserande tendens sannolikt är en specifik anpassning och inte bara vilken sidoeffekt som helst, likt sparkcykelåkning eller zappande framför teven. Att vi kan göra sådana saker beror förstås också på hur evolutionen har format oss, men den har inte format oss så *för att* det var gynnsamt att kunna göra dem.

Flera slags evidens kan åberopas för moraliskt tänkande som en evolutionär anpassning. Förmågan att tänka i moraliska termer är till exempel en viktig signal till andra att man är en pålitlig samarbetspartner. Moraliska övertygelser hjälper individen att skjuta upp belöningar eller avstå avhopp från samarbeten som ger utdelning på längre sikt genom att upplevas som en auktoritativ och extern källa till motivation.

Det är påfallande hur snabbt moraliskt tänkande går. Det är nästan som observationer: man tycker sig direkt se vad som är rätt eller fel. Samtidigt visar forskning att människor kan ha väldigt svårt att underbygga sina moraliska bedömningar med *skäl* – trots att de själva upplever sig göra bedömningen i kraft av vissa skäl. Detta liknar lite grann hur kompetenta talare av ett språk vet vad som är rätt form eller uttryck men inte kan formulera de grammatiska regler som förklarar varför. Denna analogi mellan språk och moral har visat sig vara en fruktbar ansats och är i debatten känd som ”universal moral grammar”. Tanken är att vårt moraliska tänkande följer en viss grundläggande struktur som är biologiskt grundad och därför gemensam för alla kulturer även om den kan fyllas på med vissa variationer beroende på de mer specifika omständigheterna. Men i denna ”grammatik” kommer vi överallt finna vissa regler kring och emotionella responser inför till exempel bruk av våld.

Tanken att moralen är medfödd i den mening som här preciseras får ses som ett pågående forskningsprogram, där filosofi, antropologi, psykologi, evolutionsforskning, arkeologi och så vidare ger viktiga bidrag till en ännu oavslutad diskussion. Även om mycket förblir osäkert tror jag vi med ganska stor säkerhet kan slå fast några saker som kommer visa sig av intresse i kommande kapitel:

1) vi har en uppsättning medfödda emotionella reaktioner som får oss att tycka att vissa moraliska omdömen är rimligare än andra, och
2) vi saknar ofta medveten tillgång till de principer och skäl som kan tänkas rättfärdiga dessa omdömen, vilken gör oss sårbara för självbedrägeri och rationaliseringar.

Kapitel 2

Här får vi börja med att förklara tre versioner av ett tankeexperiment som har kommit att inta en central plats i diskussionen. På engelska går de under namnet ”the trolley dilemmas”, så vi kan väl kalla dem för *spårvagnsdilemmana*. De tre versionerna kan formuleras så här:

Växeln
En förlupen spårvagn skenar fram i hög hastighet. Om den får fortsätta sin färd kommer den att köra över och döda fem människor som befinner sig längre fram på spåret. Det enda sättet att rädda dem är att växla om vagnen till ett annat spår. Längre fram på det spåret står dock en ensam människa som då dödas av vagnen. Är det moraliskt acceptabelt att växla om vagnen?

Knuffen
En förlupen spårvagn skenar fram i hög hastighet. Om den får fortsätta sin färd kommer den att köra över och döda fem människor som befinner sig längre fram på spåret. En betraktare på en gångbro mellan den framrusande vagnen och de fem människorna överväger att hoppa framför spårvagnen för att rädda de andra. Men han inser att han väger för lite, och att hans offer skulle vara meningslöst. Bredvid honom står däremot en mycket storväxt individ. Det enda sättet att rädda människorna på spåret är att knuffa ned den storväxta individen från bron och framför tåget. Den individen kommer då att dö, men hans kropp är tillräckligt tung för att tåget ska stoppas. Är det moraliskt acceptabelt att knuffa ned den storväxta främlingen framför tåget?

Öglespåret
En förlupen spårvagn skenar fram i hög hastighet. Om den får fortsätta sin färd kommer den att köra över och döda fem människor som befinner sig längre fram på spåret. Det enda sättet att rädda dem är att växla om vagnen till ett annat spår. Detta spår går i en cirkel och ansluter sedan återigen till huvudspåret strax framför gruppen av människor. På öglan står en storväxt främling som är tung nog att stoppa spårvagnen men inte tillräckligt tung för att själv överleva kollisionen. Är det moraliskt acceptabelt att växla om vagnen?

De flesta människor bedömer att det är moraliskt försvarligt att växla om spårvagnen i det första fallet, men att det däremot inte är moraliskt försvarligt att knuffa främlingen i det andra. Varför denna skillnad, fallen liknar ju varandra däri att man för att rädda fem behöver orsaka ens död? Ett tänkbart svar är att den enskildes död i det första fallet är en *oavsedd sidoeffekt* av att man räddar de fem från att krossas av spårvagnen. Naturligtvis skulle man helst vilja att stickspåret var tomt. Den som däremot knuffar någon framför en spårvagn för att rädda livet på andra använder denna människa som ett slags *instrument*, en tyngd, vars död gynnar andra människor. Även om utilitarismen naturligtvis säger att vi ska handla så att så många som möjligt överlever i fall som dessa, kan det tyckas att vi har en godtagbar skillnad mellan fallen som deontologen kan peka på: det är inte bara det totala utfallet som har moralisk betydelse, vilken *sorts* handling agenten utför har

också betydelse. Och det är, säger deontologen, en moraliskt relevant skillnad mellan att rädda fem även om man *förutser* att en dör som en olycklig bieffekt, å ena sidan, och att, å den andra, *avsiktligen* döda en person som ett *medel* för att rädda andra. Denna sorts konflikt mellan utilitarismens rekommendationer och deontologisk etik är typisk och välbekant.

Med introduktionen av öglespårsdilemmat blir läget lite mer komplicerat. En vanlig bedömning är att det är försvarligt att växla om spårvagnen till öglan där den storväxta främlingen står, åtminstone mer försvarligt än att knuffa personen från gångbron. Samtidigt verkar de principer som talade för att knuffen var oförsvarlig också tala för att det är fel att ingripa här. Främlingen på spåret dödas också avsiktligen, eftersom kollisionen med honom är en förutsättning för att rädda gruppen längre fram på spåret.

Kanske kan man försöka revidera försvaret av den moraliskt relevanta skillnaden mellan *Växeln* och *Knuffen* på ett sådant sätt att *Öglespåret* ges samma lösning som *Växeln*, trots att den har vissa grundläggande likheter med *Knuffen*. Men att det har visat sig så svårt att på ett redigt sätt ge ett moralfilosofiskt överblickbar berättigande av människors spontana bedömningar är också intressant från psykologisk synpunkt: hur lever människor med dessa spretiga bedömningar, och hur uppkommer de?

Denna fråga blev startskottet för en serie studier av de psykologiska processer som ligger till grund för människors moraliska omdömen. Filosofidoktoranden, sedermera psykologiprofessorn, Joshua Greene formulerade misstanken att orsaken till att de tre spårvagnsfallen bedöms olika inte har så mycket att göra med subtila moralfilosofiska distinktioner utan snarare förklaras av att de olika beskrivningarna ger upphov till starkare eller svagare *emotionella reaktioner* hos oss. Mer specifikt ville han skilja så kallade *personliga* från *opersonliga* dilemman åt. Opersonliga dilemman kan handla om att välja mellan olika människors död och överlevnad, men agenten behöver inte bruka våld för att åstadkomma något av utfallen. Personliga dilemman, å andra sidan, rymmer ett alternativ som i någon mening leder till ett bättre sluttillstånd (fler överlevande till exempel) men där agenten behöver bruka våld för att uppnå detta utfall.

Greene och hans forskargrupp använde en så kallad *fMRI*-kamera för att studera hjärnaktiviteten hon människor som tog ställning till olika slags beslut, inklusive moraliska dilemman av både opersonligt och personligt snitt. Fyra förutsägelser formulerades inför studien:

1) De områden i hjärnan som är dokumenterat förknippade med producerande av känslor kommer att vara mer aktiverade när vi konfronteras med ”personliga moraliska dilemman” och i mindre grad aktiverade när vi konfronteras med ”opersonliga moraliska dilemman”; 2) En majoritet kommer att bedöma att personliga överträdelser inte är acceptabla även om de skulle ge totalt sett bättre konsekvenser. 3) En majoritet kommer att bedöma att opersonliga överträdelser är acceptabla om de totalt sett ger bättre konsekvenser. 4) Den minoritet som anser att personliga överträdelser är moraliskt acceptabla om de är nödvändiga för att konsekvenserna totalt sett ska bli de bästa kommer att behöva längre tid på sig jämfört med majoriteten.

Förutsägelserna slog alla in. I kombination med många andra slags forskningsfynd har Greene argumenterat för att mänsklig psykologi i allmänhet, och moraliskt beslutsfattande i synnerhet, rymmer två slags mentala processer. En av dessa processer är explicit, reflekterande, språklig och någonting som sker medvetet. En annan del är mer automatisk, ofta omedveten, och inte sällan förankrad i emotionella processer och responser. Den förra processen är ofta tidskrävande men medger flexibilitet och nyansrikedom, medan den senare är snabb och tillfredsställande för de flesta kontexter.

Greene liknar denna tudelning med inställningsalternativen på en modern systemkamera, där användaren kan välja på manuellt läge (som kräver kompetens och tar tid men medger specifika lösningar för specifika ändamål) och ett automatiskt läge där användaren snabbt kan ta bra bilder som bygger på att kamerans programmering själv avgör vilka inställningar som passar bäst. Greenes position är att det manuella läget – ”förnuftets röst” – är benägen att rekommendera utilitaristiska lösningar på moraliska dilemman, så länge de är opersonliga. När dilemman i stället är personliga blir responsen mer emotionell, och vi har en stark känsla att det inte kan vara rätt att utföra en viss våldsam handling även om ändamålet är att rädda så många som möjligt. Detta trots att även opersonliga dilemman kan innebära att någon dör, eller mer allmänt att de totala utfallen är likvärdiga. Greene menar att denna uppkomsthistoria borde underminera vår tilltro till deontologiska omdömen, eller åtminstone att sådana omdömen inte med fog kan användas för att kritisera utilitarismen.

Det finns mycket att säga om *fMRI* och mer allmänt om denna dubbelprocessteori om det mänskliga psyket. För våra filosofiska syften kan man formulera kärnan i Greenes kritik av deontologisk etik enligt följande:

I. En viktig grupp deontologiska omdömen orsakas av emotionella processer som svarar på omständigheter som gör ett dilemma personligt snarare än opersonligt.
II. De omständigheter som gör ett dilemma personligt snarare än opersonligt är moraliskt irrelevanta.
III. En viktig grupp deontologiska omdömen orsakas av emotionella processer som svarar på omständigheter som är moraliskt irrelevanta.
IV. Deontologisk kritik av konsekventialistiska etiska teorier som åberopar omdömen som svarar på moraliskt irrelevanta omständigheter kan ges låg eller ingen vikt i diskussioner om utilitarismens rimlighet.

Tanken är alltså att utilitarismen får ett slags indirekt stöd av denna forskning, däri att en typ av vanliga argument mot utilitarismen inte bör ges den tyngd som de ibland ges. Jag försvarar i stora drag denna syn, men pekar också på sätt som deontologiska positioner kan försöka bemöta kritiken, bland annat genom att tydligare ange vilka omdömen som en deontologisk teori närmare bestämt kan tänkas ge inför fall som dessa.

Kapitel 3

En del av den utmaning som presenteras mot deontologisk etik i kapitel 3 vilar på antagandet att evolutionen har fått oss att reagera starkare, mer emotionellt, på vissa slags moraliska situationer, nämligen sådana som inbegriper personligt våld. Mer allmänt kan man kanske förvänta sig att evolutionen format vår moralpsykologi med fokus på ”den lilla världen”. Mellanmänskliga relationer, snarare än folkhälsa eller brottsprevention alltså. Denna insikt kan få oss att se mer kritiskt på vissa till synes självklara etiska uppfattningar: vi tror på dem därför att det har varit gynnsamt att vara så funtad, inte för att de är sanna. Här kan man tacksamt kontrastera med vår perceptionsapparat. Denna är ju inte perfekt, men överlag är det svårt att förstå hur till exempel syn och hörsel skulle ha varit evolutionärt gynnsamma förmågor om de inte hjälpte oss att upptäcka världen ungefär så som den är. När jag ser en tekopp framför mig är den bästa förklaringen till denna observation att det verkligen står en kopp där som orsakar dessa förnimmelser i mig. När det gäller moraliska övertygelser verkar vi inte på samma sätt behöva anta existensen av *moraliska fakta* som upprinnelse till övertygelserna.

Såna här tankegångar leder i skeptisk riktning, och filosofer som Sharon Street och Richard Joyce har utvecklat olika resonemang om hur evolutionspsykologiska rön underminerar anspråk på moraliska fakta eller andra antaganden förknippade med moralisk realism. Men kanske träffar denna evolutionära underminering vissa etiska uppfattningar mer än andra? I detta kapitel undersöker jag förslaget att utilitarismen är bättre rustad än andra etiska teorier att motstå försök till så kallad evolutionär underminering. Grundtanken kan sägas vara att eftersom utilitarismen bygger på en så krävande och oegennyttig utgångspunkt är det svårt att se hur evolutionen kunde ha gynnat att människor anammar den. Tvärtom borde det ha haft lågt överlevnadsvärde att gå omkring och tänka att främlingars väl och ve betyder lika mycket som mitt eget och bör maximeras opartiskt. Just för att det är en så altruistisk tanke finns det hopp om att vi accepterar den för att den verkligen är sann snarare än för att vi *gynnas av att tro* att den är sann.

Att en etisk teori inte kan undermineras evolutionärt är ju så klart inte ensamt ett tecken på att den är en rimlig eller sann teori. I kapitlet diskuteras förslaget från Katarzyna de Lazari-Radek och Peter Singer att utilitarismen bygger på antaganden som är *självevidenta*, vilket bland annat innebär att deras sanning är någonting som vi kan inse med förnuftet. Dessa axiomatiska utgångspunkter kommer från Henry Sidgwick och lyder:

U: En enskild individs goda är, från universums synpunkt, inte av större betydelse än någon annan individs goda.
R: Som rationell individ är jag ålagd att sikta mot det goda generellt, i den mån jag kan påverka saken, inte enbart mot en enskild del av det.

Tanken är alltså att från ett opartiskt perspektiv är vars och ens väl lika värdefullt, varken mer eller mindre. *För mig* kanske mitt väl är mer värdefullt, men så är det alltså inte ”från universums synpunkt”. Nästa steg är att rationaliteten fordrar att vi har det allmänna goda som ändamål, inte just vårt individuella goda. Att välja ut en del av det goda – mitt väl – är irrationellt på samma sätt som det är irrationellt att bry sig om vissa *ögonblick* mer än andra inom det egna livet. Att diskontera betydelsen av framtida lidande för att kunna njuta idag är ju själva sinnebilden för oförnuftigt handlande.

I kapitlet uttrycker jag skepsis mot flera delar av detta förslag. Dels kritiserar jag själva utgångspunkten att en etisk uppfattning kan vara självevident. Jag förespråkar en mer holistisk syn på rättfärdigande, enligt vilken ingen enskild uppfattning har status av axiom utan varje del i systemet motiveras med utgångspunkt i hur väl den passar in i och bidrar till helheten. Men även om vi för argumentationens skull accepterar denna utgångspunkt, kan man kritisera det specifika förslaget att Sidgwicks principer och den utilitarism som de motiverar motstår evolutionär underminering. Ett problem i dessa debatter är att man kan tänka sig en evolutionär tillkomsthistoria för i stort sett vilka dispositioner som helst. Även utilitaristiska.

Kapitel 4

Skulle du ha varit en av dem som lytt order om du levt i Nazityskland, eller skulle du kanske har gjort motstånd och tagit stora risker för att hjälpa andra? Varför agerar människor så olika? Är vissa mer empatiska och hjälpsamma och mer andra självupptagna och fega? Att det är på det viset förefaller självklart, men en betydelsefull forskningsinriktning inom socialpsykologin kan tyckas ge stöd för tanken att människors handlande inte alls förklaras av individuella karaktärsegenskaper som dem ovan utan långt mer av egenskaperna hos den *situation* som de agerar inom. Även om jag gärna vill tro om mig själv att jag är empatisk och modig så är nog sanningen den att jag hade varit en av dem som lydde order i Nazityskland om jag hade levt då. Det gjorde de flesta, och jag har ingen särskild grund för att tro att jag är annorlunda. Jag har haft tur som inte prövats så drastiskt.

Nazityskland var en extrem tid, men kanske kan denna blinda lydnad frambringas även i modern tid, i ett demokratiskt samhälle med individuell frihet att säga nej? Psykologen Stanley Milgram ville testa sin misstanke att de flesta av oss skulle lyda en auktoritetsfigur, trots att inget straff väntar för olydnad, att på given order bestraffa en annan människa, och i en serie sinnrika och omskakande experiment fick han vanliga, genomsnittliga människor att utdela vad de trodde var extremt smärtsamma eller rent av dödliga elektriska stötar till en försöksperson som misslyckades med att lära sig några glosor.

Milgrams och en hel rad andra studier gav stöd åt en *situationistisk* syn på mänskligt beteende. En sådan syn verkar vara ett hot mot etiska teorier som fokuserar på att utvärdera handlingar och människor i termer av hur *dygdiga* de är. Den situationistiska utmaningen mot dygdetik kunde formuleras så här:

1) Om mänskligt beteende till största delen förklaras av individers robusta personlighetsdrag så skulle noggranna studier finna stor samstämmighet i individens handlande över tid och mellan olika men relevant likartade situationer.
2) Noggranna studier finner ingen sådan samstämmighet.
3) Alltså: mänskligt beteende förklaras inte till största delen av individers robusta personlighetsdrag

Ett tag såg denna kritik av karaktärsdrag och dygder ut att utgöra ett allvarligt hot mot dygdetiken. Men i kapitlet försöker jag erbjuda flera olika svarsstrategier. En del av svaret handlar om att moderera ambitionsnivån, och en del är mer inomvetenskaplig och fokuserar på den stora mängd forskning som trots allt visar att det finns stora individuella skillnader mellan människor. Situationen är mindre avgörande än vad de allra mest hotande formuleringarna av situationism gjorde gällande. En del av den ”hjälp” jag erbjuder kanske ses av dygdetiker som björntjänster. Mitt försvar för biologiskt grundade skillnader i personlighetsegenskaper föder nämligen nya problem: hur kan man blir mer dygdig, och hur kan en människa vara ansvarig för sin karaktär om mycket av förutsättningarna är biologiskt givna? Jag erbjuder några svarsstrategier, bland annat den att dygder inte har någon direkt motsvarighet i modern personlighetsforskning utan måste ses som kompositer av flera olika personlighetsdrag, vilket kan öppna för möjligheten att dygder är förändringsbara och inom individens kontroll.

Avslutning

När psykologiska eller andra empiriska rön når filosofin finns ofta en förhoppning att vi nu äntligen kan bilägga dessa gamla debatter med hjälp av ”riktig vetenskap”. Men debattlandskapet framträder strax i ny skepnad, upplyst och förändrat men inte utplånat. Hur frustrerande det än är så får vi nog inse att det finns ett knippe filosofiska frågor som aldrig helt kan lösas av vetenskaperna. Likväl har de olika konfrontationer med tillgänglig empiri som jag har presenterat och utvärderat bidragit till att utveckla och bättre formulera, omvärdera och revidera filosofiska positioner. Min initiala målsättning med avhandlingen var att visa att utilitarismen skulle komma ut på andra sidan dessa konfrontationer mindre skadskjuten är konkurrenterna. Under arbetets gång har jag släppt den ambitionen något och intar istället en mer agnostisk hållning, både till utilitarismen och till styrkan i de olika undermineringsstrategier som formulerats mot deontologisk etik och dygdetik.

7 Bibliography

Abelson, Robert 1985: "A Variance Explanation Paradox: When a Little is a Lot", pp. 129-133 in *Psychological Bulletin*, vol. 97, no. 1.

Adleberg, Toni, Thompson, Morgan & Nahmias, Eddy 2015: "Do Men and Women Have Different Philosophical Intuitions? Further Data", pp. 615-41 in *Philosophical Psychology*, vol. 28, no. 5.

Ahlenius, Henrik & Tännsjö, Torbjörn 2012: "Chinese and Westerners Respond Differently to the Trolley Dilemmas", pp. 195-201 in *Journal of Cognition and Culture*, vol. 12, no. 3-4.

Al-Ubaydli, Omar, Jones, Garett, & Weel, Jaap 2013: "Patience, Cognitive Skill, and Coordination in the Repeated Stag Hunt", pp. 71-96 in *Journal of Neuroscience, Psychology, and Economics*, vol. 6, no. 2.

Alfano, Mark 2013: *Character as Moral Fiction*, Cambridge University Press.

Anderson, Gail 2006: *Biological Influences on Criminal Behavior*, Simon Fraser University Publications.

Annas, Julia 1993: *The Morality of Happiness*, Oxford University Press.

Annas, Julia 2003: "Virtue Ethics and Social Psychology", pp. 20-34 in *A Priori*, vol. 2.

Annas, Julia 2005: "Comments on John Doris's *Lack of Character*", pp. 636-42 in *Philosophy and Phenomenological Research*, vol. 71, no. 3.

Annas, Julia 2011: *Intelligent Virtue*, Oxford University Press.

Anscombe, Elizabeth 1958: "Modern Moral Philosophy", pp. 1-19 in *Philosophy*, vol. 33, no. 124.

Appiah, Kwame Anthony 2008: *Experiments in Ethics*, Harvard University Press.

Aristotle: *Nicomachean Ethics*, translated with introduction by David Ross, Oxford University Press 1992.

Ashton, Michael & Lee, Kibeom: "Honesty-Humility, the Big Five, and the Five-Factor Model", pp. 1321-53 in *Journal of Personality*, vol. 73, no. 5.

Athanassoulis, Nafsika 2005: *Morality, Moral Luck and Responsibility*, Palgrave Macmillan.

Audi, Robert 2008: "Intuition, Inference, and Rational Disagreement in Ethics", pp. 475–92 in *Ethical Theory and Moral Practice*, vol. 11, no. 5.

Awad, Edmond et al. *2018:* "The Moral Machine Experiment", pp. 59-64 in *Nature* 563.

Awad, Edmond et al 2020: "Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants", pp. 2332-7 in *Proceedings of the National Academy of Sciences*, Feb., vol. 117, no. 5.

Ayer, Alfred Julius 1936: *Language Truth, and Logic*, 2nd ed. 1946, Dover Books 2002.

Bacon, Francis 1598: "A Letter of Advice to the Earl of Essex", in *The Letters and Life of Francis Bacon,* Spedding, J. (ed.), vol. 2, 1862.

Baillargeon, Renée 1987: "Object Permanence in 3 ½ and 4 ½ Month Old Infants", pp. 655–64 in *Developmental Psychology* vol 23, no. 5.

Barbaro, Nicole & Penke, Lars 2020: "Behavior Genetics", pp. 336-54 in, Shackelford, Todd (ed.), *The Sage Handbook of Evolutionary Psychology*, Sage Publications.

Barnes, James Christopher et al. 2014: "Demonstrating the Validity of Twin Research in Criminology", pp. 588-626 in *Criminology*, vol. 52, no. 4.

Bartels, Daniel 2008: "Principled Moral Sentiment and the Flexibility of Moral Judgment and Decision Making", pp. 381-417 in *Cognition,* vol. 108, no. 2.

Bates, Tom & Kleingeld, Pauline 2018: "Virtue, Vice, and Situationism", ch. 27 in *The Oxford Handbook of Virtue*, ed. Snow, Nancy, Oxford University Press.

Berker, Selim 2009: "The Normative Insignificance of Neuroscience", pp. 293-329 in *Philosophy and Public Affairs*, vol. 37, no. 4.

Berwick, Robert, Chomsky, Noam et al. 2013: "Evolution, Brain, and the Nature of Language", pp. 89-98 in *Trends in Cognitive Sciences* vol. 17, no. 2.

Blackburn, Simon 1998: *Ruling Passions: A Theory of Practical Reasoning*, Oxford University Press.

Bloom, Paul 2013: *Just Babies: The Origins of Good and Evil*, Broadway Books.

Bonnefon, Jean-François, De Neys, Wim, and Trémolière, Bastien 2012: "Mortality Salience and Morality: Thinking about Death Makes People Less Utilitarian," pp. 379-84 in *Cognition* vol. 124, no. 3.

Bouchard Jr, Thomas 2004: "Genetic Influence on Human Psychological Traits: A Survey", pp. 148-151 in *Current Directions in Psychological Science,* vol. 13, no. 4.

Brink, David O. 1989: *Moral Realism and the Foundations of Ethics*, Cambridge University Press.

Broad, C. D. 1936: "Are There Synthetic A Priori Truths?", pp. 102-17 in *Proceedings of the Aristotelian Society,* Supplementary volume 15.

Brosnan, Sarah & de Waal, Frans 2003: "Monkeys Reject Unequal Pay", pp. 297-9 in *Nature*, vol. 425, 18 Sept.

Brosnan, Sarah & de Waal, Frans 2014: "Evolution of Responses to (un)fairness" in *Science*, vol. 346, no. 6207.

Buckwalter, Wesley & Stich, Stephen 2014: "Gender and Philosophical Intuition.", pp. 307–46 in Knobe & Nichols (eds.), *Experimental Philosophy, Vol. 2*. Oxford University Press.

Buss, David & von Hippel, William 2018: "Psychological Barriers to Evolutionary Psychology: Ideological Bias and Coalitional Adaptations", pp. 148-58. APA: *Archives of Scientific Psychology*, 6.

Carlsmith, Kevin, Darley, John, and Robinson, Paul 2002: "Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment", pp. 284-99 in *Journal of Personality and Social Psychology* vol. 83, no. 2.

Chappell, Timothy 2013: "Virtue Ethics in the Twentieth Century", pp. 149-71 in Russell, Daniel (ed.).

Chomsky, Noam 1965: *Aspects of the Theory of Syntax*, MIT Press.

Chomsky, Noam et al. 2002: "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?", pp. 1569-79 in *Science,* vol. 298, 22 November.

Christiansen, Karl 1970: "Crime in a Danish Twin Population", pp. 323-26 in *Acta Geneticae Medicae et Gemellologiae*, vol. 19, no. 1-2.

Church, Ian & Hartman, Robert (eds.) 2019: *The Routledge Handbook of the Philosophy and Psychology of Luck*, Routledge.

Churchland, Patricia 2011: *Braintrust: What Neuroscience Tells Us about Morality*, Princeton UP 2011.

Churchland, Patricia 2019: *Conscience: The Origins of Moral Intuition,* W.W. Norton & Company.

Clarke-Doane, Justin 2016: "Debunking and Dispensability", in Leibowitz, Uri & Sinclair, Neil (eds.) *Explanation in Ethics and Mathematics*, Oxford University Press.

Conway, Paul & Gawronski, Bertram 2013: "Deontological and Utilitarian Inclinations in Moral Decision Making: A Process Dissociation Approach, pp. 216-35 in *Journal of Personality and Social Psychology*, vol. 104 no. 2.

Cosmides, Leda & Tooby, John 1992: "Cognitive Adaptions for Social Exchange", pp. 163–228 in Barkow, Cosmides, and Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford UP.

Cosmides, Leda & Tooby, John 2015: "Adaptations for Reasoning about Social Exchange", pp. 625-68 in Buss, David (ed.), *The Handbook of Evolutionary Psychology*, Second edition, John Wiley & Sons.

Curry, Oliver Scott 2016: "Morality as Cooperation: A Problem-Centred Approach", pp. 27-51 in Shackelford & Hansen (eds.) *The Evolution of Morality*, Springer International Publishing.

Curry, Oliver Scott, Mullins, Daniel Austin, and Whitehouse, Harvey 2019: "Is it Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies", *Current Anthropology*, vol. 60, no. 1.

Cushman, Fiery, Young, Liane, and Hauser, Marc 2006: "The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm", pp. 1082-9 in *Psychological Science*, vol. 17.

Darley, John & Batson, Daniel 1973: "From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior", pp. 100-8 in *Journal of Personality and Social Psychology*, vol. 27, no. 1.

Darwin, Charles 1859: *On the Origin of Species by Means of Natural Selection*, Gramercy Books Random House 1979.

Davis, Lauren Cassani 2015: "Would You Pull the Trolley Switch? Does it Matter? The Lifespan of a Thought Experiment", *Atlantic*, October 9.

Demarest, Heather et al.: "Similarity and Enjoyment: Predicting Continuation for Women in Philosophy", pp. 525-41 in *Analysis,* vol. 77, no. 3.

Dennett, Daniel 1995: *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster.

DesChamps, Trent, Eason, Arianne, and Sommerville, Jessica 2016: "Infants Associate Praise and Admonishment with Fair and Unfair Individuals", pp 478-504 in *Infancy*, vol. 21, no. 4.

Doris, John 1998: "Persons, Situations, and Virtue Ethics", pp 504–40 in *Noûs*, vol. 32, no. 4.

Doris, John 2002: *Lack of Character: Personality and Moral Behavior*, Cambridge University Press.

Doris, John 2005: "Replies: Evidence and Sensibility" pp. 656-77 in *Philosophy and Phenomenological Research*, vol. 71, no. 3.

Downes, Stephen 2014: "Evolutionary Psychology", in *The Stanford Encyclopedia of Philosophy* (Edward N. Zalta (ed.), <https://plato.stanford.edu/entries/evolutionary-psychology/>.

Driver, Julia 2001: *Uneasy Virtue*, Cambridge University Press.

Driver, Julia 2006: "Virtue Theory," pp. 113-23 in Dreier, James (ed.), *Contemporary Debates in Moral Theory,* Blackwell.

Drożdżowicz, Anna 2018: "Philosophical Expertise beyond Intuitions", pp. 253-77 in *Philosophical Psychology*, vol. 31, no. 2.

Dunbar, Robin 1996: *Grooming, Gossip and the Evolution of Language*, Harvard University Press.

Edmunds, David 2014: *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us about Right and Wrong*, Princeton University Press.

Enoch, David 2010: "The Epistemological Challenge to Metanormative Realism: How Best to Understand it, and How to Cope with it", pp. 413-38 in *Philosophical Studies*, vol. 148, no. 3.

Enoch, David 2013: *Taking Morality Seriously: A Defense of Robust Realism*, Oxford University Press.

Everett, Jim, Pizarro, David, Crockett, Molly 2016: "Inference of Trustworthiness From Intuitive Moral Judgments", pp. 772–87 in *Journal of Experimental Psychology General*, vol. 145 no. 6.

Everett, Jim et al. 2018: "The Costs of Being Consequentialist: Social Perceptions of Those Who Harm and Help for the Greater Good", pp. 200-16 in *Journal of Experimental Social Psychology*, vol. 79.

Feltz, Adam & Cokely, Edward 2009: "Do Judgments about Freedom and Responsibility Depend on Who You Are? Personality Differences in Intuitions about Compatibilism and Incompatibilism", pp. 342-50 in *Consciousness and Cognition* vol. 18, no. 1.

Ferguson, Christopher 2010: "Genetic Contributions to Antisocial Personality and Behavior: A meta-Analytic Review from an Evolutionary Perspective", pp160-80 in *The Journal of Social Psychology*, vol. 150, no. 2.

Figdor, Carrie & Drabek, Matt 2016. "Experimental Philosophy and the Underrepresentation of Women" in Sytsma & Buckwalter (eds.) *A Companion to Experimental Philosophy*, Blackwell.

Fine, Cordelia 2017: *Testosterone Rex: Myths of Sex, Science, and Society*, W. W. Norton & Company.

FitzPatrick, William 2009: "Thomson's Turnabout on the Trolley", pp. 636-43 in *Analysis*, vol. 69, no. 4.

Flanagan, Owen 1991: *Varieties of Moral Personality: Ethics and Psychological Realism*, Harvard University Press.

Foot, Philippa 1967: "The Problem of Abortion and the Doctrine of Double Effect" pp. 5-15 in *Philosophical Review*, vol. 5.

Foot, Philippa 2001: *Natural Goodness*, Oxford University Press.

Frank, Robert 1988: *Passions within Reason: The Strategic Role of the Emotions*, W.W. Norton.

Frede, Dorothea 2013: "The Historic Decline of Virtue Ethics", pp 124-48 in Russell, Daniel (ed.) 2013.

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

Gettier, Edmund 1963: "Is Justified True Belief Knowledge?", pp 121-2 in *Analysis*, 23(6).

Ghelfi, Eric et al. 2020: "Reexamining the Effect of Gustatory Disgust on Moral Judgment: A Multilab Direct Replication of Eskine, Kacinik, and

Prinz (2011)", pp. 3–23 in *Advances in Methods and Practices in Psychological Science*, vol. 3, no. 1,

Gibbard, Alan 1990: *Wise Choices, Apt Feelings: A Theory of Normative Judgment*, Harvard University Press.

Global Study on Homicide: Trends, Contexts, Data, United Nations Office on Drugs and Crime, 2013.

Gould, Stephen Jay & Lewontin, Richard 1979: "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme", pp. 581-91 in *Proceedings of the Royal Society B*, 205.

Graham, Peter 2017: "Thomson's Trolley Problem", pp. 168-90 in *Journal of Ethics and Social Philosophy*, vol. 12, no. 2.

Greene, Joshua 2004: "Cognitive Neuroscience and the Structure of the Moral Mind", pp. 338-53 in Carruthers, Laurence, and Stich (eds.) *The Innate Mind: Structure and Contents*, Oxford University Press.

Greene, Joshua 2008: "The Secret Joke of Kant's soul", pp. 35-79 in. Sinnott-Armstrong (ed.) 2008c.

Greene, Joshua 2013: *Moral Tribes: Emotion, Reason And the Gap between Us and Them*, Penguin.

Greene, Joshua 2014: "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics", pp. 695-726 in *Ethics*, vol. 124, no. 4.

Greene, Joshua et al. 2001: "An fMRI Investigation of Emotional Engagement in Moral Judgment", pp. 2105–8 in *Science*, vol. 293, no. 5537.

Greene, Joshua et al. 2004: "The Neural Bases of Cognitive Conflict and Control in Moral Judgment, pp 389-400 in *Neuron*, vol. 44, no. 2.

Greene, Joshua et al. 2008: "Cognitive Load Selectively Interferes with Utilitarian Moral Judgment", pp. 1144-54 in *Cognition*, vol. 107, no. 3.

Greene, Joshua et al. 2009: "Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment", pp. 364-71 in *Cognition,* vol. 111 no. 3.

Griggs, Richard 2017: "Milgram's Obedience Study: A Contentious Classic Reinterpreted", pp. 32-7 in *Teaching of Psychology*, vol. 44, no. 1.

Haidt, Jonathan 2001: "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment, pp 814-34 in *Psychological Review*, vol. 108, no. ue4.

Haidt, Jonathan 2012: *The Righteous Mind: Why Good People are Divided by Politics and Religion*, Allen Lane.

Haidt, Jonathan, Koller, Silvia Helena, and Dias, Maria 1993: "Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?", pp. 613-28 in *Journal of Personality and Social Psychology,* vol. 65, no. 4.

Hamlin, J. Kiley 2013: "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core", pp. 186–93 in *Current Directions in Psychological Science*, vol. 22 no 3.

Harman, Gilbert 1999: "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error", pp. 315–31 in *Proceedings of the Aristotelian Society*, New Series, vol. 99.

Harman, Gilbert 2009: "Skepticism about Character Traits", pp. 235-42 in *Journal of Ethics*, vol. 13, no. 2-3.

Harris, Sam 2000: *The Moral Landscape: How Science can Determine Human Values*, Free Press.

Hartman, Robert & Bronson, Joshua: "Consequentialism and Virtue" forthcoming in Halbig and Timmermann (eds.) *The Handbook of Virtue and Virtue Ethics*, Springer

Hartman, Robert 2017: *In Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness*, Routledge.

Hartshorne, Hugh & May, Mark 1928: *Studies in the Nature of Character*, [part] 1: "Studies in Deceit": book 1, "General Methods and Results"; book 2, "Statistical Methods and Results", MacMillan.

Hauser, Marc 2006: *Moral Minds: How Nature Designed a Universal Sense of Right and Wrong*, Ecco Press/Harper Collins.

Hauser, Marc et al. 2007: "A Dissociation Between Moral Judgments and Justifications", pp. 1-21 in *Mind & Language*, vol. 22 no. 1.

Heaney, Megan, Gray, Russell, Taylor, Alex 2017: "Kea Show No Evidence of Inequity Aversion", pp. 1-8 in *Royal Society Open Science*, vol. 4, no. 3.

Henrich, Joseph and Silk, Joan 2013: "Interpretative Problems with Chimpanzee Ultimatum Game", p E3049 in *PNAS* (*Proceedings of the National Academy of Sciences*) August 13, vol. 110, no. 33.

Hollander, Matthew & Turowetz, Jason 2017: "Normalizing Trust: Participants' Immediately Posthoc Explanations of Behaviour in Milgram's 'obedience' experiments", pp. 655-74 in *British Journal of Social Psychology*, vol. 56, no 1.

Holtzman, Geoffrey 2016: Rejecting Beliefs, or Rejecting Believers? On the Importance and Exclusion of Women in Philosophy., pp. 293-312 in *Hypatia*, vol 31, no. 2

Huemer, Michael 2005: *Ethical Intuitionism*, Palgrave, 2005.

Huemer, Michael 2007: "Compassionate Phenomenal Conservatism", pp. 30-55 in *Philosophy and Phenomenological Research*, vol. 74, no. 1.

Hume, David, 1739/40: *Treatise of Human Nature*, L. A. Selby-Bigge (Editor), second edition with revisions and notes by P. H. Nidditch, Oxford University Press 1978.

Hurka, Thomas 2006: "Virtuous Act, Virtuous Dispositions" pp. 69-76 in *Analysis*, vol. 66, no. 289.

Hursthouse, Rosalind 1999: *On Virtue Ethics*, Oxford University Press.

Hursthouse, Rosalind 2013: "Virtue Ethics", in Zalta, Edward (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), <http://plato.stanford.edu/archives/fall2013/entries/ethics-virtue/>

Ichikawa, Jonathan Jenkins and Steup, Matthias 2018: "The analysis of knowledge", in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), <https://plato.stanford.edu/entries/knowledge-analysis>.

Isen, Alice & Levin, Paula 1972: "The Effect of Feeling Good on Helping: Cookies and Kindness", pp 384-8 in *Journal of Personality and Social Psychology*, vol. 21, no. 3.

Jaquet, François 2018: "Evolution and Utilitarianism", pp. 1151-61 in *Ethical Theory and Moral Practice*, vol. 21.

Joyce, Richard 2006: *The Evolution of Morality*, MIT Press.

Joyce, Richard 2007: "Is Human Morality Innate?", in Carruthers, Laurence, and Stich (eds.) *The Innate Mind: Culture and Cognition*, , Oxford University Press.

Joyce, Richard 2008: "What Neuroscience Can (And Cannot) Contribute to Metaethics", pp. 371-94 in Sinnott-Armstrong (ed.) 2008c.

Kahane, Guy 2011: "Evolutionary Debunking Arguments", pp. 103-25 in *Noûs*, vol. 45, no. 1.

Kahane, Guy 2014: "Evolution and Impartiality", pp. 327-41 in *Ethics*, vol. 124, no. 2.

Kahane, Guy et al. 2018: "Beyond Sacrificial Harm: A Two-Dimensional Model of Utilitarian Psychology", pp. 131-64 in *Psychological Review*, vol. 125, no. 2.

Kahneman, Daniel 2011: *Thinking, Fast and Slow*. Farrar, Straus, & Giroux.

Kamm, Frances 2000: "Nonconsequentialism", pp. 205-26, in LaFollette (ed.), *The Blackwell Guide to Ethical Theory*, Blackwell.

Kamm, Frances 2007: *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*, Oxford University Press.

Kamm, Frances 2009: "Neuroscience and Moral Reasoning: A Note on Recent Research", pp. 330-45 in *Philosophy & Public Affairs,* vol. 37, no. 4.

Kamtekar, Rachana 2004: "Situationism and Virtue Ethics on the Content of our Character", pp. 458-91 in *Ethics*, vol 114 no. 3.

Kauppinen, Antti: 2007: "The Rise and Fall of Experimental Philosophy", pp. 95-118 in *Philosophical Explorations*, vol. 10, no. 2, June.

Kauppinen, Antti forthcoming: "Who Should Bear the Risk When Self-Driving Vehicles Crash?", *The Journal of Applied Philosophy*.

Kelly, Daniel et al. 2007: "Harm, Affect, and the Moral/Conventional Distinction" pp. 117-31 in *Mind & Language*, vol. 22 no. 2.

Khemiri Lotfi et al. 2012: "Alcohol Dependence Associated with Increased Utilitarian Moral Judgment: a Case Control Study", pp. 1-8 in *PLoS One*, vol. 7 no. 6.

Koenigs, Michael et al. 2007: "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements", pp. 908-11 in *Nature*, vol 446 no. 7138.

Koenigs, Michael et al 2012: "Utilitarian Moral Judgment in Psychopathy", pp. 708-14 in *Social, Cognitive and Affective Neuroscience*, vol. 7 no. 6.

Kornreich, Charles et al. 2017: "Conditional Reasoning in Schizophrenic Patients", pp. 1-8 in *Evolutionary Psychology*, July-September.

Kumar, Victor 2015: "Moral Judgment as a Natural Kind", pp. 2887-2910 in *Philosophical Studies*, vol. 172, no. 11.

Larsen, Randy & Buss, David 2017: *Personality Psychology: Domains of Knowledge about Human Nature*, 6th ed., McGraw-Hill Education.

Latané, Bibb & Darley, John 1970: *The Unresponsive Bystander: Why Doesn't He Help?*, Appleton-Century-Crofts.

Latané, Bibb & Rodin, Judith 1969: "A Lady in Distress: Inhibiting Effects of Friends and Strangers on Bystander Intervention", pp. 189-202 in *Journal of Experimental Psychology*, vol. 5, no. 2.

Laurence, Stephen & Margolis, Eric 2013: "In Defense of Nativism", pp 693–718 in *Philosophical Studies*, vol. 165, no. 2.

Lazari-Radek, Katarzyna & Singer, Peter 2012: "The Objectivity of Ethics and the Unity of Practical Reason", pp. 9-31 in *Ethics*, vol. 123, no. 1.

Lazari-Radek, Katarzyna & Singer, Peter 2014: *The Point of View of the Universe: Sidgwick and Contemporary Ethics*, Oxford University Press.

Locke, John 1690: *An Essay Concerning Human Understanding*, abridged and edited by John W. Yolton, Everyman 1993.

Luncz, Lydia et al. 2018: "Costly Culture: Differences in Nut-Cracking Efficiency between Wild Chimpanzee Groups", pp. 63-73 in *Animal Behaviour*, vol. 137.

MacAskill, William, Mogensen, Andreas, and Ord, Toby: "Evolution, Utilitarianism, and Normative Uncertainty: the Practical Significance of Debunking Arguments", manuscript.

Machery, Edouard 2010: "The Bleak Implications of Moral Psychology", pp. 223-31 in *Neuroethics*, vol. 3.

Machery, Edouard et al. 2004: "Semantics, Cross-Cultural Style", pp. 1-12 in *Cognition*, vol. 92, no. 3.

Machery, Edouard et al. 2015: "Gettier across Cultures", pp. 645-64 in *Noûs*, vol. 51, no. 3.

Machery, Edouard et al. 2017: "The Gettier Intuition from South America to Asia", pp. 517–41 in *Journal of Indian Council of Philosophical Research*, volume 34.

MacIntyre, Alasdair 1984: *After Virtue: A Study in Moral Theory*, 2nd ed., University of Notre Dame Press.

Mackie, John 1977: *Ethics: Inventing Right and Wrong*, Penguin.

Malik, Salma et al. 2019: "Genetics of Autism Spectrum Disorder: An Update", pp 109-14 in *Psychiatric Annals*, vol. 49, no. 3.

Mameli, Matteo & Bateson, Patrik 2011: "An Evaluation of the Concept of Innateness", pp. 436-43 in *Philosophical Transactions of the Royal Society: Biological Sciences*, vol. 366, no. 1563.

Melnikoff, David & Bargh, John 2018: "The Mythical Number Two", pp. 280-93 in *Trends in Cognitive Sciences*, vol. 22, no. 4 (April).

Merritt, Maria 2000: "Virtue Ethics and Situationist Personality Psychology", pp. 365-83 in *Ethical Theory and Moral Practice*, vol. 3, no. 4.

Mikhail, John 2007: "Universal Moral Grammar: Theory, Evidence, and the Future", pp. 143-52 in *Trends in Cognitive Sciences*, vol. 11.

Mikhail, John 2008: "The Poverty of the Moral Stimulus", pp. 353-60 in Sinnott-Armstrong (ed.) 2008a.

Mikhail, John 2011: *Elements of Moral Cognition: Rawls's Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge University Press.

Milgram, Stanley 1963: "Behavioral Study of Obedience, pp. 371-8 in *The Journal of Abnormal and Social Psychology*, vol. 67, no. 4.

Milgram, Stanley 1974: *Obedience to Authority: An Experimental View*, with preface by Jerome Bruner, Pinter & Martin 2005

Miller, Christian 2013: *Moral Character: An Empirical Theory*, Oxford University Press.

Miller, Christian 2014: *Character and Moral Psychology*. Oxford University Press.

Miller, Christian 2017: *The Character Gap: How Good are We?*, Oxford University Press.

Miller, Christian 2020: "Empirical Approaches to Moral Character", in Zalta, Edward (ed.), *The Stanford Encyclopedia of Philosophy*, URL < https://plato.stanford.edu/archives/fall2020/entries/moral-character-empirical/>

Miller, Geoffrey 2001: *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*, Vintage.

Miller, Geoffrey 2007: "Sexual Selection for Moral Virtues" pp. 97-125 in *The Quarterly Review of Biology*, vol. 82, no. 2.

Mischel, Walter 2009: "Becoming a Cumulative Science", presidential column, Association for Psychological Science, January 1.

Mitchell, Kevin 2018: *Innate: How the Wiring of Our Brains Shapes Who We Are*, Princeton University Press.

Mogensen, Andreas 2014: *Evolutionary Debunking Arguments in Ethics*, University of Oxford D.Phil thesis.

Montealegre, Andreas et al.: "Does Maximizing Good Make People Look Bad?", manuscript.

Moore, George Edward 1903: *Principia Ethica*, Cambridge University Press 1976.

Naguib, Marc & Oers, Kees van 2013: "Avian Personality" in Carere & Maestripieri (eds.) *Animal Personalities: Behavior, Physiology, and Evolution*, University of Chicago Press.

Nosek, Brian et al. 2015: "Estimating the Reproducibility of Psychological Science, *Science*, vol. 349, no. 6251.

Nosek, Brian et al. 2018: "The Preregistration Revolution", pp. 2600-2606 in *Proceedings for the National Academy of Sciences (PNAS)*, vol. 115, no. 11.

Nyholm, Sven & Smids, Jilles 2016: "The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?", pp. 1275-89 in *Ethical Theory and Moral Practice*, vol. 19, no. 5.

Olson, Jonas 2014: *Moral Error Theory: History, Critique, Defence*, Oxford University Press.

Parfit, Derek 2011a: *On What Matters*, vol. I, Oxford University Press.

Parfit, Derek 2011b: *On What Matters*, vol. II, Oxford University Press.

Paxton, Joseph, Ungar, Leo, and Greene, Joshua 2012: "Reflection and Reasoning in Moral Judgment", pp. 163-77 in *Cognitive Science*, vol. 36, no. 1.

Pederson, Amy, King, James., & Landau, Virginia 2005: "Chimpanzee (pan troglodytes) Personality Predicts Behavior", pp. 534-49 in *Journal of Research in Personality*, vol. 39, no. 5.

Perry, Gina 2012: *Behind the Shock Machine: The Untold Story of the Notorious Milgram Psychology Experiments*, Scribe.

Pinker, Steven 1994: *The Language Instinct*, William Morrow & Co.

Plato: *The Republic*, translated by Benjamin Jowett, Vintage Books 1991.

Plomin, Robert 2018: *Blueprint: How DNA Makes Us Who We Are*, Allen Lane 2018.

Plomin, Robert, & Daniels, Denise 2011: "Why Are Children in the Same Family so Different from One Another?", pp. 563-82 in *International Journal of Epidemiology*, vol. 40, no. 3.

Plomin, Robert et al. 2016: "Top 10 Replicated Findings from Behavioral Genetics", pp. 3-23 in *Perspectives on Psychological Science*, vol. 11, no. 1.

Price, Richard 1787: *A Review of the Principal Questions in Morals*, 3rd ed, Oxford University Press 1974.

Prinz, Jesse 2007: The Emotional Construction of Morals, Oxford University Press.

Prinz, Jesse 2008: "Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Cushman", pp. 157-70 in Sinnott-Armstrong (ed.) 2008b.

Prinz, Jesse 2009: "Against Moral Nativism", pp. 167-89 in Murphy & Bishop (eds.), *Stich and His Critics*, Blackwell.

Prinz, Jesse 2012: *Beyond Human Nature: How Culture and Experience Shape our Lives*, Penguin.

Prinz, Jesse 2014: "Where Do Morals Come From? – A Plea for a Cultural Approach", pp. 99-116 in Christen, Markus et al. (eds.), *Empirically Informed Ethics: Morality between Facts and Norms*, Springer.

Proto, Eugenio, Rustichini, Aldo & Sofianos, Andis 2019: "Intelligence, personality and gains from cooperation in repeated interactions", pp 1351-90 in *Journal of Political Economy*, vol 127, no. 3.

Quine, Willard van Orman 1951: "Two Dogmas of Empiricism", pp. 20–43 in *Philosophical Review*, vol. 60, no. 1.

Rachels, James 1990: *Created from Animals: The Moral Implications of Darwinism*, Oxford University Press.

Rachul, Christen & Zarzeczny, Amy 2012: "The Rise of Neuroskepticism", pp. 77-81 in *International Journal of Law and Psychiatry*, vol. 35, no. 2.

Raine, Adrian 2013: *The Anatomy of Violence: The Biological Roots of Crime*, Pantheon Books.

Rawls, John 1971: *A Theory of Justice*, Harvard University Press.

Ritchie, Stuart 2020: *Science Fictions: Exposing Fraud, Bias, Negligence and Hype in Science*, Penguin Random House.

Ross, Lee & Nisbett, Richard 2011: *The Person and The Situation: Perspectives of Social Psychology*, 2nd ed., Pinter & Martin.

Ross, W. D. 1930: *The Right and the Good*, Oxford University Press 2012.

Ruse, Michael 1985: *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*, Blackwell.

Ruse, Michael & Richards, Robert (eds.) 2017: *The Cambridge Handbook of Evolutionary Ethics*, Cambridge University Press.

Russell, Bertrand 1946: *A History of Western Philosophy and its Connection with Political and Social Circumstances from the Earliest Times to the Present Day*, Routledge 1991.

Russell, Daniel (ed.) 2013: *The Cambridge Companion to Virtue Ethics*, Cambridge University Press.

Sabini, John & Silver, Maury 2005: "Lack of Character? Situationism Critiqued", pp. 535-62 in *Ethics*, vol. 115, no. 3.

Sandin, Sven 2014: "The Familial Risk of Autism", pp. 1770-7 in *JAMA*, vol. 311, no. 17.

Sargent, Michael 2004: "Less Thought, More Punishment: Need for Cognition Predicts Support for Punitive Responses to Crime", pp. 1485-93 in *Personality and Social Psychology Bulletin,* vol. 30, no. 11.

Schmitt, David & Pilcher, June 2004: "Evaluating Evidence of Psychological Adaptation: How Do We Know One When We See One?", pp 643-9 in *Psychological Science*, vol. 15, no 10.

Schnall, Simone et al. 2008: "Disgust as Embodied Moral Judgment", pp. 1096-1109 in *Personality and Social Psychology Bulletin*, vol. 34, no. 8.

Schneewind, Jerome B. 1994: "Locke's Moral Philosophy", pp. 199-225 in Chappell, Vere (ed.), *The Cambridge Companion to Locke*, Cambridge University Press.

Schroeder, Mark 2010: *Slaves of the Passions*, Oxford University Press.

Schuett, Wiebke, Dall, Sasha, and Royle, Nick 2011: "Pairs of Zebra Finches with Similar 'Personalities' Make Better Parents", pp. 609-18 in *Animal Behaviour*, vol. 81, no. 3.

Sellars, Wilfred 1962: "Philosophy and the Scientific Image of Man", pp. 35-78 in Colodny (ed.) *Frontiers of Science and Philosophy*, University of Pittsburgh Press.

Seyedsayamdost, Hamid 2015: "On Gender and Philosophical Intuition: Failure of Replication and other Negative Results", pp. 642-73 in *Philosophical Psychology*, vol. 28, no. 5.

Shafer-Landau, Russ 2012: "Evolutionary Debunking, Moral Realism, and Moral Knowledge", pp. 1-37 in *Journal of Ethics & Social Philosophy*, vol. 7, no. 1

Shenhav, Amitai & Greene, Joshua 2014: "Integrative Moral Judgment: Dissociating the Roles of the Amygdala and the Ventromedial Prefrontal Cortex", pp. 4741-49 in *Journal of Neuroscience* vol. 34, no. 13.

Simpsons 1994: episode 114, "Fear of Flying".

Sidgwick, Henry 1907: *The Methods of Ethics*, 7th edition, Hacket 1981.

Singer, Peter 1972: "Famine, Affluence, and Morality", pp. 229-43 in *Philosophy & Public Affairs*, vol. 1, no. 3.

Singer, Peter 1981 *The Expanding Circle: Ethics and Sociobiology*, Oxford University Press. Reissued with a new forward in 2011 as *The Expanding Circle: Ethics, Evolution, and Moral Progress*.

Singer, Peter 2005: "Ethics and Intuitions", pp. 331-52 in *Journal of Ethics*, vol 9, no. 3-4

Singer, Peter 2007: "Should We Trust Our Moral Intuitions?", *Project Syndicate*, March 14.

Sinnott-Armstrong, Walter (ed.) 2008a: *Moral Psychology volume 1: The Evolution of Morality: Adaptations and Innateness*, MIT Press.

Sinnott-Armstrong, Walter (ed.) 2008b: *Moral Psychology volume 2: The Cognitive Science of Morality: Intuition and Diversity*, MIT Press.

Sinnott-Armstrong, Walter (ed.) 2008c: *Moral Psychology volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, MIT Press.

Sinnott-Armstrong, Walter & Miller, Christian (eds.) 2017 *Moral Psychology: Virtue and Character*, vol. 5, MIT Press.

Sivan, Jonathan, Curry, Oliver Scott & van Lissa, Caspar 2018: "Excavating the Foundations: Cognitive Adaptations for Multiple Moral Domains", pp 408-19 in *Evolutionary Psychological Science*, volume 4.

Skyrms, Brian 2003: *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press.

Slote, Michael 2005: "Moral Sentimentalism and Moral Psychology" pp. 219-39 in Copp (ed.), *The Oxford Handbook of Ethical Theory*, Oxford University Press.

Slote, Michael 2007: *The Ethics of Care and Empathy*, Routledge.

Snow, Nancy (ed.) 2018: *The Oxford Handbook of Virtue*, Oxford University Press.

Sober, Elliott 2018: *Philosophy of Biology*, 2nd ed., Routledge.

Southwood, Nicholas 2011: "The Moral/Conventional Distinction", pp. 761-802 in *Mind*, vol. 120, no. 479.

Spelke, Elizabeth & Kinzler, Katherine 2007: "Core Knowledge", pp. 89-96 in *Developmental Science,* vol. 10, no. 1.

Spelke, Elizabeth 1990: "Principles of Object Perception," pp. 29-56 in *Cognitive Science,* vol. 14, no. 1.

Srinivasan, Amia 2015: "The Archimedean Urge", pp. 325-362 in *Philosophical Perspectives*, vol. 29, no. 1.

Sterelny, Kim & Fraser, Ben 2017: "Evolution and Moral Realism", pp. 981–1006 in *The British Journal for the Philosophy of Science*, vol. 68, no. 4.

Stevenson, Charles 1944: *Ethics and Language*, Yale University Press.

Stich, Stephen 1990: *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*, MIT Press

Street, Sharon: "A Darwinian Dilemma for Realist Theories of Value", pp. 109-66 in *Philosophical Studies*, vol. 127, no. 1.

Strohminger, Nina, Lewis, Richard., and Meyer, David 2011: "Divergent Effects of Different Positive Emotions on Moral Judgment", pp 295–300 in *Cognition*, vol. 119, no. 2.

Suter, Renata & Hertwig, Ralph 2011: "Time and Moral Judgment, pp. 454-8 in *Cognition*, vol. 119 no. 3.

Svensson, Frans 2010: "Virtue Ethics and the Search for an Account of Right Action", pp. 255–71 in *Ethical Theory and Moral Practice*, vol. 13, no. 3.

Svensson, Frans 2011: "Eudaimonist Virtue Ethics and Right Action: A Reassessment", pp. 321–39 in *Journal of Ethics,* vol. 15, no. 4.

Svensson, Frans & Johansson, Jens 2018: "Objections to Virtue Ethics" Snow (ed.) *The Oxford Handbook of Virtue*, Oxford University Press

Tersman, Folke 2008: "The Reliability of Moral Intuitions: A Challenge from Neuroscience", pp. 389-405 in *Australasian Journal of Philosophy*, vol. 86, no. 3.

Tersman, Folke 2017: "Debunking and Disagreement, pp. 754-74 in *Noûs*, vol. 51, no. 4.

Texier, Thibault Le 2019: "Debunking the Stanford Prison Experiment", pp. 823-39 in *American Psychologist*, vol. 74, no. 7

Thomas, Bradley, Croft, Katie, & Tranel, Daniel 2011: "Harming Kin to Save Strangers: Further Evidence for Abnormally Utilitarian Moral Judgments after Ventromedial Prefrontal Damage, pp 2186-96 in *Journal of Cognitive neuroscience*, vol. 23 no. 9.

Thomson, Judith Jarvis 1985: "The Trolley Problem", pp. 1395-1415 in *The Yale Law Journal*, vol. 94, no. 6.

Thomson, Judith Jarvis 1997: "The Right and The Good" pp. 273-298 in *Journal of Philosophy*, vol. 94 no. 6.

Thomson, Judith Jarvis 2008: "Turning the Trolley", pp. 359-74 in *Philosophy & Public Affairs*, vol. 36, no. 4.

Timmons, Mark 2008: "Toward a Sentimentalist Deontology", pp. 93-103 in Sinnott-Armstrong (ed.) 2008c.

Tosi, Justin & Warmke, Brandon 2020*: Moral Grandstanding: The Use and Abuse of Moral Talk*, Oxford University Press.

Tracy, Jessica, Steckler, Conor, and Heltzel, Gordon 2019: "The Physiological Basis of Psychological Disgust and Moral Judgments", pp. 15-32 in *Journal of Personality and Social Psychology*, vol. 116, no. 1.

Trivers, Robert 2011: *Deceit and Self-Deception: Fooling Yourself the Better to Fool Others*, Penguin Books.

Tversky, Amos & Kahneman, Daniel 1981: "The Framing of Decisions and the Psychology of Choice", pp. 453-8 in *Science*, vol. 211, no. 4481.

Unger, Peter 1996: *Living High and Letting Die: Our Illusion of Innocence,* Oxford University Press.

Valdesolo, Piercarlo and DeSteno, David 2006: "Manipulations of Emotional Context Shape Moral Judgment," pp. 476-7 in *Psychological Science*, vol. 17, no. 7.

Van de Vondervoort, Julia, & Hamlin, Kiley 2019: "The Infantile Roots of Sociomoral Evaluations", pp. 402-12 in Gray & Graham (eds.), *The Atlas of Moral Psychology*, Guilford Press.

Wason, Peter Cathcart 1968: "Reasoning about a Rule", pp. 273-81 in *Quarterly Journal of Experimental Psychology*, vol. 20, no. 3.

Watkins, Michael 1984: "Models as Toothbrushes", p. 86 of *Behavioral & Brain Sciences*, vol. 7, no. 1.

Weinberg, Jonathan, Nichols, Shaun, and Stich, Stephen 2001: "Normativity and Epistemic Intuitions", pp. 429–60 in *Philosophical Topics*, vol. 29, no 1-2.

Weinstein, David, "Herbert Spencer", in Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* <https://plato.stanford.edu/archives/fall2019/entries/spencer/>

Weiss, Alexander et al. 2017: "Personality in the Chimpanzees of Gombe National Park", *Nature: Scientific Data*, 4, 170146.

Wheatley, Thalia and Haidt, Jonathan 2005: "Hypnotic Disgust Makes Moral Judgments More Severe", pp. 780-4 in *Psychological Science,* vol. 16, no. 10.

White, Roger 2010: "You Just Believe that Because …", pp. 573-615 in *Philosophical Perspectives*, Volume 24, no. 1.

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

Williamson, Timothy 2020: *Philosophical Method: A Very Short Introduction*, Oxford University Press.

Wilson, Edward O. 1975: *Sociobiology: The New Synthesis*, Harvard University Press.

Wilson, Michael et al. 2014: "Lethal Aggression in Pan Is Better Explained by Adaptive Strategies than Human Impacts", pp. 414-7 in *Nature* vol. 513, September.

Woodward, Amanda 1998: "Infants Selectively Encode the Goal Object of an Actor's Reach,", pp. 1-34 in *Cognition* vol. 69, no. 1.

Yuko, Elizabeth 2017: "How The Good Place Goes beyond 'the Trolley Problem'", *The Atlantic*, October 21.

Zimbardo, Philip 2007: *The Lucifer Effect: Understanding How Good People Turn Evil*, Random House.

8 Index of names

ISBN 978-91-7911-354-4

Department of Philosophy

Stockholm University