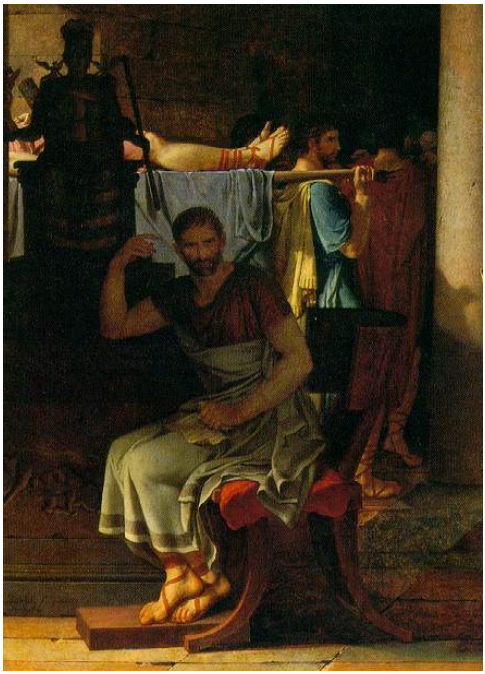
Antti Kauppinen

Essays in Philosophical Moral Psychology



Philosophical Studies from the University of Helsinki 19

Filosofisia tutkimuksia Helsingin yliopistosta
Filosofiska studier från Helsingfors universitet
Philosophical studies from the University of Helsinki

Publishers:

Department of Philosophy
Department of Social and Moral Philosophy
P. O. Box 9 (Siltavuorenpenger 20 A)
00014 University of Helsinki
Finland

Antti Kauppinen

Essays in Philosophical Moral Psychology

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Social Sciences of the University of Helsinki, for public examination in lecture room XII, University Main Building,
on 9 January 2008, at 12 noon.

ISBN 978-952-10-4474-8 (paperback)
ISBN 978-952-10-4475-5 (PDF)
ISSN 1458-8331

Helsinki University Print
Helsinki 2007

TABLE OF CONTENTS

Acknowledgements

People who don't enjoy the support of family, friends, and colleagues don't have to worry about writing an acknowledgements section – mostly because they don't finish their dissertations in the first place. I would certainly not have made it to the end without the support of many others. First of all, I want to thank Timo Airaksinen, who supervised my entire project and funded a large part of it through the Academy of Finland research project *Practical Reason and Moral Motivation*. I also want to note the support of Raimo Tuomela, who helped me along early on in my graduate studies by allowing me to participate in his *Research Project on Social Action*, even though my research interests were not an exact match. I am very grateful for that. Earlier yet, Heta Gylling and Matti Häyry hired me as a research assistant, which had the fortunate side effect that I got to know their work quite well and hopefully learn a little. Also at the Department of Moral and Social Philosophy in Helsinki, Tuula Pietilä, Karolina Kokko-Uusitalo, and Anu Kuusela offered invaluable help in practical matters throughout my studies.

I was very lucky to work on a dissertation on metaethics at a time in which I could count on the support of exceptionally talented fellow graduate students working on similar issues. Within our research project, I had innumerable conversations with Jussi Suikkanen and Teemu Toppinen on fundamental questions and current debates. They also gave me detailed feedback on all the papers I published, and on some papers I did not even try to publish because they convinced me my argument just was not good enough. I shudder to think what my metaethical views would be had I not come to know them! Of other graduate students, Pekka Mäkelä worked on partially overlapping themes like moral responsibility, and always had time to give me some sage advice. Pilvi Toppinen and Jarno Rautio helped with political philosophy, and Tomi Kokkonen and Jaakko Kuorikoski never shied away from philosophical or political argument, whether or not any of us knew anything about the topic. Vili Lähteenmaki and Tuukka Tanninen tolerate no non-sense, and forced me to clarify my ideas whenever

they came up in conversation. Merja Mähkä and Sanna Nyqvist each provided an intelligent outside perspective on the philosophical issues I worked on. Heikki Ikäheimo and Arto Laitinen from Jyväskylä recognized me, and I want to recognize them in turn. Finally, of people who studied with me in Helsinki, Pekka Väyrynen deserves a special mention here. I doubt if I would ever have got interested in metaethics without his towering example. From the very start, Pekka held himself to world-class standards, and met them; from him, I learned to ask more of myself, and though I have never risen to the level I should have, I would surely be even less of a philosopher had I not tried to follow in Pekka's footsteps.

Much of my graduate work was done during visiting scholarships at Florida State, Pittsburgh, and Chapel Hill. In Tallahassee, Piers Rawling was most kind to invite me and moreover, with David McNaughton, teach me everything I know about deontology. Al Mele, Thomas Nadelhoffer, Eddy Nahmias, and Jason Turner defended experimental philosophy, with the result that I wrote two papers criticizing it. Virginia Tice helped considerably with homesickness. The following year, Robert Brandom invited me to Pittsburgh, and taught an inspiring seminar on inferentialism. John McDowell was kind enough to have several thorough conversations on his work and read some of mine, and my fellow Finn Hille Paakkunainen helped me navigate the social world. After a year in Finland, I was able to strong-arm Geoff Sayre-McCord to invite me to Chapel Hill for a full year. This turned out to be one of the smartest things I had ever done. Not only did I enjoy Geoff's metaethics reading group, but also had a chance to attend wonderful seminars by, among others, Susan Wolf, Richard Kraut, Jesse Prinz, and my arch-enemy Josh Knobe. These seminars and conversations gave me the push I needed to write my introduction and finally finish the dissertation. It is no surprise that their graduate students like Ben Bramble and Sven Nyholm have become such great people to talk metaethics with, and tolerable pool-players as well.

Finally, I owe special thanks to three people. Lilian O'Brien first showed me how to write a dissertation, and then helped me work out my views. This is no place to discuss her other virtues; suffice it

to say that since I came to know her through philosophy, I am glad I do philosophy. That's as sentimental as I'm going to get.

My parents, Pirkko and Tapio, never had the chance to go to university, nor would it have made sense given their background and interests. That makes it somewhat unlikely, sociologically speaking, that I should have become a doctor of philosophy and an academic professional. Yet it is they who put me on the path that I took, from my mother teaching me to read and to love the library to the long philosophical and historical and psychological conversations I used to have with my father when we should already have been sleeping. I know of no other people who would possess an equal amount of practical wisdom, in every sense of the word. I am still struggling to learn from them.

I would dedicate this work to my parents, but I think I will wait for my first real book.

St Andrews, Scotland, December 2007

Antti Kauppinen

List of original publications

This dissertation consists of the following publications:

I The Rise and Fall of Experimental Philosophy
Published in *Philosophical Explorations* 10 (2), 95–112, June 2007.

II Moral Judgment and Volitional Incapacity
To be published in *Topics in Contemporary Philosophy vol. 7: Action, Ethics, and Reponsibility*, eds. Joseph Keim Campbell, Michael O'Rourke, and Harry S. Silverstein. MIT Press 2008.

III The Social Dimension of Autonomy
To be published in *The Critical Theory of Axel Honneth*, ed. Danielle Petherbridge. Brill, Leiden 2008.

IV Reason, Recognition, and Internal Critique
Published in *Inquiry* 45 (4), 479–498, December 2002.

Introduction

1 What Is Philosophical about Philosophical Moral Psychology?

> Psychology is no more closely related to philosophy than any other natural science.
>
> Ludwig Wittgenstein, *Tractatus Logico-Philosophicus* 4.1121

> Throughout the twentieth century, philosophical work in metaethics largely ignored the psychological literature on moral judgment. … Over the last twenty years, a tradition in moral psychology has developed that really does, I will maintain, help us understand the nature of moral judgment.
>
> Shaun Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment*, 4

> Thus, the student of politics must study the soul, but he must do so with his own aim in view, and only to the extent that the objects of his inquiry demand: to go into it in greater detail would perhaps be more laborious than his purposes require.
>
> Aristotle, *Nicomachean Ethics* 1102a

This dissertation consists in four essays on the necessary psychological conditions of moral judgment and moral responsibility. In addition to examining these psychological conditions and their implications for normative theories, some of the

papers discuss the proper methodology of such investigation. The emphasis on methodology is surely warranted, for the nature of claims in this area is apt to be particularly confusing, as much recent work in empirical psychology unintentionally shows. Philosophical theses and arguments for them are easily mistaken for empirical claims, unless they are formulated with particular care. Unfortunately, philosophers have often not been careful enough. Take a recent attempt at formulating the subject matter of philosophical moral psychology by Jay Wallace:

> Moral psychology […] explores a variety of psychological phenomena through the unifying prism of a concern for normativity. It studies the psychological conditions for the possibility of binding norms of action; the ways in which moral and other such norms can be internalized and complied with in the lives of agents; and a range of psychological conditions and formations that have implications for the normative assessment of agents and their lives. (Wallace 2006, 87)

Wallace is in effect saying that moral psychology studies what it takes to be a moral *subject*, someone who makes judgments about right and wrong, on the one hand, and to be an *object* of evaluative assessment, on the other. As we will see in the next sections, I believe this is along the right tracks, but talk of 'exploring psychological phenomena' or 'studying psychological formations' is dangerously ambiguous. It is very natural to read it as suggesting an empirical investigation into how human beings internalize norms, make moral judgments, or engage in moral reasoning, for example. And indeed, an increasing number of philosophers and psychologists are treating these questions as purely empirical ones. Discussing the role of emotion in moral judgment, Jesse Prinz formulates the view with exceptional clarity:

> Do our ordinary moral concepts (the ones we deploy in token thoughts most frequently) have an emotional component? This is essentially an empirical question. It's a question about what goes on in our heads when we use moral terms like 'good' and 'bad' or 'right' and 'wrong'. (Prinz 2006, 30)

If Prinz is right about the nature of the questions in moral psychology, there are two ways to look for answers: we can either speculate from the armchair what people are like, relying on introspection and anecdotal observation, or we can perform or make use of controlled psychological, neuroscientific, and social psychological experiments. After all, how else do you 'explore psychological phenomena'? Given these options, it is obvious which one to take. Empirical truths about the human mind, as well as anything else, are best discovered through the use of the scientific method and scientific evidence. Thus Prinz, for example, defends the importance of emotion to moral judgment by appeal to fMRI studies that show brain activity in areas associated with emotion while they make moral judgments or hear morally loaded stories, studies that indicate that emotions influence which subject matters people moralize about, and studies that claim that violent psychopathy is explained by a lacking capacity for negative emotions and consequently empathy and guilt.[1]

Yet philosophers studying moral judgment and moral responsibility have not, by and large, made use of such evidence. As John Doris and Stephen Stich put it,

> Until recently, the moral psychology of *philosophy* departments has been largely speculative; prominent empirical claims — about the structure of character, say, or the nature of moral reasoning — have seldom been subject to systematic empirical scrutiny. (Doris and Stich 2006)

Have philosophers simply been irresponsible in turning away from brain scanners, experimental microeconomics, psychology laboratories, and surveys? Or is it possible that the questions that philosophers have been asking are of a very different kind? As this dissertation makes clear, I believe the latter is correct. There are philosophical questions and philosophical methods that are in an

[1] I discuss these data in section 2.3 below.

important way discontinuous with scientific ones, although the latter are not irrelevant to philosophy either. But what is the difference?

To begin with a relatively clear case, when it comes to normative ethics, we have a pretty good idea about how to distinguish between philosophical and non-philosophical questions. Bracketing for the time being subjectivism and social relativism about ethics, it is one thing to ask what people think is good or right and another to ask what *is* good or right. Here, the distinction between the philosophical and the empirical coincides with that between the *normative* and the *descriptive*. (Correspondingly, it is blurred, if at all, to the extent that the distinction between the normative and the descriptive is blurred.) This is not to say that psychological data are irrelevant to normative ethics. First, insofar as normative theories are meant to describe an ideal that could actually guide us, we must be able to actually live up to that ideal, and that depends on what we are like. Typically, ethical theories are not meant for the benefit of angels, but ordinary human beings, and that gives rise to a kind of constraint by facts about our psychological capacities, as people like Bernard Williams (1981a) and Owen Flanagan (1991) have argued. Second, normative theories themselves involve various sorts of empirical commitments. Virtue ethics assumes the existence of character traits, consequentialists need to know what consequences actions have on people's welfare before they can pronounce on their normative status, and liberals of all varieties need to know what kinds of social arrangements promote or threaten autonomy to derive concrete prescriptions from their principles. Psychological results are thus relevant to normative ethics in at least two different ways. In neither case, however, is there a danger of confusing the philosophical and the empirical aspect of the enterprise.[2]

In metaethics, things are otherwise. Many of the questions it asks look on the surface very much like factual, empirical questions: What is the nature of moral reasoning? Do moral utterances express emotions? Can beliefs motivate us? What kind of dispositions are

[2] For a more detailed discussion, see section 4 of this introduction.

virtues and vices? And so on.[3] So what is the difference between the philosophical and the non-philosophical in this area? Even those who have defended the autonomy of normative ethics with respect to psychology and cognitive science have sometimes been willing to cede the fight when it comes to metaethics.[4] I will begin with a simple sort of response that must soon be qualified: as philosophers we are interested in what is *necessary* and what is *possible*, not in what is *actual*. In metaethics we want to find out what kind of psychological structures *must* be in place for an agent to count as making a moral judgment or as fully morally responsible, or what the structure of a psychological process *must* be like for it to count as moral reasoning. This is an *a priori* investigation into the truth conditions of the relevant claims or, in other words, our concepts and the practices in which they are embedded. It usually proceeds by reflecting on intuitive judgments about particular cases and drawing general conclusions on that basis. When a philosopher claims that to think it is wrong to lie is to be in a state of accepting a norm that prescribes guilt for lying, he is not making an empirical psychological generalization on the basis of observing people who think it is wrong to lie. Rather, he is saying that someone who is not in such a psychological state does not really think that lying is wrong (whether or not she claims to think so); she does not fulfil the criteria for making that type of moral judgment. Thus, the philosopher provides a target, as it were, for empirical, *a posteriori* psychological investigation: if you want to find out what makes people think lying is wrong, for example, look at what makes people accept norms that prescribe guilt for lying. This is the sort of division of labour that would be acceptable to those subscribe to the view expressed in *Tractatus* 4.1121, quoted above.

---

[3] For a clear example of someone who takes just this sort of questions to be empirical, see Johnson 1996, 50.

[4] See Held 1996, who is content to distinguish between causal explanations of moral judgments and normative questions about their correctness. This leaves no room for metaethics, which is concerned with neither.

However, as I noted, distinguishing the philosophical from the empirical in terms of the distinction between the a priori and the a posteriori is too blunt. Things are not quite so simple. First of all, the status of a priori knowledge is a very contentious matter these days. Quine's rejection of the distinction between analytic and synthetic truths cast doubt on one natural grounding for a priori knowledge, and Kripke's arguments for the existence of necessary a posteriori truths convinced many that conceptual analysis has at best a very limited scope in metaphysics.[5] There exists a vast literature on the topic pro and con, and I cannot review it here.[6] As Essay 1 shows, I do not despair of the possibility of something like a priori conceptual analysis even post-Quine and post-Kripke, so I do not find this worry from *methodological naturalism* compelling. It is not my concern to argue that there could not be, for example, *a posteriori* necessary knowledge in philosophy – if we can discover the essence of water by empirical research, we could surely, in principle, discover the essence of some philosophically relevant kind the same way. All I want and need to defend is *methodological pluralism* – as long as *some* of the important truths about necessities and possibilities are accessible to *a priori* conceptual investigation, there is room for traditional philosophical reflection.[7]

Nevertheless, second, *a posteriori* considerations can have a legitimate role in philosophical investigation when *revision* is warranted. One desideratum in a philosophical account is making sense of our ordinary practices, for which the distinction between moral responsibility and the lack of it, for example, seems to be fundamental – we take it that some people deserve praise, others

---

[5] Quine 1951, 1960; Kripke 1980. It should be noted that though the semantic issue of analyticity and the epistemological issue of apriority are connected, they are not identical. For one thing, there may be synthetic a priori truths, like, perhaps, the supervenience of moral properties on non-moral properties (defended in Zangwill 1995).

[6] For some more detail, see the summary of Essay 1 below.

[7] The worry with pluralism is that different methods can lead to different conclusions on the same issue. I believe that these priority issues must be settled case by case.

blame. But what if it turns out that our ordinary concept of moral responsibility requires the sort of capacities for reflective self-control, for example, that empirical studies show human beings just do not have?[8] Well, we could say that we should be sceptics about moral responsibility. But we could also make a revisionist move. Perhaps we cannot be responsible in quite the sense way we thought we were, but there is still a distinction to be made that makes sense of most of our pre-theoretical judgments in the area; some people *do*, as a matter of fact, have the capacities it takes to be *schresponsible* and others do not, and it is the schresponsible ones we think deserve praise and blame. If this were the case, the conclusion to draw could plausibly be revising our ordinary concept of responsibility in the light of *a posteriori* considerations.

Third, there is a class of what we could term *Moorean facts about human nature*. Like the existence of G. E. Moore's two hands, Moorean facts are contingent states of affairs of whose existence we are more certain than of any countervailing evidence.[9] I would argue that there are such things about human nature as well. To begin with, I take it that that it is *a priori* that agents (beings who act and do not merely react) somehow represent goals, have some way of ranking and selecting among them, and somehow represent their environment and the consequences of taking various means to their goals. Now, here is a Moorean fact: at least adult, healthy human beings *are* agents. It is *a posteriori* and contingent, but no conceivable evidence from psychology, cognitive science, or biology could

[8] This worry is nicely brought out by the recent work of Eddy Nahmias on 'neurotic compatibilism' (Nahmias forthcoming). It is very different from the sort of problem that Galen Strawson (1994, 2002) claims we have, namely that our concept of moral responsibility is simply incoherent, so that nobody could be free or ultimately morally responsible, whether determinism is true or not; if Strawson were right, empirical facts would not matter at all. See section 3.2 for more discussion.

[9] Moore's argument is in Moore 1939. The term 'Moorean fact' was introduced by David Lewis, according whom they are "those things we know better than any philosophical argument to the contrary" (Lewis 1999, 418). I see no reason to limit the counterarguments to philosophical ones.

convince us otherwise. A further Moorean fact about human nature: human beings have various emotions. They also daydream, play games, desire respect from others, make love and war. If someone claimed that they did not, we would be entitled to respond with an incredulous stare. So, there is a class of basic truths about human psychology that are accessible without scientific study and that scientific study could not overturn. Philosophers, too, are entitled to appeal to these truths, and have certainly not shied away from doing so. However, this should not be taken as a license for unchecked armchair speculation about human nature, which, unfortunately, has also been a favourite pastime for many philosophers – Hobbes, Rousseau, and Nietzsche spring immediately to mind. Caution and judgment are called for. It is not easy to draw the line around acceptable Moorean appeals, but the more specific and contentious the claims become, the more important systematic empirical confirmation becomes. For example, we cannot simply assume that psychological egoism is false (altruism is not a truism!) or that broad character traits like honesty exist – both either are or involve explanatory hypotheses about human behaviour.[10] As a rule of thumb, the fewer empirically unsupported appeals to contingent a posteriori truths a philosophical account makes, the better.

Finally, philosophers do not just analyze, but also systematize, explain and justify. The method of reflective equilibrium calls for balancing judgments about particular cases with general principles. It is surely central to philosophical work in many areas, but cannot be straightforwardly classified as *a priori* or *a posteriori*. Nor is it conceptual analysis. Relatedly, concepts can be vague and their application to novel situations uncertain. Perhaps it is indeterminate whether small children's recognition of non-conventional status of some norm violations pre-theoretically counts as moral judgment – in some ways it does, in some ways it does not.[11] In such a case, we are surely entitled to fix the borders of the concept guided by further

[10] For details, see the discussion of empirical and philosophical work on these issues in section 4.1.

[11] I discuss studies on the development of the moral/conventional distinction below in section 2.3.

theoretical purposes. And finally, it is a desideratum for philosophical accounts that the explanations they offer are *consistent* with the results of natural and social science.[12] For example, other things being equal, an account of how we can come to know moral truths that only appeals to naturalistically acceptable mechanisms that do other explanatory work as well is superior to an account that postulates a capacity not known to existing science. Rejecting *methodological* naturalism in favour of pluralism does not necessarily mean giving up on *substantive* naturalism, the ontological view that there are no supernatural properties.[13]

The picture that emerges is that while there is an important discontinuity between philosophical and scientific questions and methods, a posteriori considerations can be relevant to philosophical inquiry in a number of distinct ways. Philosophers cannot simply dismiss what science says about their area of interest, even if it is unlikely that the empirical data as such will settle philosophical disputes. *Methodologically*, I thus reject both Wittgensteinian exceptionalism and naturalistic assimilationism. Aristotle's view, as usual, seems the most wise. As to the *subject matter*, I take it that the questions of philosophical moral psychology fall under three main categories:

1. What are the necessary psychological conditions for making moral judgments – that is, what is the nature of moral thinking?
2. What are the necessary psychological conditions for being morally responsible and thus fit to be praised and blamed?
3. What are the implications of facts about human psychology for normative ethical theory?

---

[12] One way to cash this out is to say that a *wide* reflective equilibrium is preferable to a *narrow* one. See Rawls 1971 and Daniels 1979 for discussion.

[13] I formulate substantive naturalism in terms of rejection of the supernatural rather than in terms of affirming the existence of only natural properties, since it makes sense to classify non-naturalists in ethics as (potentially) substantive naturalists. But this is a terminological issue.

For reasons discussed in the next section, the first category has traditionally been central to metaethics, so I will call it *narrowly metaethical*. Questions in the second category can be termed *broadly metaethical*, since they concern the nature of key elements of our ethical practices, but are not themselves normative questions, at least not directly. The final category obviously falls under *normative ethics*, though the psychological claims involved are either simply empirical or belong to either of the two metaethical categories. As it turns out, there are a variety of links between these categories. It is very plausible that being morally responsible requires being able to make moral judgments, so the metaethical categories are, to a degree, interdependent. Second, the capacities required for moral responsibility are plausibly also necessary if not sufficient for autonomy. This means that they have implications for most normative ethical and political theories, and thus for questions of the third category.

In the rest of this introduction, I will present some of the background of the essays comprising the dissertation and summarize their main arguments. Together, they touch on all of the central issues in philosophical moral psychology. Simplifying things somewhat (since none of the papers is limited to discussing questions of a single category), essay 1 discusses the distinctive a priori methodology of philosophical moral psychology, essay 2 narrowly metaethical questions, essay 3 broadly metaethical issues, and finally essay 4 the normative implications of the metaethically relevant psychological facts. Though I do not claim to offer a comprehensive theory of philosophical moral psychology in this collection of articles, I hope that their joint effect is to demonstrate that the recent explosion of interest in the field is not altogether unjustified and that philosophy still has something distinctive to contribute.

The traditional wisdom has it that philosophers are cheap: they do not need a laboratory, statistics programs, or a microscope, just a laptop and a subscription to JSTOR. This, of course, is because the knowledge they seek has been taken to be accessible *a priori*, in some sense independent of experience. An important part of this knowledge is conceptual in nature. (Hardly anybody would claim that all of it is; there seem to be (in Kantian terms) synthetic *a priori* truths, like the knowledge that nothing can be both blue and yellow all over.) Conceptual or analytic truths are such, it used to be said, by virtue of the meanings of the words involved. Some, like "Vixens are female foxes", wear their status on their sleeve. Others, like the (alleged) conceptual truth that someone who could not have done otherwise that she in fact did, did not act freely, are unobvious and may not be recognized as such. However, the traditional view has it, if there are conceptual truths, they are knowable *a priori*, with no need for experience beyond what is needed to grasp the concepts involved. One need not go out and catch a lot of vixens and run them by a vixen-sexer to discover that they are all female. Nor does one need to go out and observe (*per impossibile*) that none of the people who could not have acted otherwise acted freely. Sufficient justification for the beliefs, if it exists, is available to reflection of competent concept-users, and unavailable on the basis of experience.

Why do philosophers care about conceptual truths? A simple answer is suggested by the discussion in the previous section: because philosophers want to get at the essence of things, at what is necessary and what possible, and conceptual truths seem to offer at least part of the answer to such questions. It is not possible for a vixen not to be a female fox; in every possible world, if there are vixens, there are female foxes. (Though of course they may not be called 'vixen', and other things may.) To vary the example, if moral judgment internalism is true, it is not possible for a person to make a genuine first-personal moral judgment without being motivated to some degree to act accordingly. What happens in these cases is that we make the move, in Carnapian terms, from the formal mode to the material mode – from the observation that our concept of moral

judgment would not apply to a certain sort of psychological state to the conclusion that the state in question is not a moral judgment, or from the truth conditions of attributions of moral judgment to the shape of the fact that constitutes it. If internalism is true, motivation is part of the essence of moral judgment.

So far, so good, but here problems arise. What if a vixen does not feel at home in the body of a foxy temptress and undergoes what we call a sex change operation? Is it still female? Is it still a vixen? And most importantly, do our answers to these two questions necessarily go together? Quineans think they do not.[14] We could decide to call something a vixen even if we granted it was now a male. For Quine (1951), all truths are open to revision in light of new *a posteriori* beliefs, famously even mathematical and logical truths. There are no conceptual truths. He is no friend of essences. Kripke (1980), by contrast, is. His challenge to traditional conceptual analysis is drawing apart three distinctions that the logical positivists assumed to coincide, the analytic/synthetic, a priori/a posteriori, and necessary/contingent distinctions. Kripke is particularly concerned to show that there can be a posteriori necessary identities, such as the identity of Hesperus with Phosphorus or water with H_20. His work has inspired most contemporary metaphysicians to talk about *de re* necessities and forget about conceptual truths (though Kripke himself does not deny their existence). In contrast to Quine's heirs, however, the Kripkeans feel free to engage in *a priori* speculation about the nature of the world.

Fortunately, I do not have to take a stand on this debate in 'The Rise and Fall of Experimental Philosophy'. It turns out that at some point in their argument, both defenders and critics of conceptual analysis appeal to conceptual intuitions about various scenarios – would we say that some animal is a vixen, or that some sample of liquid is water? These intuitions are taken to be shared with other speakers and thinkers who have the concept in question. Within the practice of conceptual analysis, too, intuitions serve as evidence one

[14] But let us not forget the well-known Yiddish proverb: Az di bobe volt gehat beytsim volt zi geven mayn zeyde (if my grandmother had balls, she would be my grandfather).

way or another. Does moral responsibility require alternative possibilities? Well, if we can construct a scenario in which we would happily describe someone as morally responsible in spite of lacking alternative possibilities – this is what Harry Frankfurt (1969) tries to do – then it does not. If it turns out that the original scenario fails to rule out alternative possibilities (as some have claimed in Frankfurt's case), it does not support compatibilism after all. This is how much debate within analytic philosophy is still conducted. But what kind of claim is it that an agent in a scenario is *intuitively* responsible? It seems to be a claim about ordinary people's judgments concerning the case in question, a claim about how they would classify the case, or simply what they would say.[15] But what is the source of entitlement for claims of this kind? That depends on how exactly we construe the claim. If it is an empirical hypothesis about the linguistic behaviour of the majority of speakers, we need *a posteriori* empirical evidence to decide on it. If it is a claim about competent users of a shared concept would say in suitable conditions, *a priori* entitlement comes for free with conceptual competence and being in suitable conditions.

The cornerstone of a new school of philosophical methodology commonly called experimental philosophy is that claims about intuitions are empirical hypotheses about observable linguistic behaviour. Consequently, experimentalists have conducted a host of surveys measuring people's responses to carefully constructed scenarios. They have discovered, for example, that most people say that a person can be morally responsible for robbing a bank even if the world is deterministic, and that most people say a psychopath can think that something is morally wrong and yet feel no compunction for doing it. In 'The Rise and Fall of Experimental Philosophy', I challenge this understanding of philosophical appeals

[15] The term 'intuition' has very many uses in philosophy, even in this area. George Bealer's view is that intuitions are "*sui generis* propositional attitudes," fallible intellectual seemings that serve as the source of *a priori* knowledge (e.g. Bealer 2003, 73–75). Bealer is clearly coming from the philosophy of logic and mathematics, where appeals to intuition may well play a different role.

to intuition. I argue that they are claims about what the rules constituting our public concepts require us to say in particular cases, which is to say that they are claims about what competent speakers in sufficiently ideal conditions would say if they ignored pragmatic considerations. I argue, further, that surveys do not and could not provide the sort of data that would settle the truth of this sort of claims. They cannot rule out insufficient grasp of the concept in question (which is all the more likely since the scenarios presented tend to be non-paradigmatic cases), mistakes in application due to inattention or emotional factors, or the influence of pragmatic considerations, such as wanting to avoid undesirable implicatures. This is why they can only get at what I call 'surface intuitions', which are not legitimate evidence in philosophical debates.

Positively, I argue that there are two sources for knowledge about shared concepts. One is simply reflection. What one would say oneself can be a guide to what other users of the same concept would say, since we all have a history of interactions with other speakers. In those interactions, there is pressure for uniformity, since otherwise we would be speaking past each other all the time. There are also sanctions, which may consist in nothing more than misunderstanding. To be sure, reflection is not as easy as it looks, and conditions may be less than ideal for the very reason that one typically has an interest one way or the other. That is why there is a role for the second source of knowledge about shared concepts, good old-fashioned Socratic dialogue. This is a type of dialogue that aims at creating suitable conditions for the responder's judgments to match her own rules. In dialogue, one can vary the scenario in question, compare and contrast it to others, and so draw the attention of the respondent to the presumably relevant features. Thus, someone who is first inclined to say that the psychopath can make genuine moral judgments may change her mind when she is brought to consider everyday cases in which lack of motivation and guilt defeats the attribution of moral judgment. Of course, one's own initial sense of things may be wrong; if the respondent in a Socratic dialogue persists in a judgment that is contrary to one's theoretical commitments, and especially if many do so, one should conclude that one's own intuitions may be corrupted. For example, the folk

concept of moral judgment may indeed turn out to be externalist. An internalist about moral judgment, then, would be proposing a revisionist account on the basis of some other philosophical virtues, and in effect saying that we should change our practice. After all, there is much more to philosophy than conceptual analysis.[16]

[16] Along with my paper, *Philosophical Explorations* will publish two responses. In 'The Past and Future of Experimental Philosophy', Eddy Nahmias and Thomas Nadelhoffer accept that surface intuitions do not suffice, but defend the possibility of getting at robust intuitions by surveys that are better designed. In 'Experimental Philosophy and Philosophical Significance', Joshua Knobe seems to make a U-turn and reject the importance of conceptual analysis to philosophy. The sort of experimental philosophy he now defends uses 'intuitions' to discover how the mind actually works rather than what our concepts are.

2 The Role of Moral Psychology in Metaethics

Metaethics, narrowly conceived, is the study of the nature and presuppositions of ethical thought and its linguistic expressions. As it is often put, metaethics asks second-order questions about ethics, not first-order ones. It does not ask whether, for example, bombing civilians is morally wrong but whether it can be objectively true that bombing civilians is morally wrong and what kind of facts, if any, would make it the case that it is so (*moral metaphysics*), whether the linguistic expressions of the moral judgment in question are in the business of stating facts or (perhaps in addition) conveying attitudes about bombing civilians (*moral semantics*), how is it that we come to know that bombing civilians is morally wrong (*moral epistemology*), and finally, what it is to think that bombing civilians is morally wrong and what is distinctive of the psychological processes that lead to such thoughts (*moral psychology*).

Answers to these questions are obviously not independent of each other. If, as non-cognitivists in moral psychology say, to think that bombing civilians is wrong is to have some kind of negative attitude toward it, then it is natural to suppose that linguistic expressions of moral judgments convey this attitude, and there is little point in looking for facts that would make the judgment true. This makes non-cognitivism metaphysically and epistemologically very undemanding, but raises well-known questions about the apparent objectivity of moral judgments and the apparent logical relationships between moral judgments, for example. If, on the other hand, cognitivists are right and thinking that bombing civilians is wrong is having a belief about its properties, we can straightforwardly ask when such beliefs are justified and what, if anything, makes them true. This puts moral judgments and

discourse on par with other domains in which questions have correct answers, but raises notorious problems about either explaining how *sui generis* moral properties fit in a physical world or how moral properties can be identical with natural ones, as well as issues about the apparent motivational and emotional importance of moral judgments. The history of metaethics is the story of a balancing act of trying to fit all the apparent features of moral thought and reality in one coherent account starting either from cognitive or non-cognitive side.

2.1 Why Moral Psychology Matters to Metaethics

Both cognitivism and non-cognitivism are in the first instance doctrines in philosophical moral psychology, though the terms 'cognitivism' and 'non-cognitivism' are all too often applied to views in moral metaphysics as if they were interchangeable with 'realism' and 'irrealism' (or 'anti-realism'). (This is highly misleading, not least because there are forms of irrealist cognitivism, namely error theory and many varieties of fictionalism.) Why have these moral psychological terms come to designate the main options in metaethics as a whole? It seems that while metaethical problems in different areas can be approached piecemeal, there is a natural order of dependence between them. Insofar as moral semantics studies what moral utterances convey and how, it seems obvious that an answer to this question depends on the answer that we give to the question of what moral judgments consist in – that is, what it is to think something is right or wrong – since it is those very judgments that sincere moral utterances give expression to. In other words, the following Expressive Identity Thesis (EIT) holds:

> (EIT) What moral judgments consist in = what moral utterances express

As it is sometimes put, moral judgments provide the *sincerity conditions* for moral utterances – an utterance of 'Bombing civilians is

wrong' is sincere only if the speaker really thinks bombing civilians is wrong, whatever having that thought consists in. (It may be the case that moral utterances convey more about the speaker's psychological states than just their sincerity conditions, perhaps by way of pragmatic implicatures. The study of moral language, therefore, does not reduce to moral psychology, though the latter is essential to it.) Further, insofar as moral metaphysics studies what ontological commitments moral utterances involve and whether the world meets them or not, the answers it gives seem to depend in part (but essentially) on answers given by moral semantics. So, the following Metaphysical Identity Thesis (MIT) holds:

> (MIT) What moral utterances commit us to = what kind of facts (if any) would make moral utterances true

Given EIT, what moral utterances commit us to is inherited from moral judgments, so given MIT, psychological study of moral judgment is central to the answers of moral metaphysics. For example, if to think that bombing civilians is wrong is to ascribe a non-natural property to bombing civilians, what would (or does) make bombing civilians wrong would be its having such a property. (Of course, there is the further metaphysical question of whether the sort of facts moral judgments ascribe exist and what their nature is. While moral psychology is essential to moral metaphysics, the latter does not by any means reduce to moral psychology.[17]) Similarly, moral epistemology could hardly get off the ground before we have an understanding of whether there is moral knowledge in the first place and what it consists in. Thus, epistemology again points back to questions about the nature of moral judgment. In this way, moral psychology has a certain limited explanatory priority in metaethics, and it often makes sense to classify comprehensive metaethical positions as cognitivist or non-cognitivist.

---

[17] Error theory makes this point vivid: according to Mackie (1977), for example, non-naturalist cognitivism is the correct view in moral psychology, but there simply are no facts that would make the moral beliefs true.

However, it is still possible that *methodologically*, some other branch of metaethics provides a more fruitful entry point to the interlinked questions. Traditionally, and for a good reason, it is moral semantics that has enjoyed this sort of methodological priority. The question about what moral utterances express is, arguably, a question about public linguistic norms that are accessible to philosophical reflection and so offer the prospect of intersubjective agreement. In such reflection, we can exploit EIT in the other direction: since moral judgments are what moral utterances express, knowing what moral utterances express is knowing what moral judgments are. This is why knowing whether moral utterances are disguised imperatives, for example, has direct implications for moral judgment, and someone like Hare can be unhesitatingly classified as a non-cognitivist, in spite his focus on moral language rather than moral psychology. However, we can also ask more directly what counts as taking a moral stance while still exploiting the publicly available character of conceptual norms by asking what makes *attributions* of moral judgments true. That is, we can inquire into the truth conditions of the following sorts of claims:

Jordan thinks that she morally ought to go home.
Paul thinks giving money to charity is morally good.
James thinks Paul is generous.
Michael thinks it would be dishonest to take the money.
Anne thinks she owes gratitude to Michael.

In each case, what makes the attribution true is something about the psychological condition of the person involved – his or her beliefs, desires, intentions, emotions, and so on.[18] For a long time, metaethicists have focused on the first two kinds of judgments, assuming that other sorts of moral judgments can be reduced to some combination of them and factual judgments. This non-accidentally parallels the focus in normative ethics on duties and obligations, right and wrong, or good and bad. As normative

[18] This is, perhaps, not a truth universally acknowledged; see Knobe and Roedder 2006 and my response in Kauppinen 2006.

ethicists have come to have a richer picture of the ethical landscape, emphasizing the notions of virtue and vice and the plurality of deontic concepts, metaethicists are slowly beginning to follow.[19]

2.2 Perspectives on Practical Reasoning and Moral Motivation

Any reflection on moral phenomenology reveals that moral considerations strike us – most of us, anyway – as having a particular kind of *authority* over us in deliberation. We experience them as demanding or compelling, as independent of our will or whims. This phenomenology provides one entry point into further questions in metaethical moral psychology: What kind of motivational and cognitive structures must be in place for this kind of experience to be possible? Do moral considerations really have the sort of authority we experience them to have, or is the experience in that respect an illusion? If they do, what is the source of the authority? These are questions about the appearance and reality of the demands of morality, and virtually every classic of Western philosophy has tried to answer them. As I will try to show, this inquiry, though it draws on some Moorean facts about human nature, has certain distinctively philosophical characteristics. First, it is concerned with possibility and necessity, not just actuality. Second, and most importantly, it always has in view the veracity of the phenomenology, so to speak: the explanation of moral motivation will be either *vindicating*, if it is compatible with the felt authority being warranted, or *undermining*, if the psychological

[19] I defend a cognitivist view of moral judgments involving 'thick' concepts like generosity and gracefulness in my 'Kind Words and Cruel Facts' (in preparation).

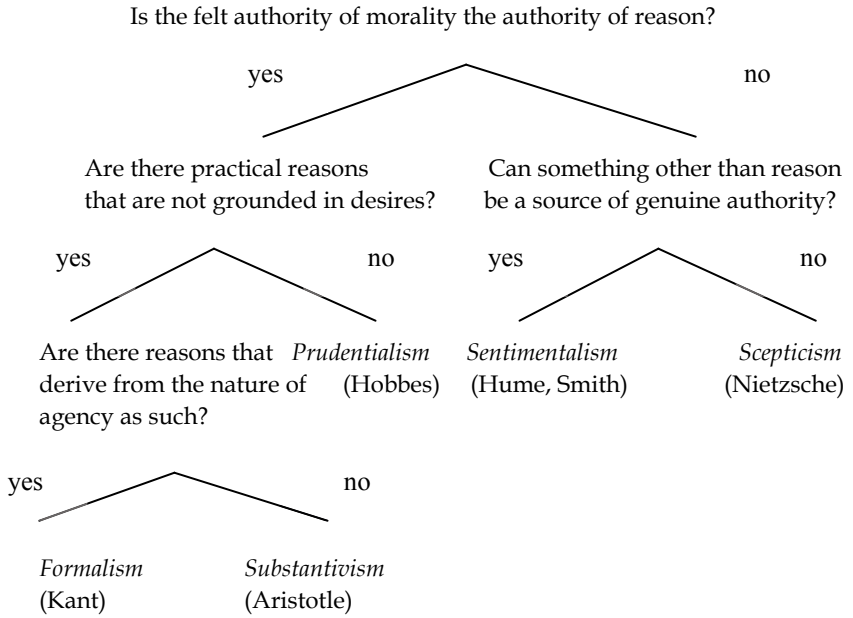
mechanisms it appeals to make it implausible that moral considerations have the sort of standing we take them to have.

To chart the available options in this area, I will organize them in a tree structure. The basic division I make is between those who argue that the authority of morality is also the authority of reason and those who deny that the demands of morality are rationally compelling. The most radical form of this denial is skepticism about the authority of morality. According to these *skeptics*, the explanation of how we come to have the sense of moral obligation undermines it. Thus, on most readings, Nietzsche and Freud each offer debunking accounts of how we come to feel that we ought to do something – in broadest terms, we come to punish ourselves for the things that others punish us for. But denial of the rational authority of morality need not lead to skepticism, as *sentimentalists* like Hume and Smith show. For them, the fact that the felt authority of morality has its source in the social sentiments of the human animal does not mean it is any less normatively compelling, given the sentimentalist conception of normativity. It would be begging the question against sentimentalists to insist that vindicating the demands of morality requires showing they are rational or at least backed up with reasons.

On the rationalist side, the first division is between those who share Hume's skepticism about the power of reason to evaluate final ends and those who have confidence in the practicality of reason. *Prudentialists* like Hobbes take it as given that the pursuit of self-interest is rational and try to show that moral behaviour is in everyone's enlightened self-interest. If that were the case, the authority of moral demands would be underwritten by the authority of our own future good. This story is only partially rationalist, however, since the ends to which morality is a means remain beyond rational assessment. There are two ways in which practical reason could extend beyond an instrumental role. *Formalists* like Kant argue that the very nature of rational agency is a source of normative demands for everyone regardless of their existing desires. The faculty of reason in its practical use allows us to recognize these demands and can give itself rise to motivation. *Substantivists* like Aristotle refrain from making the assumption that all rational agents

would find moral demands compelling. But they do believe that there are reasons for doing the right thing and that being a virtuous agent is a matter of being able to recognize these reasons and being appropriately motivated by this recognition.

The philosophical options in making sense of the appearance and reality of moral demands can be summed up in the following tree:



Choosing between these positions requires answering a number of questions that have been at the center of recent metaethical debates: What are the roles of belief, desire, and emotion in motivation? Is the source of desires expected pleasure, expected good, or something else? Must normative reasons be grounded in existing motives? What makes a psychological transition an instance of practical reasoning? Can practical reason be the source of moral motivation?

What role do principles play in moral deliberation? What is virtue? All of these can be seen as questions about moral judgment, provided that judgment is understood in its process sense rather than the product sense, as an activity rather than a state. A full theory of moral judgment would integrate theories of the process and the product of judging into a coherent whole.

In this section, I will give an overview of how different traditions in philosophical moral psychology answer questions about the process of moral judgment. However, I will discuss sceptical views only to the extent that some empirical theories count as such. A thorough discussion of the views of Nietzsche and Freud would lead too far away from mainstream debates. Since the purpose of this introduction is simply to provide a broader background for the articles comprising the dissertation, I will not try to adjudicate between the competing accounts either.

2.2.1 The Sentimentalist Tradition

The distinctive feature of sentimentalism is that the authority of morality is grounded in sentiments rather than reason. It thus combines pessimism about the powers of practical reason with optimism about human nature. I will begin with the conception of practical reason that the sentimentalist and prudentialist traditions share in its essentials.

Reasoning in general is a process of arriving at new psychological states (or retaining old ones) by way of drawing out the implications of old ones. It is thus something one *does*, consciously or unconsciously. For a psychological process to count as reasoning, the transitions from old to new states must somehow be underwritten, and perhaps also guided, by broadly speaking logical relations among their *contents*.[20] In the simplest case, a person who begins with the beliefs that it will rain tomorrow if the barometer

[20] This sketch draws on the work of Gilbert Harman (1999a) and John Broome (1999). See also Wallace 2003.

falls below 980 and that the barometer reads 979, and infers that it will rain tomorrow counts as engaging in (good) reasoning, since the content of the new belief deductively follows from the contents of the old ones. The *causal* transition between psychological states mirrors the abstract *logical* relationship among the contents, insofar as the agent is rational. This simple case is surely an exception: much reasoning is based on relations that are looser than the deductive, and all too often people engage in *bad* reasoning that only remotely resembles drawing out logical consequences.

Practical reasoning differs from theoretical reasoning in at least two main respects: its subject matter is not establishing how things are but what one is to do, in virtue of which is it essentially first-personal, and its conclusion is correspondingly a psychological state with practical relevance, either a belief about what one ought to do or an intention to act. The simplest kind of practical reasoning is *instrumental reasoning*, which is (roughly) reasoning from intentions (or desires) and beliefs about the most efficient means to new intentions to take the most efficient means. On the *Humean* picture of practical reason, this is the only sort of practical psychological transition that merits the title of reasoning.[21] Relying on what is called the 'Humean Theory of Motivation'[22], according to which only psychological states with a world-to-mind direction of fit, paradigmatically desires, can motivate us to act, Humeans argue that if practical reasoning is to lead to action, its *conclusion* must be a desire-like state. Relying on what has been termed the 'desire-in, desire-out principle'[23], according to which practical reasoning can

[21] To talk about a 'Humean' view is not necessarily to talk about Hume's own view, and the same goes for other classics. The relationship between Hume and Humeans is some sort of family resemblance, the most distinctive feature of which is Hume's assertion that "Reason is, and ought only to be the slave of the passions." (Hume 1739-1740/1978, 415).

[22] For a well-known defense of the view, see Smith 1987. For the notion of direction of fit, see also Anscombe 1958a, Searle 1983. Smith's version of the Humean theory is criticized and rejected by Schueler 1995, Dancy 2000, and Tenenbaum (forthcoming).

[23] Wallace 1990, 370.

give rise to new motivation only if there has been some antecedent "motivation for the agent to deliberate *from*", as Bernard Williams puts it[24], Humeans argue also that the *starting point* of practical reasoning must ultimately be a desire that is itself beyond rational assessment. Consequently, desires can be rationally criticized only insofar as they are based on false factual beliefs or conflict with other, more fundamental non-rational desires.[25] It is a short step from here to the familiar *maximizing conception of rationality*, according to which what it is rational for us to do is what would maximize our expected utility, where utility is understood as satisfaction of existing desires. If what we have *reason* to do is what we would do if we were fully rational, this has implications for what reasons we have as well. Based on the connection between rationality and reasons, Williams has argued that what an agent has reason to do is constrained by her existing motivations, however those have come about. In his terms, we can have only *internal reasons*, considerations that would motivate us after sound practical deliberation proceeding from our existing motivational set.[26] Negatively, the claim is that there are no *external reasons*,

---

[24] Williams 1981b, 109. It should be noted that Williams himself goes beyond the narrowly Humean picture by including among an agent's "motivational set" not just desires but also "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be called, embodying commitments of the agent" (ibid., 105), and among ways of practical deliberation not just means-end reasoning but also finding constitutive means, harmonizing and ranking ends, and imagining what the realization of ends comes to (ibid., 104). Compare Hume, *Treatise* 3.1.1, 296–298.

[25] What 'more fundamental' means in this context is not a simple question. It is cannot be just 'causally stronger', since that would leave no room for normativity – the causally strongest desire is by definition the one I actually act on. Several alternatives are open: perhaps fundamental desires are those on which others depend for their point (for example, desire to borrow a book probably lacks a point in the absence of a desire to read the book) or those that feature centrally in the agent's self-conception (for example, Ronald Reagan's desire that Communism fail).

[26] Williams 1981b, Williams 1995, 39. See also Smith 1995.

considerations that would be normative for an agent regardless of her existing motivational states. It would be "bluff" or "bullying" someone to insist that they have a reason to do something that does not serve their desires or projects.[27]

The Humean view has clear implications for the relationship between morality and rationality. If our reasons depend on our contingent desires, it may well be the case that a dictator has no reason to refrain from torturing his opponents. If, as seems plausible, what we ought to do is what we have most reason to do, it may thus be the case that the dictator positively *ought* to torture his opponents.[28] Morality and reason come apart. This raises a worry about the authority of morality. There are two basic reactions to this threatening fact within the instrumental or maximizing conception of rationality: denying it or giving an alternative explanation for why morality is important. Hobbes and his followers like Gauthier take the first route and try to show that at the end of the day, it pays off to be moral, so it is rational in this sense after all. Since the pursuit of one's enlightened self-interest does not look very much like moral behaviour even on those occasions in which the two coincide, I will

[27] Williams 1981b, 111, Williams 1995.

[28] This conclusion is embraced by Gilbert Harman, according to whom it is not the case that Hitler ought not have ordered the extermination of Jews, given his values, even if we can say, deploying our own values, that Hitler was evil (Harman and Thomson 1996, 60–62; Harman 1975). Other Humeans like Simon Blackburn avoid this conclusion by rejecting the connection between reasons and oughts, on the one hand, and rationality, on the other; consequently, they can say that though it was not *irrational* for Hitler to order the genocide, he had no *reason* to do so and *ought not* to have done so (Blackburn 1998, ch. 9).

leave this *prudentialist* option aside here.[29] At best, it vindicates the authority of morality in an ersatz sense.[30]

The second alternative is more promising. Hume and Adam Smith begin with a strict division of labour between reason and sentiment: the former tells us what the probable consequences of our actions are and the latter independently makes us choose those that benefit general happiness. The central question thus becomes how we can and do come to have a desire, or more precisely a sentiment of approbation toward actions that benefit others, even when it runs contrary to our own perceived interests. On their story, we start our approving actions that give us pleasure and disapproving those that cause us pain. However, it is a natural fact about us that we *sympathize* with the pleasure and pain of others as well, and thus to an extent share them. (It is important that this is not a prescription but a descriptive claim: we *cannot help* sympathizing with others.[31]) This kind of first-order sympathy already disposes us to disapprove of actions that cause pain to others – unless we take them to deserve it, or, perhaps, if our own interests are at stake. For such cases, at least, a further source of motivation is needed. Here sympathy enters the story a second time: we also come to feel the approval and disapproval of others toward our own actions. This process of *internalizing* the attitudes of others is the source of the sentiments of shame and guilt, which can be motivationally very effective.[32] Of

---

[29] As Kant points out, "[t]he maxim of self-love (prudence) only *advises*; the law of morality *commands*" (Kant 1788/1996, 169). In his example, if you win a game by cheating, prudence congratulates you, while the moral law tells you to despise yourself. Nor do we think people deserve to be punished if they act against their self-interest, unlike in the case of moral violations.

[30] In fairness to Hobbes in particular, he is not concerned with moral obligation or our sense of it, but with the justification of *political* obligation.

[31] "These sentiments are so rooted in our constitution and temper, that without entirely confounding the human mind by disease or madness, 'tis impossible to extirpate and destroy them." (*Treatise* 3.1.2, 305)

[32] Both Hume and Smith provide very persuasive examples of the motivational role of shame, in particular. Hume mentions the mortification that a man feels when another complains of his bad breath, though it clearly

course, we are aware of the fact that particular others may approve or disapprove of us without our meriting it, perhaps because their interests conflict with ours. But we can also *imagine* how an *impartial spectator* would react to our action, and thus come to feel that the approval or disapproval is warranted. As Smith puts it,

> we either approve or disapprove of our own conduct, according as we feel that, when we place ourselves in the situation of another man, and view it, as it were, with his eyes and from his station, we either can or cannot entirely enter into and sympathise with the sentiments and motives which influenced it. […] We endeavour to examine our own conduct as we imagine any other fair and impartial spectator would examine it. (Smith 1759/1976, Part III, chap. 1.)

Along the same lines, Hume notes that if we call someone an 'enemy' or a 'rival', it is understood that we are expressing our particular sentiments, but if we call him 'vicious' or 'depraved', we are expressing sentiments with which we expect our audience to concur, sentiments that would be endorsed from a 'common point of view'.[33] When moralizing, we distance ourselves from our immediate sentiments of praise and blame, since we are aware that morally irrelevant facts about our situation, such as mood, spatiotemporal position and self-interest, can distort them, and reflectively correct for such factors to arrive at more impartial judgments. As Hume puts it, "We make allowance for a certain degree of selfishness in men; because we know it to be inseparable from human nature, and inherent in our frame and constitution. By this reflection we correct those sentiments of blame, which so naturally arise upon any opposition."[34] This is how we may even

---

it is as such no inconvenience to himself (*Treatise* 3.3.1) and Smith that of a man who looks around and finds he is the only one laughing at his joke (Smith 1759/1976, 1.1.2).

[33] *Enquiry* IX, 252.

[34] *Treatise* 3.3.1, 372. Hume also notes that this impartial correction is an aspiration in which we often fail: "Sympathy, we shall allow, is much fainter than our concern for ourselves, and sympathy with persons remote from us

admire our enemies.[35] This is no different in principle to what we do with our sense perceptions when we correct for the effects of distance or lighting conditions. In both cases, as Geoffrey Sayre-McCord puts the Humean position, we are able to distinguish between reality and appearance by defining "a set of standard conditions occupied by a standard observer and [taking] her reactions (her sense perceptions or sentiments) as setting the standard for ours"[36]. If we did not take those steps, morality could not perform its function of social coordination of attitudes and actions.[37] Specifically moral language would lose its point. Thus, in short, our natural, non-rational dispositions to sympathize, internalize, and imagine together give rise to recognizably moral preferences, which we then tend to project as features of the world itself.

As Simon Blackburn has emphasized, the Humean story is not a sceptical one, even though it denies that the demands of morality are those of rationality and denies the explanatory reality of moral properties. The fact that the felt authority of morality derives from internalizing the disapproval of an impartial spectator does not undermine that authority. It does not make the moral demand

---

much fainter than that with persons near and contiguous; but for this very reason it is necessary for us, in our calm judgments and discourse concerning the characters of men, to neglect all these differences and render our sentiments more public and social." (*Enquiry* V, 220)

[35] *Treatise*, 303.

[36] Sayre-McCord 1994. Sayre-McCord argues that for the purposes of social coordination of moral judgments, we are better off referring to a standard rather than ideal (omniscient, perfectly impartial, limitlessly sympathetic etc.) observer, whose reactions we could not possibly anticipate, given that we as a matter of fact always lack her knowledge and psychological capacities.

[37] In Hume's words, "When we form our judgments of persons, merely from the tendency of their characters to our own benefit, or to that of our friends, we find so many contradictions to our sentiments in society and conversation, that we seek some other standard of merit and demerit, which may not admit of so great variation." (*Treatise* 3.3.1, 373; cf. *Enquiry* V, 219–220) For a contemporary version, see especially Gibbard 1990.

arbitrary or unwarranted – that is the point of idealizing the disapproval. To be sure, it does not vindicate morality by showing that its demands are rational, but the sentimentalists reject such restriction on the notion of vindication. To think that morality is vindicated is itself to have a positive sentiment toward it, and Hume argues at the very end of the *Treatise on Human Nature* that this is precisely what should happen when we come to see that moral thought has its source in our natural social sentiments:

> All lovers of virtue (and such we all are in speculation, however we may degenerate in practice) must certainly be pleas'd to see moral distinctions deriv'd from so noble a source, which gives us a just notion both of the generosity and capacity of human nature. It requires but very little knowledge of human affairs to perceive, that a sense of morals is a principle inherent in the soul, and one of the most powerful that enters into the composition. But this sense must certainly acquire new force, when reflecting on itself, it approves of those principles, from whence it is deriv'd, and finds nothing but what is great and good in its rise and origin. (*Treatise* 3.3.6, 394)

2.2.2 The Kantian Perspective

For the sentimentalists, those of us who treat moral demands as authoritative do so only because of contingent facts about human nature – most of us happen to be attuned to the sentiments of others. Kant found this an intolerably shaky foundation for morality, as well as one that does not sit well with our commonsense conviction that the demands of morality obligate everyone. For him, if the will is determined with respect to a contingent goal of an agent, it always aims at some expected pleasure.[38] To be sure, the source of that

[38] Kant is thus a psychological hedonist when it comes to motives not derived from pure reason. To this extent he is in accord with Hume, according to whom "'Tis from the prospect of pain or pleasure that aversion or propensity arises towards any object" (*Treatise* 2.3.3, 266), and a long line of others. In the *Metaphysics of Morals* he emphasizes that pleasure can be

pleasure may be the well-being of others, but insofar as it is just a pleasure among others, what he will do is unbearably contingent from a moral perspective. If some other pleasure appears greater or easier or longer-lasting, a benefactor "can even repulse a poor man whom at other times it is a joy for him to benefit because he now has only enough money in his pocket to pay for his admission to the theater"[39]. In Kantian terms, the Humean view has room only for *hypothetical* reasons to be moral. Kant himself famously held that as we ordinarily understand it, the demands of morality are both rational and *categorical*, imperative for anyone regardless of their desires. To vindicate the authority of morality, we must be able to show that even Hume's 'sensible knave' who, lacking sympathy or imagination, desires to take advantage of his fellow-men when able to do so with impunity, has reason to honour the moral law.

Establishing that the knave has reason to act morally will naturally take a lot of argument. Perhaps the cleverest arguments against the Humean view aim to show that instrumental reasoning is only normative or reason-providing if there are categorical reasons to have the desires that serve as its premises. Otherwise it is not the case that one *ought* to take the means, or has reason to do so. Thus, John Broome has argued that Humeans commit the *fallacy of detachment* parallel to a similar mistake in modal reasoning: just as it would be a mistake to infer "Necessarily q" from "Necessarily, if p then q" and p, it is a mistake to infer "I ought to φ" from the instrumental principle "I ought to make it the case that if I desire to ψ, I take the best means φ" and desiring to ψ.[40] In the modal case, for it to be necessarily the case that q, it must also be *necessarily*, not

---

associated with desire in two ways, either as cause or effect (Kant 1797/1996, 373–374). In the latter case, the pleasure cannot determine the desire, which must instead be directed by pure practical reason itself. This sort of pleasure is not pathological but moral: "[P]leasure that must be *preceded* by the law in order to be felt is in the *moral order*." (Kant 1797/1996, 511; emphasis in the original)

[39] Kant 1788/1996, 157.

[40] What I present here is a simplified and modified version of the argument in Broome 1999.

merely contingently, the case that p, even if q necessarily follows from p. Similarly, in the case of oughts, for it to be the case that I ought to φ, it must also be the case that I *ought* to desire to ψ, even if rationality requires φ-ing if one desires to ψ.[41] If there is a sound argument of this type, reasons cannot *bottom out* in non-rational desires.[42]

Even if the Humean view of practical reasoning is inadequate, it does not yet follow that a Kantian view must be correct. The distinctively Kantian approach is to argue that categorical reasons derive from the very structure of practical reasoning itself. Like Humeans, Kantians thus understand *reasons* for action in terms of what it would be *rational* for an agent to do. For Kant, practical reasoning is a matter of deciding which *maxim* of action to adopt; a maxim is a subjective principle of action of the type "In circumstances C, I will perform action A in order to achieve end E"[43]. A maxim is like a policy or law that one gives to oneself. Many contemporary Kantians argue that having such policies, as opposed to merely doing what one most desires, is constitutive of being a person as such. The argument is that being a person is not being at the mercy of passing desires (like Harry Frankfurt's 'wantons') but having a vantage point that transcends particular moments.[44] Persons are capable of taking a stand on things on the basis of

---

[41] The fallacy can be formulated in terms of the *scope* of the relevant operator. In the case of necessity, it is a matter of reading □ (p→q) as p→□q, which are clearly not equivalent. Using 'O' for 'ought to', the practical reasoning case can be formulated as mistaking O (ψ→φ) for ψ→Oφ. The latter principle would obviously allow for detaching the consequent.

[42] Christine Korsgaard presents a different argument in the same vein in Korsgaard 1997. It is criticized in Hubin 2001.

[43] It is an indication of the contested nature of Kant interpretation that there is no consensus on such a fundamental issue as what a maxim is for him. The formulation I use captures all the elements in at least one of Kant's own examples, namely "[E] From self-love [A] I make it my principle to shorten my life [C] when its longer duration threatens more troubles that it promises agreeableness." (*Groundwork* II, 74)

[44] See Korsgaard 1996 and Velleman 2006 for this line of argument.

reflection, and taking a stand consists in adopting policies and principles.

Famously, Kant argues that adopting a maxim is rational only if it could at the same time be *willed* by *all* as a *universal law*. What this amounts to and exactly how it is meant to work is the subject of much controversy. A few things should be relatively clear, however. First, this formulation of the Categorical Imperative[45] is meant to be a *formal* criterion, a filter through which any maxim must pass to be rational. The criterion must be formal, because reasons are essentially the sort of considerations that have authority for *any* rational agent. Rationality cannot depend on the *material* of the maxims, contingent desires and goals, since people have, or at any rate could have, all sorts of desires or goals.[46] By the same token, the maxim must be universalizable – if it is binding on me *qua* rational agent, it is binding on everyone *qua* rational agent. As many have recently argued, for a policy to be rational, it must be *justifiable* to any rational agent, suitably qualified – for Habermas, to any agent willing and able to enter into an uncoerced discourse[47], for Scanlon, to any agent who is motivated "to find principles for the general regulation of behavior that others, similarly motivated, could not reasonably reject"[48].

Second, here as elsewhere, the fundamental formal principle of rationality is that of *avoiding contradiction*. Kant offers two different tests for this. The *contradiction in conception* test asks whether the

---

[45] I follow the current convention and capitalize the Categorical Imperative when talking about the test that maxims must pass in order to determine which are "categorical imperatives" in the plural, i.e. prescriptions that obligate unconditionally. I will here discuss only the Formula of Universal Law version and leave aside the issue of whether the other formulations are equivalent, as Kant claims.

[46] "And how should laws of the determination of *our* will be taken as laws of the determination of the will of rational beings as such, and for ours only as rational beings, if they were merely empirical and did not have their origin completely a priori in pure but practical reason?" (*Groundwork* II, 63)

[47] See Habermas 1983, Habermas 1996.

[48] Scanlon 1998, 6.

maxim is even conceivable as a universal law; if not, it will not be justifiable to any rational agent. Kant's most famous and persuasive example of this is the case of deceitful promises. Suppose I am considering the possible maxim "I will make a deceitful promise in order to advance my interests". If it were a universal law, it would be something like "Everyone will make a deceitful promise in order to advance his interests." But if everyone made deceitful promises when it suited them, the whole institution of promising would collapse: no one would take anyone's word for anything.[49] Thus, the maxim cannot be coherently universalized. A maxim whose contrary fails the contradiction in conception test identifies a *perfect duty* or unconditional obligation. Since failing to keep promises fails the test, keeping promises is such a duty – for Kant, it is never acceptable to break a promise. The other test Kant offers for identifying rational maxims is *contradiction in will*. According to him, there are maxims that can be conceived as universal laws but cannot be coherently *willed* as such. One of his examples is failing to develop one's talents. The idea seems to be that given some basic facts about human beings, they cannot achieve any goals without either themselves or others having the skills they require. If nobody had any skills, no goals would be reached. Assuming that as agents we are always seeking one goal or another, we cannot coherently will a world in which no goals are reached.[50] If this sort of admittedly contrived reasoning holds water, we have an *imperfect duty* to develop our talents. Imperfect duties obligate only sometimes and to some extent – we need not spend all our time developing our talents.

Suppose that we do, in fact, get determinate content for the moral law by way of the categorical imperative test, and so as a result of pure reasoning. (Since I am not here concerned with

---

[49] As Kant puts it, "the universality of a law that everyone, when he believes himself to be in need, could promise whatever he pleases with the intention of not keeping it would make the promise and the end one might have in it itself impossible, since no one would believe what was promised him but would laugh at all such expressions as vain pretenses." (*Groundwork* II, 74)

[50] Compare Kant 1797/1996, 522–523.

arguing for any particular position, I will adopt an agnostic stance on whether or not Kant's tests yield determinate results.) Why should anyone care about this – that is, why should we adopt only rationally acceptable policies? Well, if, as it seems, the question about why one should care about something is a question about what *reasons* we have to care about something, it is already answered. In David Velleman's words, there is "something self-defeating about asking for a reason to act for reasons"[51]. Similarly, when we ask if something really has authority, we seem to be asking for reasons to obey its commands, and we already know that reason commands us to obey the commands of reason. So while we *can*, as a matter of fact, act contrary to morality, as we all too often do, its normative authority over us is inescapable and the demands it makes upon us genuine. When we experience something as having this sort of authority, we have the subjective feeling of respect (*Achtung*) for it: "What I cognize immediately as a law for me, I cognize with respect, which signifies merely consciousness of the *subordination* of my will to a law without the mediation of other influences on my sense."[52] Kant hastens to add that respect is not a cause of the moral law (that is, the will is not presented with a hypothetical imperative "If you respect the moral law, do x") but an effect of the law. Kant does not present this as an empirical hypothesis; rather, consciousness of a law one has given oneself is what respect *is*.[53] Thus, insofar as moral

[51] Velleman 2006, 19.

[52] *Groundwork* I, 56n. This kind of respect combines elements of fear and inclination: the law gives rise to something like fear insofar as it may command us to act against our own happiness, and something like inclination, since it is something that we ourselves will as rational agents, not something external.

[53] Thus, to say that one acts out of respect for the law, as Kant sometimes does, is just to say one's action is determined by consciousness of the law, and since the latter results from an exercise of pure reason, reason itself can be practical. I emphasize this, because otherwise there is a temptation to read Kant as going back on his claim that reason itself can be practical when he says things like "neither fear nor inclination, but solely respect for the law, is the incentive which can give an action moral worth" (*Groundwork* II).

demands are reason's own laws, both their experienced and normative authority is explained.

2.2.3 The Aristotelian Alternative

It takes some violence to fit Aristotle into the framework of moral motivation I have been working with. The Greeks, as is well known, were not terribly keen on moral obligation and guilt. Nor were they, like Kant, haunted by the possibility that morality might not be authoritative for everyone. But Aristotle does contrast acting 'for the sake of the fine' with other motives and links it with avoiding shame, and believes that the demands of virtue are those of reason. This is clear both in his general metaethical remarks and the discussion of particular virtues. For example, he says of the brave person that "though he will fear even the sorts of things that are not irresistible, he will stand firm against them, in the right way, *as reason prescribes*, *for the sake of the fine*, for this is the end aimed at by virtue"[54] (all emphases in the paragraph are mine). Standing firm in the face of danger requires that "the brave person's actions and *feelings accord with what something is worth*, and follow what reason prescribes"[55]. Moreover, when a virtuous person does what reason prescribes, "he will do this *with pleasure*, or at any rate without pain; for action in accord with virtue is pleasant or at any rate painless, and least of all

---

Such passages could suggest that an incentive (*Triebfeder*) in addition to reason is needed for moral action in Kant's moral psychology. Along the same lines, Philip Stratton-Lake analyzes respect as "reverential awareness of the moral law" (Stratton-Lake 2000, 36) and defends the view that this sort of awareness *constitutes* being morally motivated rather than usurping the practical role of the moral law itself.

[54] *NE* 1115b, 41. I will be referring to Irwin's translation of *Nicomachean Ethics* using both the Bekker and Irwin page numbers. Since I am not working from a Greek edition, I will not use line numbers.

[55] *NE* 1115b, 41.

is it painful."[56] Practical reason, emotions, desires, and pleasure are thus all involved in the process of moral judgment, but Aristotle's understanding of each is very different from his early modern counterparts. I will focus on the puzzles of his understanding of practical reason and its relationship to virtue and emotion.

Aristotle has a kind of Humean theory of action – for animals. Animals experience pleasure and pain, and expectation of pleasure or pain gives rise to desire, which gives rise to action.[57] But in human beings, desire or wish (*boulesis*), in contrast to mere animal urge, is essentially oriented to what the agent takes to be good. Of course, if the agent thought that pleasure was the only good, this would almost coincide with the Humean account of desire.[58] But since for Aristotle there is more to well-being or *eudaimonia* than pleasure, the virtuous, who have a correct conception of the good, have a different motivational structure.[59] They desire in accordance with reason, and deliberate about the best means (which may be constitutive) to bring about the desired end. The structure of deliberation, for Aristotle, can be represented by means of a practical syllogism. The first or major premise is the 'premise of the good', which is the content of a desiderative (orectic) state, and the second

---

[56] *NE* 1120a, 50. This is in the context of a discussion of generosity.

[57] See Irwin 1980, 42. Desire in animals is "appetite (*epithumia*), nonrational desire for the pleasant" (Irwin 1980, 44).

[58] I say it would 'almost' coincide, since for Aristotle, pleasure is not a single phenomenal state. Instead, there are many kinds of pleasure arising from the exercise of our various capacities, in short, activities. As Aristotle often says, the sort of pleasure that is characteristic of an activity, like the joy of solving a mathematical problem, *completes* it, while an alien pleasure disrupts it. See *NE* 1174a–1176a, 157–161.

[59] It follows from the nature of pleasure (see previous footnote) that it does not make sense to aim at happiness by way of finding the best means for maximizing pleasure. This is just not a meaningful goal, since the 'means' and the 'end' are not independent of each other in the way that such deliberation would require. The best life *is* the most pleasant, too, but that is just a consequence of the fact that it involves the exercise of characteristic human capacities with respect to their proper objects. Compare *NE* 1099a, 11.

or minor premise is the 'premise of the possible', which is the content of a cognitive state, and the conclusion an action or a decision to act. For deliberation to be good, both of these premises must be correct. In good deliberation, "virtue makes the goal correct, and prudence makes the things promoting the goal [correct]."[60] (This statement creates a number of puzzles, which I will return to in a moment.) Decision (*prohairesis*), in turn, is just a "deliberative desire to do an action that is up to us" (*NE* 1113a, 36). This deliberative desire, then is the "principle of the action – the source of the motion" (*NE* 1139a, 87), unless weakness of the will intervenes.

Given that Aristotle sometimes suggests that the scope of prudence or practical wisdom (*phronesis*)[61] is limited to "things open to deliberation"[62], there is some temptation to read him as having a Humean conception of practical reason as an ability to go from one desire to another.[63] But this cannot be true. Aristotle also unambiguously says that prudence is "a state of grasping the truth, involving reason"[64]. Practical wisdom is distinct from mere cleverness, which is "such as to be able to do the actions that tend to promote whatever goal is assumed and to attain them" (*NE* 1144a, 97). The *phronimos*, the practically wise person, has the right aim as well as the right means to it. That is why it cannot be had without the virtues of character.[65] At the same time, full virtue cannot be acquired without *phronesis*; virtues of character do not merely accord with right reason, but involve it.[66] There is thus a puzzle at the heart

---

[60] *NE* 1144a, 97. Gloss by Irwin.

[61] 'Prudence' is the term used in Irwin's standard translation of *NE*. 'Practical wisdom' seems often more appropriate. I will use both terms here. Prudence is the virtue of the part of the rational part of the soul that is concerned with things that can be otherwise.

[62] *NE* 1141b, 91. Cf. also *NE* 1139a, 86, according to which prudence is the virtue of the rationally calculating part of the soul, and "deliberating is the same as rationally calculating".

[63] For this kind of reading, see Audi 1989.

[64] *NE* 1140b, 89, 90.

[65] *NE* 1144b, 98-99; *NE* 1178a, 165.

[66] *NE* 1144b, 98.

of book VI of the *Nicomachean Ethics* and Aristotle's theory of practical reason: how can we reconcile the claim that we do not deliberate about the ends with the claim that virtues, which set the ends, cannot be had without *phronesis*? One way to understand these remarks is to understand prudence functioning in two stages: first specifying what would best promote *eudaimonia* in the present situation (providing the premise of the good), and then deliberating about the best means to that end. This sort of two-part conception of the process seems to accord well with *NE* 1142b: "If, then, having deliberated well is proper to a prudent person, good deliberation will be the type of correctness that accords with what is expedient for promoting the *end* about which *prudence is the true supposition*" (my emphasis). This is the line taken by Terence Irwin, who believes that when we deliberate about what promotes happiness, "we discover its constituents, and so we have a more precise conception of happiness", which can then be "the basis for further deliberation about what to do"[67]. But how can this be reconciled with the claim that "we deliberate about things that promote the end, not about the end" [68]?

Some interpreters believe that the only way to make Aristotle coherent is to take the correct end that prudence requires to be provided by some other capacity. We know that what makes an end correct is that it contributes to *eudaimonia* or happiness or living well, which everyone can agree in the abstract is the one thing that we do not pursue for the sake of anything else and for the sake of which (at least in part) we pursue all other things. But this is not to say anything until we have a concrete view of what eudaimonia consists in.[69] How do we reach that? David Reeve argues that ethical principles, generalizations that hold for the most part about what constitutes *eudaimonia* or promotes it, are not a matter of deliberation, though they are arrived at by a rational process: "We do not deliberate about what *eudaimonia* really is. We discover what it is, as we discover what a crab is, by experience, empirical

[67] Irwin 1999, 249.
[68] *NE* 1113a, 36.
[69] Cf. *NE* 1095a, 3.

investigation, and dialectic."[70] Practical wisdom gets its major premise "second-hand from the scientific part"[71]. How does the scientific part, then, get at the principles of ethics? Reeve's leading idea is that ethics is much more like science for Aristotle than many think. Like science, ethics has first principles that cannot be derived from other principles. Instead, their source is in the first instance *induction* (epagoge) *from experience*.

In ethics, induction works to begin with by way of experiences of pleasure and pain, which inform us whether our existing ends are such that they contribute to a satisfactory life – whether what we happen to desire actually is desirable.[72] We learn, for example, which foods it is good to eat and how much. This process does not begin with a blank slate, since we start out with *natural tendencies* that are roughly in the right direction – children do not find crude oil appetizing, for example.[73] Nor does it take place in a social vacuum: we already have generations of 'experiments in living' (Reeve borrows Mill's term) behind us, and so *culturally transmitted knowledge* of what is good for being like us, such as recipes for Thai red curry. (Some people get lucky in this respect and get better brought up than others in the ways of human happiness; more on this below.) These experiences give rise to candidate conceptions, which can then be tested in *dialectic* argumentation against the *endoxa* on the area, the views held by "everyone or by the majority or by the wise, either by all of them or by most or by the most notable and

[70] Reeve 1992, 82. Compare Reeve 2006, 205: "that happiness is our end is not up to us, since, as something determined by our function or essence … it does not admit of being otherwise".

[71] Reeve 2006, 208.

[72] Reeve 2006, 204, 214. He points out that these experiences also inform us of what the best means to satisfy our desires are. Compare *NE* 1172a, 153: "[W]hen we educate children, we steer them by pleasure and pain."

[73] In line with his general teleological worldview, Aristotle talks about "natural virtue" in *NE* VI, 13. As Reeve puts it, "without natural virtue we will not have the kind of experience from which the truth about *eudaimonia* can be reached by induction or habituation." (Reeve 1992, 89)

reputable"[74]. Dialectic in general, not just in ethics but all the sciences, is a matter of trying to solve the *aporiai* or apparent contradictions among the *endoxa* by removing ambiguities, identifying false assumptions, explaining how people could have come to make mistakes, and showing that the principles arrived at can account for the remaining *endoxa*.[75] Aristotle's summarizes some of these features at the beginning of his discussion of weakness of will:

> As in other cases, we must set out the appearances, and first of all go through the puzzles. In this way we must prove the common beliefs about these ways of being affected – ideally, all the common beliefs, but if not all, most of them, and the most important. For if the objections are solved, and the common beliefs are left, it will be an adequate proof. (*NE* 1145b, 100)

Aristotle's argument against the widely held belief (and thus *endoxon*) that happiness is bodily pleasure is an example of this sort of dialectic at play – basically, he argues that people who have limited experience of the good life take one constituent of it, bodily pleasure, for the whole thing.[76] This makes intelligible why the many make a mistake, and allows Aristotle to delete the view with good conscience from the list of those that the best theory must accommodate.

Dialectic, certainly, is a rational process of justification, but it is not practical reasoning – it aims to discover how things really are, to give us theoretical knowledge or *nous* of *eudaimonia*. Of course, not everyone engages in dialectic, but in any case, once we have a conception of what happiness is and what promotes it, whether derived from personal experience, education, or dialectic, in Reeve's picture, "*phronesis* or practical wisdom uses perception to apply a universal … supplied by *nous* to guide a particular action."[77] On this

---

[74] *Topics* I.1.100b21-23, I.11.104b32-34 (quoted in Reeve 2006, 200).

[75] See Kraut 2006; also Reeve 1992, ch. 1.

[76] See *NE* 1153b-1154a.

[77] Reeve 1992, 59.

kind of view, the universal, which functions as the major premise of the practical syllogism, is something like "Giving to the needy is part of the good life", and the minor premise, supplied by a kind of perception, is something like "My signing this check is giving to the needy".[78] It is important that the 'middle term' of the syllogism ("giving to the needy") is itself cast in terms of ethically neutral properties whose instantiation can consequently be grasped by anyone, including those not having had a good ethical upbringing. The ethical weight, so to speak, is then carried by the major premise; it is what one must get right in order to be a virtuous person.[79]

---

[78] The minor premise must involve indexical and demonstrative elements, since it is meant to lead to immediate action or at least decision concerning action. See also Gottlieb 2006.

[79] Aquinas's theory of practical reason seems to work along these lines. He clearly subscribes to the division of labour model: *prudentia* (his Latin version of *phronesis*) "applies universal principles to the particular conclusions of practical matters" (*Summa* II/II/Q47.6). These principles are known to the understanding by a "special natural habit, which we call 'synderesis'." (*Summa* I/Q79.12), a notion related to conscience. There is a twist to the story, however: for Aquinas, understanding is in a sense a *part* of *prudentia*: "Now every deduction of reason proceeds from certain statements which are taken as primary: wherefore every process of reasoning must needs proceed from some understanding. Therefore since prudence is right reason applied to action, the whole process of prudence must needs have its source in understanding. Hence it is that understanding is reckoned a part of prudence." (*Summa* II/II/Q47.7) The sense in which understanding is a part of prudence seems to be that it is one of the abilities needed for practical wisdom. According to Aquinas, *prudentia* has eight such 'quasi-integral' parts: memory (since it is needed to learn from experience), understanding (*nous*, right estimate of a principle, which is needed to get the major premise right), docility or deference (since we must learn from others), shrewdness (since we must be quick in seeing similarities to find the right minor premise), inference (since we move from premises to conclusion), foresight (since we must anticipate consequences), circumspection (since we must fit the means to our circumstances) and caution (since appearances of the good are often deceptive). (*Summa*, II/II/Q.48-49)

There is, however, a different reading of Aristotle that gives the perceptual nature of *phronesis* more emphasis than Reeve allows. It highlights Aristotle's remark that practical wisdom "is about the last thing, an object of perception, not scientific knowledge"[80]. On this view, defended by John McDowell, virtue does not consist in having the right principles, for there are no such things to be had – the moral world is too complex to be captured in finitely graspable rules, as Aristotle suggests several times in the *Ethics*[81] – nor does practical wisdom, correspondingly, consist in selecting the action that best promotes the end specified by the principles. Instead, as Myles Burnyeat puts it, when Aristotle talks about moral learning, he is pointing to "our ability to internalize from a scattered range of particular cases a general evaluative attitude which is not reducible to rules or precepts."[82] A virtue of character, Aristotle says, is a state or disposition (*hexis*) to feel pleasure, pain, anger, pity, and other emotions "at the right times, about the right things, toward the right people, for the right end, and in the right way"[83]. If this is the case,

---

[80] *NE* 1142a, 93.

[81] *NE* 1094b, 2; *NE* 1098a, 9. See also *NE* 1137b, 83-84, where Aristotle is discussing why laws, which are inherently universal, will sometimes lead to error in individual cases but are not the worse for that: "And the law is no less correct on this account; for the source of the error is not the law or the legislator, but the nature of the object itself, since *this is what the subject matter of actions is bound to be like*." (my emphasis)

[82] Burnyeat 1980, 72. Martha Nussbaum has defended a similar view, arguing that moral principles are "summaries or rules of thumb, highly useful for a variety of purposes, but valid only to the extent to which they correctly describe good concrete judgments, and to be assessed, ultimately, against these." (Nussbaum 1990, 68)

[83] *NE* 1106b. Aristotle does famously characterize the right end, right way, and so on, as a mean between extremes, but this is a singularly unhelpful characterization. I find Sarah Broadie's reading of this 'doctrine of the mean' persuasive. Though she is too reverent to come straight out and say that it is useless (she thinks that virtues are dispositions that protect from "excesses and deficiencies of feeling and impulse" (Broadie 1991, 101) that lead astray), she notes that some passages in Aristotle suggest that "one could discover independently that such and such a possible response would

McDowell argues, there is really not much substance to the so-called major premise or premise of the good:

> Having the right end is not a mere aggregate of concerns; it requires the capacity to know which should be acted on when. If that capacity cannot be identified with acceptance of a set of rules, there is really nothing for it to be except the capacity to get things right occasion by occasion: that is, the perceptual capacity that determines which feature of the situation should engage a standing concern. So the premise of the good, and the selection of the right feature of the situation to serve as premise of the possible, correspond to a single fact about the agent, which we can view indifferently as an orectic state or as a cognitive capacity. (McDowell 1998b, 30)

McDowell's reading emphasizes that virtue and *phronesis* really do go hand in hand, as *NE* 1144b says. This has a number of consequences. First, when it comes to moral learning, "the moulding of character *is* (in part) the shaping of reason"[84] When we learn to be virtuous, we are learning what reasons there are for doing things and what is really worthwhile. When we fail to become virtuous, we are also blind to (some) reasons there are. Thus, though moral education is not a rational process, it can be necessary for opening one's eyes to reasons. For Aristotle, you cannot convince everyone by rational argument that they should be virtuous. If they cannot see a reason for doing something that, say, decreases their pleasure though it would crucially benefit others, their failure is not one of rationality. There is no incoherence or contradiction involved. The strongest support for this kind of reading comes from passages like the following:

---

be intermediate independently of knowing that it would be right, and from this deduce that it would be right" and goes on to point out, rightly, that "there seems to be no independent sense of 'intermediate' such that every response is right to which that sense applies." (ibid., 100)

[84] McDowell 1996/1998, 184n33. Cf. McDowell 1998b, 40.

> What argument, then, could reform people like these? For it is impossible, or not easy, to alter by argument what has long been absorbed as a result of one's habits. … Arguments and teaching surely do not prevail on everyone, but the soul of the student needs to have been prepared by habits for enjoying and hating finely, like ground that is to nourish seed. (*NE* 1179b, 168)

It may thus take good luck to be able to appreciate what one has reason to do. Aristotle self-consciously addresses his treatise to the lucky: only those who already have a decent grasp of 'the that' (what is good) will benefit from an examination of 'the because' (why it is good), both because only they have the necessary cognitive grasp and because only they have the necessary motivation for the reflection to make a practical difference to their lives.[85] It should be emphasized that this in no way contradicts the view that the demands of virtue are the demands of reason.[86] All it means is that reasons and rationality can come apart, so that we can have reason to do something even if we could be perfectly rational while failing to recognize it. This is a decisive departure from the tradition of Hume and Kant, which is reaffirmed in our day by Bernard Williams, as discussed above. As McDowell sees it, one can come to be sensitive to reasons, including reasons one always had, by means of non-rational processes like being influenced by moving rhetoric, inspiration, and conversion.[87] For him, the point of the contrast that Aristotle draws between virtue and practical wisdom is that though one cannot have practical wisdom without having the right goal (and so recognizing genuine reasons), it is not exercise of practical wisdom that makes it the case that one's goal is right, but the shaping of one's character in good upbringing.[88]

---

[85] Compare *NE* 1095b, 4 and Irwin 1999, 176–177.

[86] It does mean, though, that insofar as moral responsibility requires being able to respond to reasons, bad luck in upbringing can lead to reduced responsibility. See also the discussion in section 3.2.

[87] McDowell 1995/1998, 100.

[88] McDowell 1998b, 31–32.

Secondly, if the ability to see reasons and being virtuous go together, Aristotle's so-called 'function argument' must be seen in a new light. In Book I of the *Ethics*, Aristotle argues that the good for anything that has a function (*ergon*) depends on what the function is – what it is to *do well* as a carpenter is determined by the function or characteristic activity of a carpenter as such. It also provides a criterion for which qualities of an individual *qua* a member of a functional class are true excellences or virtues (*aretai*) – precision is a virtue of the carpenter because it is required to build well, while wit is not his virtue as a carpenter, since it is not required for that. So if there is a function or characteristic activity of a human being, something "it is the business of a human being to do"[89] as McDowell has it, it will tell us what is good for human beings and what the true virtues are. Aristotle thinks there is such a thing. After ruling out merely staying alive (which is shared with plants and so could not be specifically human) and sense perception and movement (which are shared with animals), he concludes that "the human function is activity of the soul with reason or requiring reason"[90]. Doing well as a human being, then, is engaging in activities that involve the use of reason over the course of an entire lifetime.[91] The qualities that are truly human excellences or virtues are those that serve or constitute this kind of life. This, Aristotle admits, is just a sketch to be filled in later; sadly, he never explicitly does so in the works preserved to us. So it is unsurprising that the function argument has been taken to be many different things.

One understanding of it is as providing a metaphysical grounding for ethics in human nature – if the best life for human beings can be specified in advance of taking an ethical stance, and the traditional virtues best promote it, it can be shown to be in everyone's interest to be traditionally virtuous. Someone who cheats

---

[89] Ibid.

[90] *NE* 1098a, 9.

[91] As Nussbaum emphasizes, this should not be taken to exclude physical activities; rather, it means that in a good life rational activity "is the distinctive and guiding feature that gives life its characteristic overall shape" (Nussbaum 1995, 113–114).

in a business deal, for example, harms herself in a way that can be specified in non-ethical terms. McDowell rejects this reading. For him, Aristotle's talk of human nature should be seen as a "rhetorical flourish"[92], a way of framing the issue about what kind of life is best in a way that connects it to the question of which character traits really are virtues. When Aristotle says that the human *ergon* is engaging in rational activity, all this does is exclude a brutish or solitary life from being the good life.[93] The argument does not provide an Archimedean point outside the fray of moral argument for choosing between various kinds of activities involving rationality, like selfish pursuit of riches and fame or devotion to finding a cure for cancer. There is no external validation for ethics, and no need for it. The fact that it is in one's self-interest to be ethical is not visible, so to speak, from outside an ethical perspective – it is good for you to be honest, but that is not to say that honesty best satisfies the desires of any human being or guarantees the most pleasure to her, regardless of her upbringing. Instead, in reflecting on what the true virtues are or what it is the business of human beings to do, we may and must draw on our substantial ethical convictions. McDowell likes to borrow Neurath's famous coherentist image of repairing a ship while still at sea – one can remove and replace a particular plank only while relying on other planks, which may in turn be removed and replaced later.[94] Just so, we can vindicate the status of honesty as a genuine human excellence only within a scheme of values that determines, among other things, what are characteristically human activities. That is why those who lack the grasp of 'the that' cannot grasp 'the because', the kind of justificatory story that shows the point of having a particular virtue by placing it in a larger scheme of things, either. By contrast, "[s]omeone who is well brought up has the beginnings [needed for ethical inquiry – AK], or can easily acquire them."[95]

---

[92] McDowell 1980/1998, 19.

[93] McDowell 1980/1998, 13. Compare McDowell 1998b, 35.

[94] McDowell 1996/1998, 191–195; McDowell 1998b, 36–40.

[95] *NE* 1095b, 4.

The third consequence of linking virtue and practical wisdom closely is a distinctive picture of what moral judgment amounts to. On McDowell's view, virtue and practical wisdom manifest themselves first in the correct morally loaded perception of the situation; certain features of it appear to the agent as calling for a response, as contributing to the moral shape of the situation. Moral reasoning is a matter of coming to recognize these demands and balancing them without the aid of principles, and its aim is to arrive at the correct conception of the situation. But how can a correct conception of the situation, a kind of belief, be motivationally efficacious? McDowell's influential reading has it that for the virtuous person, the thing to do in the situation stands out as practically salient, and that is both a cognitive and motivational state – nothing else seems *appealing* to her, as we might say. So, for example, someone who lacks the concern that goes with (is one with) the virtue of temperance, and to whom, say, a tryst with an intern thus seems appealing, is not in the same cognitive state (does not perceive the world in the same way as) the truly temperate person.[96] Lacking the proper concern (motivational state), he has at best a limited grasp of what virtue calls for in this situation, since the knowledge in question is not knowledge of principles but a situation-specific sense of what really matters.

As McDowell puts it, the competing concerns are "silenced" by the virtue.[97] In the Aristotelian tradition, there is thus a fundamental difference between the virtuous person and the continent (enkratic or strong-willed) person who is tempted by vice but succeeds in resisting it. The virtuous are not tempted by vice. The idea of an extramarital relationship, for example, does not seem appealing to him. The continent person, in contrast, feels the pull, and so perceives the world differently, but has enough of a hold of virtue to stay on the straight and narrow. The incontinent (akratic or weak-willed) person, then, gives in, though he knows better – on Aristotle's explanation, he does not attend to his knowledge because of an unruly desire, and has it only in the way a drunk knows how

[96] See, for example, McDowell 1998b, 47.

[97] McDowell 1979/1998, 56.

to walk straight.[98] His perception could be the same as the continent person's one, but virtue is not as deeply ingrained in his habits.[99] Finally, of course, there is the vicious person who lacks not only the situation-specific sense of what matters but also the shallower understanding of principles that the continent and incontinent have. Actions that are in fact bad appear as good to the vicious person, and if not weak-willed[100], he will decide to pursue them.

Let us now return to the original question about the authority of morality. It should be clear by now that Aristotle indeed sees the demands of virtue as the demands of reason, though not of rationality itself. But how about their felt authority? The virtuous

---

[98] *NE* 1146b-1147a, 103. More precisely, the incontinent person's general knowledge of the good (such as "fatty foods are bad for you") does not get activated, because as a result of expected pleasure or pain, he does not focus on the fattiness of the pizza in question and so lacks the minor premise ("this pizza is fatty") needed to draw the inference: "Since the last premise is a belief about something perceptible, and controls action, this is what the incontinent person does not have when he is being affected." (*NE* 1147b, 104) The problem is thus one of perception of the particular case. That is why the incontinent person is "like a city that votes for all the right decrees and has excellent laws, but does not apply them" (*NE* 1152a, 113). This way Aristotle tries in a dialectical spirit to save the Socratic view that no one knowingly does what is bad – in one sense, the incontinent person knows what is good (so he is not vicious), but in another he does not.

[99] It could also be that the weak-willed person is lacking in what Philip Pettit and Michael Smith call 'executive virtues', virtues that are not about having the right end but about being motivated to pursue it in the right way (Pettit and Smith 1993, 76–77). Their examples of such virtues are temperance, courage, fortitude, and impartiality across times and persons (a kind of justice).

[100] Huckleberry Finn, who believed that black people were below whites (and thus was to that extent vicious) but could not help helping his friend Jim, is a now-classic example of 'inverse akrasia', weakness of will in pursuit of a bad goal. See especially Arpaly 2003. Aristotle himself brings up inverse akrasia in *NE* 1146a, 101, though he thinks that acting against one's best judgment is not incontinent if the pleasure that causes it is not shameful but fine (*NE* 1151b, 112).

person, or even the imperfectly virtuous person, which is what most of us are, does not just choose the right action, the action that contributes best to her well-being over the course of a lifetime and to that of her city. She also chooses it 'for the sake of the fine', as we saw above. The 'fine' is *kalon*, also beautiful and noble. Gabriel Richardson Lear argues that things are *kalon* in the Aristotelian sense when they visibly manifest order, symmetry, and boundedness (being limited to the right amount, not too much or too little) in effectively serving an end, which is naturally *eudaimonia* in the case of activity.[101] Contemplation of fine things gives pleasure, which is pride if the object is one's own action. Now, virtuous actions, especially those that benefit others, are fine in the Aristotelian sense – think of the admirable skill and the resulting pleasure of contemplating the actions of someone giving just the right kind of gift or holding a forward position under enemy fire for just the right time or saying just the right words in apologizing. Since virtuous actions are fine, doing them 'for the sake of the fine' does not stand opposed to doing them for their own sake. It certainly does not involve virtuous agent thinking "X would be fine, so I will do X". Rather, it seems, 'for the sake of" indicates that the virtuous actions get their *point* from contributing to a well-ordered, happy life, rather than contributing to pleasure or status. Actions that depart far enough from this goal are ugly or shameful; the brave person "stands firm against what is and appears frightening to a human being; he does this because it is fine to stand firm and shameful to fail"[102]. Again, it is tempting to read passages like this as presenting the virtuous person as being motivated by the fear of shame. Perhaps this is occasionally the case for the less virtuous, but in the case of the practically wise, we should rather say that virtuousness *shows itself* in the fact that one feels ashamed after doing something that falls obviously short of an ideal, just as one takes pride and pleasure in actions that meet it. On the Aristotelian picture, then, the felt authority of morality is explained by the fact that the virtuous have been brought up to feel pride in doing the right thing in the right

[101] Richardson Lear 2006.

[102] *NE* 1117a, 44.

way and shame for falling short of it because of expectation of pain or pleasure.

2.3 Empirical Study of Moral Thinking and Its Philosophical Implications

As I have emphasized, it is sometimes difficult to distinguish metaethical questions from empirical ones. In recent years, interest in empirical moral psychology has burgeoned, and claims of philosophical significance have not been wanting. Classic philosophical views are often dismissed as outmoded speculation on matters that are now at long last brought within scientific scrutiny. In the following, I will take a brief look at some empirical claims concerning the nature and causes of moral judgment, and discuss their potential philosophical relevance. My aim is to clarify the areas of overlap and difference in the explanatory questions that philosophers and psychologists are asking and articulate some desiderata that can help us decide between the answers.

2.3.1 The Moral/Conventional Distinction

It would appear fairly obvious that the question about the nature or essence of moral judgment in the product sense cannot be answered *a posteriori*. However, there is a research tradition in developmental psychology that sometimes, in addition to very legitimate questions, tries to do just that. Inspired by Lawrence Kohlberg's notion of stages of moral development[103], Elliot Turiel and his followers have

[103] See e.g. Kohlberg 1981. For a devastating review and critique see e.g. Shweder, Mahapatra, and Miller (1987, 11), who conclude that "[w]hat Kohlberg has firmly established empirically is that, with his interview

studied the emergence of the distinction between judgments about moral and conventional violations in children. Kohlberg, whose main interest was in the justifications that children of different ages offered for their verdicts on various moral dilemmas (most famously, should Heinz steal the drug necessary to save his wife or not?), found that small children explained their answers by reference to a punishing authority or self-interested reciprocity (stages 1 and 2) and teenagers by reference to functioning of the society (stages 3 and 4, or 'conventional morality'). Only later in the teens did people start to refer to rights and universal principles of justice (stages 5 and 6, or 'post-conventional morality'), adopting a critical perspective on parental authority and existing social structures.[104]

Turiel and colleagues challenged Kohlberg's reliance on explicit justifications and wanted to show that the distinction between conventional and moral rules emerges much earlier than Kohlberg's view allows. In broadest terms, their method is as follows: Children of different ages and ethnicities are presented with scenarios (in the form of stories or drawings) in which a rule that they are presumably familiar with, such as not hitting another child, not stealing an apple, sitting on a rug during story time, or putting a toy in its designated place, is violated, and they reliably judge these transgressive behaviours to be "not OK". Further questions are then asked: would the behaviour be OK if an authority figure, such as a teacher or parent, said it was? Would the behaviour be OK in another time and place? How serious is the violation in question? It turns out that in a variety of different cultures and communities, the children's answers form a distinctive pattern. According to children as young as 3 ½ years old (Smetana 1981), behaviour that involves hurting others,

---

methodology and scheme of concepts, children are more likely than adults to justify action verbally by reference to the subjective feelings of the self, and that adults make more reference to social and political institutions … in discussing their obligations." But that is it – children and adults (and adults in different socioeconomic groups) offer different kinds of verbal justifications, and even those cluster on Kohlberg's stages 2 and 3 (for children) and 3 and 4 (for adults).

[104] See Kohlberg 1981 for detailed discussion of these studies.

such as throwing sand on another child's face, is judged to be wrong regardless of what the authority figure says – even if the authority figure is God and the community in question is very religious.[105] It is also judged to be wrong in other places (such as other schools), and a serious violation. Justifications that are offered refer to others' welfare, fairness, and rights (the latter two kinds of justifications are increasingly offered as children get older). By contrast, behaviour that involves violating a social convention, such as wearing pyjamas to school, is considered by children to be okay if an authority figure, such as a teacher or parent, says so. It is also judged to be acceptable in other places, and engaging in it is not taken to be a serious violation. Justifications offered for these judgments appeal to obeying authority, avoiding punishment, or need for social coordination.[106] According to a recent overview by Turiel, there exist at least a hundred studies, conducted not only in the United States but also non-Western countries like India, Korea, Nigeria, and Zambia, confirming these results.[107]

This body of research, then, suggests that children distinguish between (at least) two different kinds of rules and violations.[108] I will call the first type of normative judgments involving authority-independence, generalizability, seriousness, and characteristic justifications in terms of harm Type 1 normative judgments, and the second Type 2 judgments. The social psychologists doing this research have not hesitated to label Type 1 judgments as *moral*. From a philosophical perspective, the first question this raises concerns the

---

[105] Nucci (1986) studied Amish children and found that though all of them believed that it would be all right to work on Sunday if God had not forbidden it, 80% believed that it would still be wrong to hit another person. This suggests that regardless of upbringing, children go with Socrates rather than Euthyphro.

[106] Smetana 1993, 115.

[107] Turiel 2002, 110–111.

[108] A third type of rule is prudential, concerning the agent's own welfare. Prudential rule violations are not taken to be as serious as moral violations, even when harm to others is small and harm to self is large (Tisak and Turiel 1984).

distinction between the *form* or functional role of moral judgment and its *content*. Are all *functionally* Type 1 norms – norms that are taken to be authority-independent, general, and serious – norms that have to do with harm, justice, and rights? What follows for the nature of moral judgment if they are? One answer, suggested by the work of Stich and colleagues[109], is that *if* all or most functionally Type 1 judgments concern harm and related considerations, these judgments form a psychological *natural kind* that is plausibly identified with moral judgment. For them, this would settle a long-standing philosophical debate about whether functional role or content defines what makes a judgment moral. Natural kinds here are understood as 'homeostatic cluster properties' in Richard Boyd's sense.[110]

This suggestion, not really defended by anyone (except perhaps Nichols 2004), does not work, for reasons familiar from earlier metaphysical disputes. To cut a long story short, if a particular content such as harm was part of the essence of moral judgment, it would follow that it would be *impossible* for something that does not have that content to count as a moral judgment. The truth of the claim would have modal consequences, indeed consequences for every possible world. So let us imagine a Twin Earth that is otherwise much like ours, except that people there make judgments with Type 1 functional role about giving gifts, which they take to be seriously wrong everywhere, regardless of what any authority says. They do not believe, let us assume, that giving gifts causes any physical or psychological harm to anyone; it's just wrong as a type of action, like callously breaking a promise is for us. The question is: is it possible that they take gift-giving to be *morally* wrong? It seems obvious that it is, and that they are indeed making moral judgments concerning it. If they were not doing so, we could not morally *disagree* with them, as Hare pointed out a long ago with respect to a similar scenario.[111] But then it cannot be the part of the essence of moral judgment that it concerns (what is taken to be) harm. Thus,

[109] Nado, Stich, and Kelly (forthcoming).

[110] Boyd 1988.

[111] Hare 1952.

even if Turiel and his colleagues are right and people *actually* only moralize what they take to be harmful behaviours, it does not follow that harm or any other content is part of the essence of moral judgment.

In fact, we do not have to go to Twin Earth to find that people moralize things that do not involve harm, justice, or rights. If that is so, moral judgment is not a psychological natural kind either, even given Boyd's liberal conception of natural kinds. Lockhart, Abrahams, and Osherson (1977) found that children considered certain social conventions, such as meanings of words, rules of hide-and-seek, eating with one's hands, or even driving on the right side of the road as functionally Type 1 norms, in spite of their evident arbitrariness and lack of basis in harm. Similarly, children in traditional Arab Israeli villages studied by Nisan (1987) treated violations like coed bathing and calling a teacher by the first name as Type 1 violations. Perhaps most extensive research in this vein has been carried out by Shweder and his colleagues in India and America. In the orthodox Hindu town of Bhubaneswar, Shweder, Mahapatra, and Miller (1987) asked both Brahman (high caste) and 'untouchable' (casteless) children and adults a series of questions about 39 different violations, including a widow eating fish, a woman sleeping in the same bed with her husband during her menstrual period, a son addressing his father by his first name, and eating beef.[112] All of these mentioned violations were treated by both children and adults as authority-independent and serious, though adults were more likely to be contextualist about the status of some violations (for example, although son using his father's first name is non-conventionally wrong, it is acceptable in the American context where the circumstances are different). Indeed, few violations were thought to be conventional. As Shweder and colleagues interpret the results, the Hindus moralize violations that do not have to do with harm, justice, or rights. To be sure, it is possible to see many of the violations as harmful in the Hindu belief context[113], but this is to

[112] Shweder, Mahapatra, and Miller 1987, 40.

[113] For example, Turiel points out in response that "it is believed that if a widow eats fish regularly it will cause offense to her husband's spirit. …

stretch the notion of 'harm' to triviality – in this sense, anything at all can be 'harmful', as long as it runs counter to some cultural belief or norm.

So, there is plenty of evidence against identifying Type 1 norms with harm norms. But are all violations of harm norms, at least, considered Type 1 violations? Recent research by Stich and his associates suggests that this is not the case. In an Internet survey of mostly American subjects, Kelly, Stich et al. (forthcoming) asked people to rate the wrongness of behaviours like whipping a sailor and keeping slaves while varying the time and place. As one might expect, most participants judged the behaviours were wrong when they were described as taking place in our day, but more than half rated whipping a sailor 300 years ago as okay. (Interestingly, only 11% found ancient slavery acceptable.) To test for authority-independence, they also asked about the wrongness of physically abusing military trainees to prepare them for interrogation in two conditions, when it was permitted by superiors and when it was forbidden by them. Around 60% of participants thought physical abuse in this context was acceptable if permitted by authorities, while only fewer than 10% thought so if it was forbidden. Kelly, Stich, and al. conclude that a significant number of people do not take harm-related violations to be authority-independent or generalizable. Thus, even if we adopted a homeostatic property cluster model of natural kinds, type 1 norms with harm content would not form a natural kind.[114]

In a broadside against the Turiel school, Howard Gabennesch argues that results like the preceding show that moral norms do not,

---

Adherence to these practices among Indians is connected to harm and its prevention – in these cases to nonearthly and nonobserved entities." (Turiel 2002, 172)

[114] It must be said that the Kelly, Stich, et al. (forthcoming) results are not particularly robust. Participants might well read additional morally relevant features into the scenarios they present – centrally, both sailors 300 years ago and contemporary military recruits might well be taken to have consented to harsh physical discipline when enlisting, and in both cases the rationale for such treatment is ready to hand.

after all, have a special status. Instead, all norms are deep down *conventional* and so basically arbitrary and changeable, but the conventional origin of some is simply masked by their age, applicability to everyone, unfamiliarity of function, complexity, lack of obvious utility, unvarying application in different contexts, relative stability over time, support from agencies of socialization like parents, teachers, and the legal system, lack of public deviance, and ideological support.[115] Consequently, the more 'transparent' the origin of norms becomes, the more likely it is that they are perceived as conventional, Gabennesch argues.[116] However, while Gabennesch's view has some explanatory force – highly educated people tend to think of more norms as conventional – its problems are all too obvious: not all norms lose their moral status when their origin becomes more transparent, and as philosophical moral realists like to note, it is no accident that some norms apply to everyone, are stable over time, and receive support from parents, the law, and ideology. Sociological facts about attitudes to norms support moral scepticism no more than they do moral realism.

Nevertheless, Gabennesch's critique of social psychological cognitivists segues into the second philosophically relevant question I want to discuss: why do children pick out certain norms as non-conventional and others as conventional? There is a variety of possible explanations relevant to moral epistemology and psychology. It could be that children have an innate moral sense that allows them to pick out norms that prohibit harm to others as having a special status.[117] It could also be that they get different feedback –

---

[115] Gabennesch 1990, 2054–2057. This is an argument against what is misleadingly called 'moral realism', the view that children (and adults) take harm norms to be Type 1.

[116] In his response to Gabennesch, Shweder argues that "the very idea that the social order is a conventional order is an expression of a culture-specific worldview", namely that of a subculture of academic liberalism (Shweder 1990, 2064). If so, it is hardly to be expected that everyone would, as an empirical matter of fact, converge on total conventionalism were the origin of moral norms transparent to them.

[117] This accords well with moral grammarian accounts, discussed below.

different sort of punishment or praise – for the norms they regard as non-conventional.[118] Turiel's own suggestion is that it is the children's own experience of pain and observation of the pain-behaviour of others that leads them to think that actions causing harm are wrong, while conventional violations are recognized as such because of adult reactions to breaches.[119] Perhaps the most straightforward account, however, is that norms that are regarded as non-conventional are those whose violation gives rise to affective reactions. This view is defended by Shaun Nichols, who calls his view the Sentimental Rules account. The idea is simple. Causing harm to other people gives rise to a strong affective response in most human beings. As a result, those social norms that happen to forbid causing harm to others get picked out as particularly important – they are "affect-backed norms", in Nichols's terms.[120] An advantage of this view is that it explains why people in many cultures moralize also actions that do not involve harm to others, but instead give rise to the affect of *disgust*. In Nichols's study, subjects regarded actions like spitting into a glass before drinking from it as both disgusting and non-conventionally wrong.[121] Further evidence for this view comes from the fact that psychopaths, who lack normal affective reactions to the suffering of others, fail to make the moral/conventional distinction – that is, they see nothing special about moral norms. Nichols quotes the psychopathic killer Ted Bundy, whose list of wrongs is entirely indifferent to the moral or conventional status of actions:

> It is wrong for me to jaywalk. It is wrong to rob a bank. It is wrong to break into other people's houses. It is wrong for me to drive

---

[118] This sort of view is defended by Prinz (forthcoming).

[119] See Turiel 1983. A basic problem with this account is that children's moral judgments are not limited to actions causing physical harm even in our culture, not to mention the Indian subculture studied by Shweder and colleagues.

[120] Nichols 2004, 21.

[121] Nichols 2004, 20–25. Compare also the research by Haidt and colleagues, discussed below.

> without a driver's license. It is wrong not to pay your parking tickets. It is wrong not to vote in elections. It is wrong to intentionally embarrass people.[122]

An alternative explanation for the deficit in psychopaths is offered by James Blair, who postulates a violence inhibition mechanism (VIM) that gets activated in normal humans in response to distress cues.[123] Blair also comes up with an adaptive rationale for VIM. It is as plausible as any just-so story in evolutionary psychology: surely people who refrained from beating crying babies to death left more offspring. On Blair's account, VIM is what underlies the moral/conventional distinctions – the rules whose violations activate VIM get designated as having a special status, and normal subjects often explain the wrongness of the violation by reference to the distress of others. While this explanation works for psychopaths[124], it fails to explain the non-conventional status of disgust-backed norms, for example, so Nichols's account must be regarded as superior to the extent that these norms really are taken to be moral by the subjects – the other alternative is to say that not all functionally Type 1 norms are moral.

On Nichols's view, the norms and the affects that back them are distinct from each other. This explains why affect does not need to be "online" every time someone makes a moral judgment – the norm is

---

[122] Nichols 2004, 112. Nichols is quoting from Michaud and Aynesworth, *Ted Bundy: Conversations with a Killer*. New American Library, New York, 1989, 116.

[123] Blair 1995, Blair 2006.

[124] To be sure, in Blair's study, the psychopaths did not, as expected, treat all norms as conventional, but as *moral*. He tries to explain this away by appeal to ulterior motives that the subjects may have had: "These subjects were all incarcerated and presumably motivated to be released. All wished to demonstrate that the treatments they were receiving were effective. They therefore would be motivated to show how they had learned the rules of the society." (Blair 1995, 23) The hypothesis is thus that faced with these incentives and unable to distinguish moral from conventional violations, the psychopaths erred on the side of caution and said that all violations were wrong in an authority-independent manner.

still marked as special due to its past association with affect. Nichols does not, however, want to subscribe to a purely developmental account, according to which affect only needs to be present during a period of moral learning; his hypothesis is that if the relevant emotions were to be eradicated, "over time, the tendency to treat harm norms as distinctive would wane."[125] At the same time, the dissociation between the affect and the norms raises the issue of why they seem to go together as a rule. Here Nichols's hypothesis is that the relationship to affect serves to explain the persistence of certain norms, including moral and etiquette norms – prohibitions that happen to coincide with independently specified emotional reactions like disgust get remembered and propagated over time, while others fall into the mists of time.[126] Were this kind of story correct, it would call into question the moral realist alternative, according to which change in moral norms concerning slavery and the treatment of women, for example, is a result of growing awareness of moral facts.[127]

As attractive as Nichols's simple story is, it has some obvious weaknesses as well. First, are affects and norms really independent from each other in the way that his account requires? If what we find distressing or even disgusting depends on what we regard as prohibited, the direction of explanation runs in the opposite direction: what is considered wrong is regarded as distressing or disgusting. This seems to be the best explanation of some empirical data: surely if people find cleaning a toilet with a flag disgusting[128], it is not because the action is inherently such, but because it is regarded as inappropriate given the norms requiring respect for what the flag symbolizes. Similarly, physical harm that is regarded as just punishment surely does not give rise to the same emotional reaction as the same harm would if it was regarded as unjust. Second, philosophical discussions of rule-following throw into

[125] Nichols 2004, 99.

[126] Nichols 2004, chs. 6 and 7.

[127] For the realist explanation of moral change, see Brink 1989 and Sturgeon 1988.

[128] As shown in Haidt, Koller, and Dias 1993.

question the Platonistic model of norms that Nichols's account seems to presuppose – that is, it may not be the case that once we latch onto rails that are out there, we can go on the same way without having appropriate emotional propensities.[129] This has some empirical support, too: while psychopaths and those suffering from acquired sociopathy are able to classify correctly paradigmatic norm violations, they get lost when a more-fine grained response is called for, suggesting reduced moral competence. A more complex story is thus needed. Perhaps recent theories of the causes of moral judgment will help, though they are not focused on the moral/conventional distinction.

2.3.2 The Process of Moral Judgment

Why is it that people respond to particular cases as they do, say by judging the behaviour in question to be wrong? That is, what kind of psychological process is moral judgment in the process sense? One explanation would be that people consciously hold certain general moral principles that have to do with welfare, harm, and justice, recognize the case in question as falling under one (or more), and come to a verdict as a result. This seems to be the assumption in the tradition deriving from Piaget and Kohlberg. However, various different experiments have called this simple rationalist model into question. I will next discuss the best-known new alternative models, which all draw heavily on experimental results.

Affectivist Accounts of Moral Judgment

The first sort of evidence comes from various 'dumbfounding' studies conducted by Jonathan Haidt and his colleagues. Here is perhaps their most famous case:

[129] The *locus classicus* for this kind of criticism is McDowell 1981.

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love? (Haidt, Björklund, and Murphy 2000; Haidt 2001, 814)

Most people say that Julie and Mark's incestuous night is wrong. But when they are asked for reasons why, their answers are confused. A typical subject mentions the dangers of inbreeding, even though Julie and Mark use multiple forms of birth control. When the researcher points this out, the typical subject comes up with a different reason, such as the emotional problems associated with incest. When the researcher reminds the subject that in the story there are no such problems, she either gropes for yet another reason or says something like "I just know it's wrong". In Haidt's terms, they are 'dumbfounded': they cannot explain why they make the judgment they do, but they hold on to it nonetheless, and come up with bad reasons if they are asked to explain. Haidt and his colleagues have found the same pattern in a number of studies featuring cases like masturbating with a dead chicken or cleaning the toilet bowl with a flag, in which, according to their hypotheses, no harm is caused by the actions.[130] What best explains this?

---

[130] For these cases, see Haidt, Koller, and Dias (1993). The assumption that the cases involve 'no harm' is problematic, however. It requires restricting 'harm' to concrete physical or psychological damage, which is fair enough insofar as the target of criticism is the Turiel school. However, for the subjects studied, desecrating the flag may well involve *symbolic* harm, and masturbating with a chicken may well be taken to indicate a damaged sexual psychology, whatever the researchers say. They may thus well have a kind of harm-based rationale for their moral judgments, contrary to what the studies assume.

Drawing on a large body of work in social psychology, Haidt argues that there are two kinds of cognition issuing from what he calls the 'intuitive' and 'reasoning' systems, often also known as 'system 1' and 'system 2'.[131] The intuitive system is fast, effortless, automatic and unintentional, and only its products but not processes are accessible to consciousness. Many of its elements are probably evolutionary adaptations. The reasoning system, by contrast, is slow, requires effort and attention, and involves at least some consciously accessible and controllable steps and verbalization. Haidt's thesis, crudely put, is that moral judgments issue from the intuitive system rather than the reasoning system. In particular, the intuitive process involved in moral judgments works by way of affect. Haidt endorses Antonio Damasio's 'somatic marker hypothesis', according to which prudential and moral decision-making proceeds in normal subjects on the basis of associations of experiential stimuli with bodily feelings.[132] The evidence for this comes mainly from subjects with damage to their ventromedial prefrontal cortex, whose function appears to be integrating the feelings in question to decision-making. After injury, these subjects make erratic judgments in spite of having their abstract reasoning capacities intact. Further, the studies by Haidt and his colleagues – as well as Nichols's work discussed above – suggest that the affect of disgust drives many non-harm-based judgments (for example in the case of masturbating with a chicken). What is more, the affective state of the subject need not have anything to do with the object of evaluation. Wheatley and Haidt (2005) found that hypnotizing susceptible participants to experience disgust at the sight of random words resulted in a difference to their moral judgments about written scenarios.[133] Valdesolo and DeSteno

---

[131] Haidt 2001, 818–819. For the general picture of two very different cognitive systems see also Zajonc 1980, Bargh 1994, Bargh and Chartrand 1999, Bargh and Ferguson 2000, and Wilson 2002.

[132] See Damasio 1994.

[133] It is important that the hypnotized disgust was entirely rationally irrelevant to the moral status of the events of the story; for example, in one case, it was aroused by reading the word 'often'. The actions themselves were innocent, like fostering good discussions.

(2006) had people watch five minutes of comedy (Saturday Night Live) to put them in a good mood, and found that it made people more likely to judge that it is morally appropriate to push a fat man in front of a trolley to save five others (see below for the trolley dilemmas).

In short, on this sort of view, moral judgments are caused by automatic, non-rational affective reactions. The reasoning system is activated as a rule only in interpersonal contexts of attitude modification, in which people are called to 'rationalize' their intuitive judgments post-hoc, by appeal to reasons and principles that have little or nothing to do with their original judgments but have currency in their social environment. As Haidt puts it, "moral reasoning does not cause moral judgment; rather, moral reasoning is usually a post hoc construction, generated after a judgment has been reached"[134]. Though in rare cases, particularly in the ones in which gut reactions conflict, conscious reflection may make a difference to moral judgments, the chief causal effect of explicit moral reasoning is on the judgments of *other* people.[135] Since these theories claim that moral judgments are arrived at in virtue of primitive affective reactions rather than any rational process and that reasoning has at best a secondary role in moral thinking, I will call these theories *affectivist*.[136] Somewhat more modest versions of this type of theory, such as the of Joshua Greene and colleagues, allow for conscious reasoning to be effective with respect to impersonal judgments, but still maintain that emotional reactions drive personal moral judgments; I will call this kind of view *semi-affectivist*.[137] The simplest versions of affectivist theory concern moral judgments made about other people's actions, but it is easy enough to extend the model to arriving at first-person moral ought-judgments. Presumably,

---

[134] Haidt 2001, 814.

[135] See especially Haidt and Björklund (forthcoming).

[136] Haidt labels his view 'social intuitionist', but this since this is apt to be very misleading, given that the view has little to do with the philosophical views called intuitionism. I will use a more descriptive term instead.

[137] For the personal/impersonal distinction, see below.

deliberation must proceed in something like the following manner. First, we imagine the outcomes of various possible actions.[138] Second, we have different affective reactions to these imagined outcomes. Third, we pick the one that generates the most positive (or least negative, as it may be) affective reaction. And finally, if asked, we come up with a story that purports to justify (show that there is most reason for) the alternative we have chosen.[139]

Is There a Universal Moral Grammar?

A rival experimentalist school argues that there exists an innate, universal moral 'grammar' that operates automatically and unconsciously in a moral faculty, producing 'ethicality' judgments, just like the Chomskian innate language faculty is meant to come up with grammaticality judgments. Chomsky famously argued that whatever feedback children get from their environment, no amount of behaviourist learning can possibly be sufficient to give rise to our knowledge of the grammaticality of a potentially infinite number of novel sentences (the poverty of the stimulus argument). Language learning is fast compared to learning in general and has age-specific stages or critical periods that seem to universal. Moreover, if we look at all the languages across the world, we find that they employ only a small portion of the theoretically possible grammatical structures. To explain these phenomena, Chomsky postulated an innate language faculty with built-in abstract principles whose parameters are set by the child's linguistic environment, giving rise to the variety of languages we have.[140] To put it crudely, in some sense, the

[138] Presumably this will rely on some heuristic about which alternatives are relevant – for the model to have any plausibility, it cannot require that we go through any very large number of the physically possible alternative outcomes.

[139] This description of affectivist deliberation is intended to capture general features of the view, not to paraphrase any particular theory. At least Haidt (2003, 198) comes close to explicitly endorsing this sort of picture.

[140] This 'principles and parameters' view, first articulated in Chomsky (1981), is the just one incarnation of Chomsky's theory.

child already knows, for example, that all complete sentences must have a subject (even if it is not always explicitly mentioned, it comes out in transformations of the sentence); the linguistic environment tells where in the sentence to put the subject relative to predicate expressions.

The original inspiration for moral grammarians comes from Rawls, who drew an analogy in *A Theory of Justice* between the work that linguists do with linguistic intuitions and moral philosophers do with moral intuitions. Rawls suggested that normative ethics could be seen as in part articulating the tacit principles that guide everyday moral judgments.[141] John Mikhail, as well as Marc Hauser and his colleagues, take the linguistic analogy much farther. Armed with a Chomskian model, they claim that moral competence is partly innate, a product of a module that contains universal principles as well as parameters that are set by the child's moral environment. The basic argument is simple. Various empirical studies, including those in the moral/conventional paradigm, suggest that even young children are able to make complex moral distinctions about the sorts of cases they have never before encountered (even if their moral *performance* does not always match these judgments, given underdeveloped capacities for self-control and mind-reading, for example[142]). However, when people, children or adults, are called upon to justify the moral choices they make, they are stumped, often pointing to features that could not possibly explain their decisions. As Hauser puts it, "When people give explanations for their moral behaviour, they may have little or nothing to do with the underlying principles. Their sense of conscious reasoning from specific principles is illusory."[143] The best explanation for why they

---

[141] Rawls 1971, 46–47.

[142] For the competence/performance distinction in ethics, see Hauser 2006, ch. 5.

[143] Hauser 2006, 67. Cushman, Young, and Hauser (2006) found that subjects' ability to articulate the principle guiding their responses depended on the principle in question – most people were able to say that action is worse than omission, but few could explain that they judged a case more severely when a bad consequence was intended rather than a side effect.

nonetheless make sophisticated distinctions, according to Mikhail and Hauser, is that (normal) individuals possess a moral grammar, a system of tacitly known rules, concepts, and principles that enables them "to determine the deontic status of an infinite variety of acts and omissions"[144]. Just as linguistic grammar makes possible quick, automatic judgments of grammaticality of novel linguistic expressions (what Chomsky calls "language perception"), moral grammar makes possible quick, automatic judgments of moral status ("moral perception"). Mikhail, who follows the linguistic model most closely, goes so far as to break down moral perception to three parts analogous to the linguistic case: deontic rules ("intentionally causing bodily harm is prima facie wrong", "causing bad consequences that are known but not intended is more acceptable than intending harm"), structural descriptions of actions in the abstract terms in which the deontic rules are defined ("action x is a case of intentionally causing bodily harm"), and conversion rules that get from perceptual stimulus to the morally loaded structural descriptions ("Joe pushed Jack off the bridge" to "Joe intentionally caused bodily harm to Jack").[145]

If, indeed, our moral judgments result from a complex, automatic computational process of the sort grammarians describe, the next question is how we could possibly *acquire* such a mental system. The moral grammarians argue that, just like in the linguistic case, there is a *poverty of moral stimulus* – children are not *taught* to make all the fine distinctions they do make, and indeed could not learn to make them on their slim experiential basis. Thus, they postulate an innate moral module that is specialized in the sort of analysis and computation that the moral grammar requires. It works independently of both general reasoning capacities and emotional reactions, though it does require input from other subsystems (like mindreading) and its output may give rise to emotional reactions.

---

[144] Mikhail (forthcoming). Cushman, Young, and Hauser (2006) endorse a multiple systems model, in which the moral grammar module is only a part of the story.

[145] For a detailed description of these various rules and principles, see Mikhail 2000, Mikhail (forthcoming).

Trolleyology: Deciding Between Empirical Accounts

How can we decide between rationalist (Kohlberg, Turiel), affectivist (Haidt, Greene), and computational (Mikhail, Hauser) accounts of the processes that give rise to particular moral judgments? One way is to look at patterns in the judgments that people make in response to carefully constructed cases, and in particular how changing the cases gives rise to variations in intuitive judgments. For this purpose, Greene, Mikhail, and Hauser, together with their colleagues, have collected a large amount of data of people's intuitive responses to the so-called trolley problems. (A lot of this data is generated through the Moral Sense Test on the Internet: http://moral.wjh.harvard.edu/.) Trolley problems are moral dilemmas that were originally introduced by Philippa Foot and Judith Jarvis Thomson to get at intuitions about the moral status of actions and omissions, intentions and side effects, agents and bystanders, and so on. Canonical variations include the following:

> (Switch) A trolley is about to run over five people on the tracks[146] and kill them. John happens to be walking by and notices that he could save the five people by hitting a switch that turns the trolley on another track. However, there is someone on the other track as well, so saving the five would mean bringing about the death of the one. Should John hit the switch?
>
> (Fat Man) A trolley is about to run over five people on the tracks and kill them. John happens to be crossing a footbridge where a fat man is standing over the tracks. If John were to push him over the edge, his heft would suffice to stop the trolley before it reached the five people.

[146] In fact, trolleys do not run on tracks but on wheels, but I will follow the philosophical tradition and pretend that they do!

> However, this would mean the death of the fat man. Should John hit the switch?

Foot and Thomson used these cases to provide intuitive support for the doctrine of double effect, the claim that a knowingly bringing about a bad outcome (killing a person) as a *side effect* of bringing about a good outcome (saving five) can be morally permissible, while bringing about a bad outcome as a *means* to a good end is not. The experimentalists, in contrast, are not directly interested in normative theory, but in the processes that underlie intuitive judgments and variation in them.

To begin with, why do people give different responses to Switch and Fat Man? The rationalist views, especially Kohlberg, would appeal to consciously held justificatory principles, but the existing empirical data provides little support for this view – virtually nobody cites anything like the doctrine of double effect to justify their differential responses. Affectivists like Haidt have not (to my knowledge) tried to explain trolley intuitions, but the semi-affectivist or dual process model of Joshua Greene and his colleagues is developed partly to deal with them. Greene et al. suggest that the difference between the cases lies in the personal/impersonal dimension: while Switch involves deflecting an existing threat toward the one from the five, Fat Man involves creating a threat of 1) physical harm 2) through one's own agency 3) to a particular person.[147] They hypothesize that violations that are personal in the sense of meeting these three conditions give rise to evolutionarily basic and early emotions that inhibit harming actions (compare Blair's VIM). This explains why people think it is wrong to push the Fat Man down.[148] Functional magnetic resonance imaging (fMRI)

---

[147] Greene et al 2001, Greene and Haidt 2002.

[148] On Greene's view, this amounts to emotions interfering with rationality, since the rational thing to do in both situations would be to sacrifice the one to save the five. He even goes so far as to claim that empirical evidence shows that deontological theories are attempts to rationalize gut reactions post hoc, while consequentialist theories involve genuine reasoning, a rational and cognitive process! Here is a representative

studies conducted by Greene and his colleagues appear to support this. Briefly, when people make the more abstract and utilitarian choice in Switch, the dorsolateral prefrontal cortex, an area of the brain associated with conscious problem-solving, is particularly active, and when they are faced with Fat Man, areas associated with emotion (the posterior cingulate cortex, the medial prefrontal cortex, and the amygdala) are more active.[149] Also, the reactions of people who say it is all right to push down Fat Man are slower than the reactions of those who say it is not, which Greene and colleagues hypothesize to result from it taking time for reasoning to overcome an initial emotional verdict.[150]

Moral grammarians disagree. According to them, the essential difference is the one uncovered by philosophers: in Switch, the death of the innocent bystander is a known side-effect, but in Fat Man, it is a necessary means. Of course, few people explicitly entertain the doctrine of double effect, but it is alleged to form a part of the universal moral grammar that can be constructed on the basis of people's reactions to these artificially constructed cases. To remove possible confounders, Mikhail and his colleagues added a different variation on the Moral Sense Test, a 'loop' case in which a large person, heavy enough to stop the trolley, is walking on a side track

---

passage: "Talk about rights, respect for persons, and reasons we can share are natural attempts to explain, in "cognitive" terms, what we feel when we find ourselves having emotionally driven intuitions that are odds with the cold calculus of consequentialism. Although these explanations are inevitably incomplete, there seems to be "something deeply right" about them because they give voice to powerful moral emotions." (Greene (forthcoming a))

[149] Greene et al. 2001, Green et al. 2004.

[150] Greene et al. 2004, 393. Further evidence for this is that according to some preliminary results, burdening subjects with a cognitive workload slows down their utilitarian judgments, but not deontological ones, which is to be expected if the latter are automatic and affective (Greene (forthcoming b)). This is not to say that other explanations could not be found – it is hardly surprising that it is quicker to discover what rules that forbid certain types of actions require than what rules that tell you to calculate and compare outcomes require.

that loops back to the track on which the five are. Here it is a matter of deflecting an existing threat, not creating a new one, but at the same time, killing the large person is a necessary means for saving the five. Since the case is impersonal, semi-affectivist views like Greene's predict that people would judge killing the one to be permissible, while if people's judgments are subconsciously guided by the doctrine of double effect, they should find it impermissible. The results are mixed: 55% say it is permissible, in contrast to 89% in Switch and only 11% in Fat Man.[151]

If variations in *cases* do not yield a verdict on the psychological mechanisms leading to moral judgment, another possibility is to vary the capacities of the *subjects* and see if that makes a difference. Particularly interesting here are psychopaths and acquired sociopathy patients, whose emotional reactions are known to be abnormal. If their reasoning capacity and moral module (assuming one exists) are intact, but moral judgments differ, we may conclude that the presence of emotions at least partially explains the judgments of normal subjects. It turns out that these people *do* think it is okay to push the Fat Man. Koenigs, Young, et al (2007) presented six patients whose social emotions (particularly guilt, shame, and empathy) were seriously impaired due to damaged ventromedial prefrontal cortex (VMPC) with both personal and impersonal moral scenarios. In impersonal and low conflict scenarios, their judgments matched with normal control subjects. But in personal high conflict scenarios, like the choice between suffocating a crying baby or revealing the location of five people to a death squad, VMPC patients were significantly (4.7 times) more likely to choose the utilitarian option. Koenigs, Young, et al. conclude that "[i]n the

[151] For these data, see Hauser, Young, and Cushman (forthcoming). The doctrine of double effect –story does receive some additional support from a fourth variant, in which a heavy weight on the side track behind the one person is added to the otherwise identical picture. In that case, the existence of the weight is a necessary means for saving the five, while the death of the one person is merely a foreseen side-effect. When killing the one is no longer a means but a side effect, 72% instead of 55% of people judge the action to be permissible.

absence of an emotional reaction to harm of others in personal moral dilemmas, VMPC patients may rely on explicit norms endorsing the maximization of aggregate welfare and prohibiting the harming of others"[152]. Conversely, the results suggest that social emotions play a significant role in normal people's judgments in cases that involve personal engagement, though not in impersonal scenarios. Should we then conclude that a dual process model like Greene's is superior? This would be too quick. It could be that *also* moral reasoning capacities or the hypothesized moral module are affected by the condition, though general reasoning abilities and intelligence are apparently intact. By trolleyological standards, the empirical debate remains open.

2.3.3 Philosophical Implications

What implications does all this recent social psychological and cognitive science research on moral judgment have for metaethics? Let us begin with a crucial distinction I mentioned in passing above, namely the distinction between the *process* and *product* senses of moral judgment. The psychological research has focused almost exclusively on the process of moral judgment, while contemporary metaethicists have focused almost exclusively on the product. As the discussion in section 2.2 showed, this was not the case for the classics of moral philosophy: they were interested both in the nature of the process and the nature of the product. What is more, they were not just interested in the process and product separately, but, as one might say, in their *interdependence*. On the one hand, it is the nature of the process of judging that explains the motivational and representational properties of the product it gives rise to. On the other hand, since we also have an independent grasp of what constitutes a moral stance, not just any way of arriving at it counts as a process of moral judging, especially as moral deliberation (which I

[152] Koenigs, Young et al. 2007.

will understand simply as the process of arriving at a first-personal moral ought-judgment, whether or not it is a process of reasoning). For example, a distinctively moral process of judging might involve some kind of reflective correction of known bias, while taking a pill that changed your moral views would not count as a process of moral judgment in this sense, in spite of its causal outcome. If we always formed judgments about good and bad by taking some sort of pill or on the basis of simple affective reactions like disgust, it would be utterly baffling why we invest those judgments with a special authority and, for example, feel guilt if we act against them, or think that they can be justified to others.

To say that not just any way of arriving at a moral judgment is a process of moral judging is not to use the term in an honorific or success sense – for example, a racist may engage in reasoning that counts as moral in this sense and yet arrive at the *wrong* answer. But neither should philosophical claims about what constitutes moral judgment in the process sense be understood as merely descriptive or statistical. Rather, they describe the necessary conditions of the sort of process that makes the essential properties of the product intelligible. It need not be the case, nor does any philosopher claim it to be the case, that we engage in that sort of process of reasoning or emotional correction or moral perception *every time* or even most of the time we moralize. Rather, there is a relationship of *asymmetric logical dependence*: a person who *never* engaged in the sort of judging the account describes would not count as making moral judgments, having utterly failed to grasp the point of making them, whether we understand the point as arriving at correct judgments about practically relevant features of the world, as realists take it, or as contributing to social coordination, as many expressivists see it.[153] Snap judgments made on the basis of simple affect or mood are thus *parasitic* on judgments made in the favoured way. Think of people who rant and rave about whatever makes them feel bad and fail to see any need for their reaction to be justifiable to others. Sometimes we think of these people as just bad moralizers, but at some point we

[153] I owe the idea of this sort of dependence to Evan Simpson (1999).

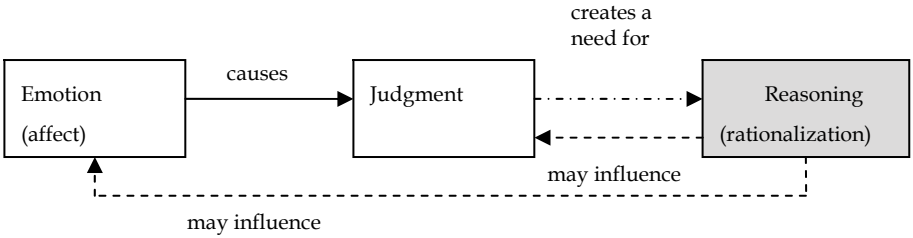
just have to say that in spite of speaking the words, they are just not engaged in the practice of moralizing – the 'disagreement' between us and them is not about what to do, morally speaking, since they do not really hold moral views, but just a difference in what we actually do.

The interdependence of the process and the product gives rise to a number of desiderata for theories of the process, whether psychological or philosophical. The description of the process should

- make intelligible the *motivational role* of moral judgments
- make intelligible the *felt authority* of moral demands
- make intelligible either how the judgments it gives rise to have genuine *normative authority* or how people have come to make the mistake that they do
- be compatible with Moorean facts about the difference between *first-personal deliberation* and moral evaluations of others
- make intelligible the *ubiquity and variety of moral argument* and the related belief that our moral judgments, unlike mere likings, are at least in principle *justifiable to others*
- be compatible with well-grounded theories of the nature of *desire, belief, emotion, reasoning, and perception*
- be compatible with *ecologically valid experimental results*

With these desiderata in mind, I will next quickly review the strengths and weaknesses of some recent empirical accounts and discuss the challenges the experimental results may present to philosophical theories. I will use diagrams of the various models to focus on their distinctive features. As a rule, solid arrows represent causal connections, dashed arrows optional connections, and shaded boxes exercises of psychological capacities (rather than simple mental states).

Haidt's 'social intuitionist' or affectivist model can be captured as follows[154]:



In short, the claim is that as a rule, non-rational affects like disgust give rise to moral judgments, which are subsequently rationalized if the social context calls for it. (The rationalizations themselves may serve as input to other people's judgments, a step not diagrammed here.) It does not matter what gives rise to the affect (hypnosis will do), so I have not diagrammed its antecedent. The dashed lines represent the possibility that on a rare occasion, conscious reasoning processes may make a difference to judgment or emotion; usually, though, they are mere "confabulation".

The affectivist model does not fare well with the desiderata. It does not even begin to make intelligible the felt authority of moral demands nor their distinctive motivational role. Neither much resembles the motivational push of the affect that is hypothesized to give rise to moral judgment – why would doing something disgusting give rise to *guilt*, when our judgments of what is disgusting do not involve a commitment to justifiability to others? The affectivist view thus leaves the nature of moral judgment in the product sense entirely mysterious. The issue of normative authority is not in view, and the picture of moral deliberation as consideration of potential affective consequences of actions is implausible and without empirical support. Terms like 'emotion' are carelessly thrown around without attention to the relationships among

[154] Cf. Haidt 2001, 815.

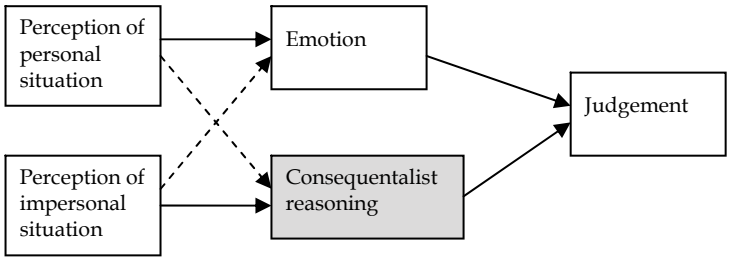
cognitive, affective, and motivational components of emotional states. Finally, while Haidt talks about the role moral argument, it is unclear why on his picture "moral reasons passed between people have causal force"[155]. If moral judgments are not influenced by reasons, why construct reasoned arguments when trying to persuade others? And indeed, on closer look, Haidt is not really talking about argumentation at all: "The reasons that people give to each other are best seen as attempts to trigger the right intuitions [i.e. affective reactions – AK] in others."[156] Philosophers will be reminded of Stevenson's early emotivism, which similarly elided the distinction between rational and non-rational persuasion.[157] Of course, there is no denying that persuasion is often non-rational and strategic, but surprisingly often arguments for moral positions are at least valid, if not so often sound. Nor is rhetorical flourish simply antithetical to argument. On the desiderata I outlined, then, the affectivist model seems like a failure.

[155] Haidt and Björklund (forthcoming).

[156] Ibid.

[157] For Stevenson, "*[a]ny* statement about any matter of fact which *any* speaker considers likely to alter attitudes may be adduced as a reason for or against an ethical judgment. Whether this reason will in fact support or oppose the judgment will depend on whether the hearer believes it, and upon whether, if he does, it will actually make a difference to his attitudes" (Stevenson 1945, 114–115).

Does Greene's 'dual process' or semi-affectivist model fare any better? It can be summed up in the following diagram (all arrows indicate causation or possible causation):

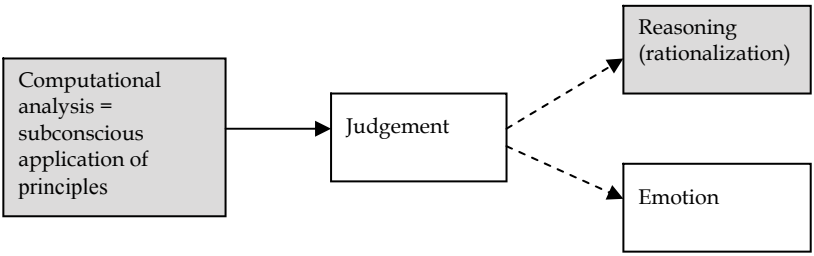


On this picture, affect drives mainly judgments triggered by personal situations, while mainly cool utilitarian reasoning is employed to settle impersonal issues. The basic problems of the affectivist model remain here: there is no explanation, and no attempt to explain, the phenomenology and functional role of moral judgment, either in personal and impersonal cases. Nor is the model of first-personal deliberation any more plausible. However, dual process models could in principle do better at explaining or undermining the normative authority of moral demands. For Greene, in brief, the intuitive judgments arising from emotional processes are irrational and unjustified, though there is an evolutionary story to be told why we enjoy punishing, for example. He argues that deontologists like Kant are therefore unwittingly in the business of rationalizing their gut reactions, and that realizing this should shift the balance of debate in normative ethics to a consequentialist direction.[158] Greene thus has the normative implications of his account of moral judging in view, and concludes that they are undermining in the case of affect-based judgments. But how about impersonal judgments made on the basis of conscious

[158] Greene (forthcoming a).

cost-benefit reasoning – is their felt authority warranted? This is a difficult question to answer, since Greene's account does not explain why these judgments – in contrast to other cost-benefit calculations – are experienced as being grounded in desire-independent demands in the first place.

The final theory I will review is Mikhail and Hauser's moral grammarian or computational model[159]:



The basic problems with the account are by now familiar: knowing that designation of something as morally wrong, say, results from a complex subconscious computational process does not give us any insight into why thinking that something is morally wrong has the sort of phenomenological, motivational, and deliberative role it does. This model simply has to assume that there is some other story to be told of why we feel guilt, for example, for doing something we think is wrong, a story that explains the properties of the judgment without making them intelligible. But in addition, the grammarian model raises questions the affectivist ones do not. There is no doubt that we have emotions and that they often make a difference to our judgments. But do we really have a 'moral module' analogous to the language faculty many linguists postulate? If the analogy fails to hold, what is left of the model? Let us begin with Steven Pinker's simple characterization of the Chomskian view of language:

[159] Cf. Hauser 2006, 45.

> 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. For these reasons some cognitive scientists have described language as a psychological faculty, a mental organ, a neural system, and a computational module. (Pinker 1994, 4–5)[160]

Can we substitute 'morality' for 'language' in such a story? There are a number of reasons to believe we cannot. First, to be sure, moral judgment is a complex skill, but does it really have to be the case that there are *principles*, conscious or subconscious, underlying every complex, intelligent performance? Take a game of chess. Surely it is possible that given a set of chess problems, a skilful player will be able to come up with effective solutions without being able to articulate any principles guiding his choices. It may also be possible for a researcher to come up with principles that match the choices at hand. But does it follow that the same principles, or any principles, must have guided the chess player's original choices? As it happens, this is hotly disputed in the literature on skills. In classic work on skill acquisition, Hubert Dreyfus has long maintained that rules and principles play a role primarily at the first, 'novice' level, when one has to rely on cues accessible to non-experts.[161] As one's expertise develops, there is less and less reliance on rules, whose place is taken by refined perception and emotional and even bodily reactions. One could object that there must still be rules at an unconscious level, perhaps a computational one. Dreyfus can point to the failures and limitations of rule-based artificial intelligence as evidence against this.[162] Connectionists in the philosophy of mind have independent reasons for the same conclusion. For connectionists, the mind is a complex network of neural networks whose inputs are not connected

[160] Compare Cosmides and Tooby 2006, 186.

[161] See, for example, Dreyfus 1990 and Dreyfus 1992. For application of this kind of views to ethics, see also Dancy 1999.

[162] See Dreyfus 1992.

to outputs by way of any sort of computational rules. From this perspective, Paul Churchland argues that the alternative to a rule-based account of moral capacity is "a hierarchy of learned prototypes, for both moral perception and moral behavior, prototypes embodied in the well-tuned configuration of a neural network's synaptic weights."[163] The principlist assumption is thus very much open to question, and hangs in part on the debate between computationalist and connectionist theories of the mind.[164]

Second, does moral judgment really develop spontaneously, without conscious instruction by parents and other authorities, that is, is there a poverty of moral stimulus?[165] This is important for the innateness assumption of the grammarians. There seems to be a clear difference between the sorts of instruction involved in teaching language and teaching ethics. For example, children are punished for moral violations, but only occasionally admonished for linguistic errors.[166] Moreover, the punishment seems to be qualitatively different from punishment for conventional violations, which potentially allows the child to come to make the moral/conventional distinction on the basis of experience.[167] And of course, for both language and morality, imitation accounts for much of the learning. It thus seems like an empirically open question whether and what sort of moral capacities would have to be innate. A moral grammar, even if there was one, could perhaps be learned. Third, and related,

---

[163] Churchland 1996, 101. Compare Clark 1996.

[164] In addition, some of the complex principles and transformations that the grammarians postulate as the operations of the moral module, like perception of certain movements as actions and analyzing the consequences of action into intended results and side effects, also serve other needs like social coordination and planning. They are thus not specifically *moral* skills, and could have developed or been learned independently.

[165] Much of the following is based on unpublished and forthcoming work by Jesse Prinz.

[166] (According to Prinz) Hoffman 2000 estimates that the behaviour of children between the ages of 2 and 10 is corrected every 6 to 9 minutes by caregivers.

[167] See Smetana 1989, Nucci and Weber 1995, and Prinz (forthcoming).

in the face of moral diversity, the idea that there would be a *universal* moral grammar requires putting a lot of weight on the distinction between principles and parameters: just like some languages indicate location with a suffix and some with a preposition, some moralities set the 'killing permitted' switch to 'any out-group members' and others to 'convicted murderers'.[168] This is a possible way of thinking of moral differences, to be sure. But it easily trivializes the 'principle' involved. In the example, all it amounts to is that there needs to be some regulation of whose killing is permissible. That indeed seems like a universal truth, but it hardly takes an innate module to figure that much out. And in areas in which there is cross-cultural convergence, there are also competing explanations, for example in terms of common needs, emotions, and problem situations.

Finally, we *are* conscious of our moral principles (that is, "aware of their underlying logic") to a much larger extent than of our grammatical principles.[169] This opens up the possibility of using general ('system 1') reasoning capacities to make moral decisions, and also calls into question the modular nature of the process. Two well-known tests for modularity are the effects of 'cognitive load' and the existence of selective deficits.[170] Adding cognitive load by making test subjects engage in some pointless but resource-demanding activity slows down tasks that require conscious reasoning but does not interfere with automatic, modular processes like face recognition. Joshua Greene (forthcoming a) reports that some preliminary studies suggest that while moral judgments in

---

[168] See Hauser 2006, 71–75.

[169] Hauser does assert that "having conscious access to some of the principles underlying our moral perception may have as little impact on our moral behavior as knowing the principles of language has on our speaking" (Hauser 2006, 67), but provides no evidence. This is trivially true if access to principles does not lead to reflective adjustment, like the adjustment that people make when they decide to become vegetarian for moral reasons. If people do come to reject principles they tacitly held, the claim that consciously adopting new principles has no impact becomes far less trivial indeed; witness the vegetarians around us.

[170] See Prinz 2006b.

personal scenarios are not slowed down by adding cognitive load, impersonal judgments are. This is bad news for moral modularists. It suggests that general rather than modular reasoning goes on in impersonal cases, and in personal scenarios, the judgments are plausibly triggered by affective reactions rather than modular analysis. However, it is important to bear in mind that this does not yet mean a victory for affectivists. Automatic processes (the sort of processes that are relatively undisturbed by cognitive load) can be *learned* and *intelligent*, rather than the sort of evolutionarily primitive affective reactions that Greene takes them to be. Dreyfus's work on skills provides a clear example:

> We recently performed an experiment in which an international [chess] master, Julio Kaplan, was required rapidly to add numbers presented to him audibly at the rate of about one number per second, while at the same time playing five-second-a-move chess against a slightly weaker, but master level, player. Even with his analytical mind completely occupied by adding numbers, Kaplan more than held his own against the master in a series of games. (Dreyfus 1990)

Chess skills, surely, are not primitive or affective or modular, though they evidently withstand cognitive load. As to selective deficits, they do exist for modular processes like face recognition, but that does not seem to be the case for morality – psychopaths, for example, suffer from a variety of problems, centrally with respect to social emotions, as we have seen, not deficits *just* in moral judgment.[171]

In brief, then, the analogy between moral and linguistic judgment does not seem very close, and hardly warrants postulating a specialized moral module, as long as there are alternative and cheaper explanations of the data.

---

[171] Prinz 2006b points out that selective deficits are closely related to another indication of modularity, anatomical localization, since the deficits are caused by damage to certain areas of the brain. Brain-imaging studies such a Greene et al. 2001 have not found a specific moral region of the brain.

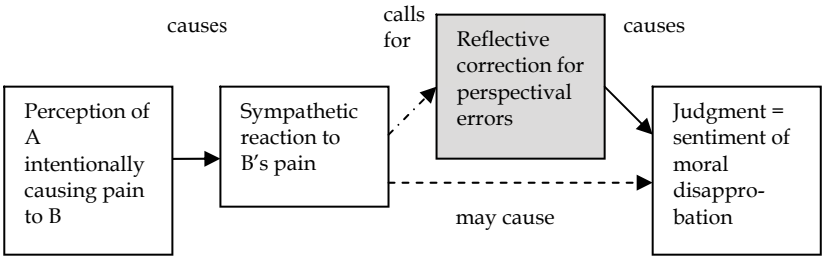
It appears that the leading contemporary psychological models do not fare well on reasonable desiderata for an account of moral judgment in the process sense. What, then of philosophical views? As the discussion in section 2.2 should have shown, each of them passes most of the desiderata with flying colours. However, some of the experimental results present a challenge. The data points that must be accommodated include the following:

- people are not, as a rule, able to articulate principles that would explain their judgments in complex dilemma cases
- people do, however, come up with confabulations when asked to explain their judgments
- emotional reactions and mood influence moral judgments, at least third-personal judgments about hypothetical cases
- people lacking in affect, such as psychopaths, make abnormal moral judgments
- even small children distinguish between moral and conventional rules
- there is considerable cross-cultural variety in the content of moral judgments

The data for these claims have been collected in surveys and structured interviews, both familiar social scientific methods. As in similar studies on other issues, there are concerns of ecological validity, that is, of how well the data correspond to people's thought and behaviour in everyday contexts. What seems especially pressing in the case of moral judgment is the difference between important first-personal moral decisions and quick evaluations of hypothetical cases. The studies measure the latter, while the philosophical interest – quite naturally, given the practical purport of ethical theory – has focused on the former. It seems very likely that conscious reasoning and argument play a more important role in first-personal judgments and in evaluations of others that potentially call for action

(such as punishment or voting) on one's own part. The existing data should thus be taken with a grain of salt.

It may in any case be worth it taking a quick look at how the philosophical accounts handle the data. With some inevitable simplification, they can also be summed up with diagrams. Let us begin with a Humean account of a simple case in which one person is morally disapproved for hurting another:



The Humean story, as I am construing it here, begins with beliefs about other people's mental states (what is sometimes called mind-reading or cognitive sympathy), which give rise to a sympathetic affective reaction. This affective reaction gives rise to a sentiment of moral disapprobation either directly or after being corrected for various perspectival errors like the effects of self-interest, mood, and distance. Having internalized the reactions of others provides at least part of the motivation for this corrective move. It is this second alternative that makes intelligible the functional and phenomenological role that moral disapprobation has, including its role in social coordination and feelings like disapproval of those who fail to disapprove of the wrongdoer. Being fallible and imperfectly sympathetic, we often fail to take the general point of view, and rush into judgment – "the passions do not always follow our corrections", Hume notes.[172] But we can understand the role that even those

[172] *Treatise* 3.3.1.

rushed judgments play by reference to the ones that result from corrected sentiments – the uncorrected judgments are *parasitic* on the corrected ones in this sense.

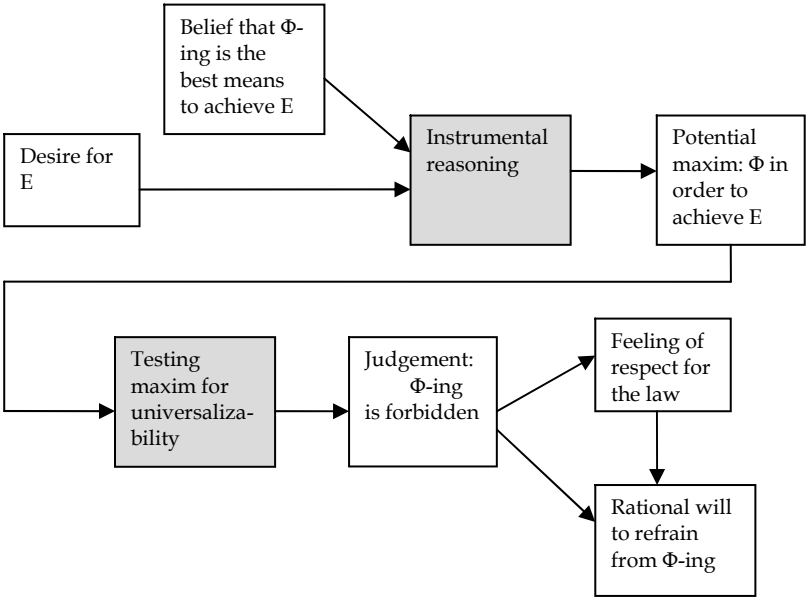
How does the Humean view square with the empirical data generated by recent studies? To begin with, it can easily accommodate all the data that support affectivist views – he explicitly cites similar observations in support of his own view. His view also predicts that various sorts of reflective correction are less likely in cases where nothing much is at stake for the agent, such as the ones the studies probe. It is also harder when immediate affective reactions are more vivid. Something like this might be at play in the trolley studies. In Fat Man, Hume, as a proto-utilitarian, might say that reflective correction is not successful – sympathy for the man one would have to push is more vivid than sympathy for the five who will die, which is why many choose not to push. The minority who allocate sympathy more impartially are willing to sacrifice the one. Secondly, as to the data concerning the inaccessibility of principles, while Hume talks about "the durable principles of the mind"[173], he clearly does not conceive of them as discursive, but rather as steady dispositions to respond to situation-types with sentiments of approbation or disapprobation.[174] It is thus predictable that people would sometimes fail to be able to articulate their content, and might confabulate if asked to state them. Finally, with respect to the moral/conventional distinction and its cultural variability, Hume's empiricism about learning is tempered by his assumption of universal human capacities and dispositions. Given our psychological make-up and need for social coordination, it is predictable that we would come to moralize certain things, particularly those that give rise to sympathetic reactions in normal human beings. At the same time, the importance of habit and the influence of the opinions of close associates serve to explain why different cultures end up moralizing different things, within the limits of our natural tendencies. After all, for him some virtues are natural and others artificial.

---

[173] *Treatise* 3.3.1.

[174] Compare Blackburn 1998, chap. 1.

As to Kant, his focus is explicitly on first-personal moral decision-making. Here is a simple picture of how he seems to conceive the process of moral judgment in a case in which what is desired fails to pass the Categorical Imperative test:



On this construal, the process of practical reasoning begins with a desire for an end, say increasing profits. When this is combined with a means-end belief in instrumental reasoning, the result is a potential maxim for action, for example cheating a customer when possible to do so without getting caught in order to increase profits.[175] If the agent is rational she will test the maxim for whether it could be a universal law, and if it is not, arrives at the judgment that the proposed action is morally proscribed. This gives rise to a feeling of

[175] This may be a controversial reading of the role of instrumental reasoning in formulating maxims of action.

respect for the majesty of the law – "[W]hat in our own judgment infringes upon our self-conceit humiliates. Hence the moral law inevitably humbles every human being when he compares with it the sensible propensity of his nature"[176] – and sets the will against the course of action in question, here cheating a customer. It is not entirely clear what Kant takes the role of respect for the law to be in moral motivation, so I leave room for both direct determination of the will by the conclusion of practical reason and indirect determination via respect.[177]

Of all the historical philosophical views discussed here, Kant's has been the hardest hit by many of the empirical studies, given the extraordinary importance he seems to place on explicit impartial reasoning in arriving at moral judgments. How could he possibly account for the fact that, to put it mildly, we do not always engage in the sort of testing involved in Categorical Imperative when we make moral judgments? In fact, most of us never do so! Perhaps surprisingly, Kantians have a number of alternative strategies available. The considerations I mentioned earlier regarding the difference between first-personal and third-personal judgments can function as softening up arguments in favour of a plausibly larger role of explicit reasoning in the case of the former. Kant certainly does not deny that pleasures and affects make a difference to our *likings*, quite the contrary.

But in fact, Kant may not need to defend the role of explicit reasoning in the first place. After first formulating the Categorical Imperative in the *Groundwork*, he says that common sense "agrees completely with this in its practical appraisals and always has this principle before its eyes"[178]. Common sense (*gemeine*

---

[176] Kant 1788/1996, 200.

[177] In some places, Kant's suggestion seems to be that it merely reduces the influence of competing desires; it "lessens the obstacle to pure practical reason and produces the conception of the superiority of its objective law to the impulses of the sensibility; and thus, by removing the counterpoise, it gives relatively greater weight to the law in the judgement of reason (in the case of a will affected by the aforesaid impulses)" (Kant 1788/1898, 168).

[178] *Groundwork* I, 57.

Menschenvernunft) does not need to be taught anything new to distinguish good and evil – the philosopher merely needs, "as did Socrates, make it attentive to its own principle"[179]. Kant's explanatory strategy is thus closer to the moral grammarians – deep, underlying moral principles can be reconstructed from our particular judgments. (This is not to say that the moral law is at the mercy of whatever judgments we happen to make – Kant believes that the laws derived *a priori* from the concept of duty itself coincide with those implicit in our judgments, if not in our behaviour.) In fact, he is explicit that we can *never* know for certain what the principles of our actions are: "[N]o certain example can be cited of the disposition to act from pure duty; that, though much may be done *in conformity with* what *duty* commands, still it is always doubtful whether it was really done *from duty* and therefore has moral worth."[180] Since the moral worth of our actions depends on "those inner principles of actions that one does not see"[181], our judgments about ourselves and others must remain uncertain. Kant acknowledges our tendency to rationalize judgments that are in fact driven by pleasure and affect – "we like to flatter ourselves by falsely attributing to ourselves a nobler motive"[182]. That is just why moral philosophy and explicit reflection are needed – "[i]nnocence is indeed a glorious thing, but it is very sad that it doesn't take care of itself, and is easily led astray."[183]

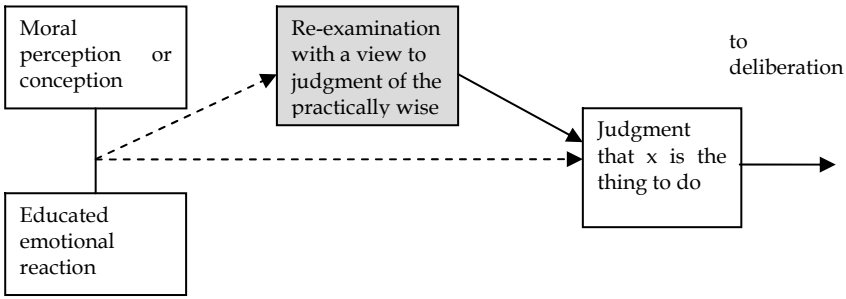
There is no basis, therefore, for the claim that Kant denies the frequent influence of affects on our moral judgments or thinks that we must always be conscious of the process of reasoning that leads to judgment, even in the lucky cases in which the reasoning conforms to the moral law. To show that he is mistaken one would need to make the case that it is *impossible* for us to reason consciously in the way he recommends, or that it is not *necessary* to do so to understand and perhaps vindicate the authority of morality. (The

[179] *Groundwork* I, 58.
[180] *Groundwork* II, 61.
[181] *Groundwork* II, 62.
[182] *Groundwork* II, 61.
[183] Kant 1785/1898.

latter, obviously, is the project of competing philosophical accounts, like those of Hume and Aristotle.) One could perhaps try to argue for this indirectly by observing the anthropological variety of moral systems – if people did reason, consciously or not, according to the Categorical Imperative, we would not expect to see such a variety. But the problem for this kind of arguments is to show that there indeed exists such variety among rational agents who make genuine (even if sometimes mistaken) moral judgments after we have eliminated both differences in factual background beliefs (such as beliefs about the afterlife) and derived moral judgments that make sense only in light of a particular cultural context (such as the judgment that it is wrong to show up drunk at a funeral, which is derived from a more general and putatively universal *pro tanto* norm of not insulting people's feelings). Here Kant could turn the rarity and relative abstractness of the duties derivable by using the Categorical Imperative test to his advantage. Suppose that the list of duties pure practical reason generates looks much like W. D. Ross's list of seven *prima facie* duties – fidelity, reparation, gratitude, beneficence, non-maleficence, justice, and self-improvement.[184] It is far from obvious that results like those of Shweder and associates show that there is disagreement at this level, even though what constitutes fidelity or how gratitude is manifested varies from culture to culture. If so, the observed cultural variation in concrete moral judgments is compatible with the judgments resulting from implicit or explicit use of pure practical reason.

[184] Ross 1930, 21–22. Robert Audi (2004) argues that a Rossian list of duties can indeed be inferred from the intrinsic-end formulation of the Categorical Imperative ("always treat another person as an end in itself, never as a mere means").

For Aristotle, finally, much of the work of moral judgment has already been done when there is a need to react to a particular case. Nevertheless, we could probably sketch a diagram something like the following for decent but not perfectly virtuous agents:



On the Aristotelian picture, we perceive (or conceive, if contemplating a possible action) the world in thick value terms – actions strike us as courageous, silly, bold, cheap, noble, and so on. These qualities provide *pro tanto* reasons toward an overall verdict about the action and the agent. Corresponding to this perception there is an emotional reaction, informed by our experience and education, and shaped by our moral views. For Aristotle, emotions make a reference to the external world not just by way of having a target (such as the thing that one is afraid of) but also having what is now called a formal object (such as being dangerous), so they are subject to assessment in terms of appropriateness.[185] As I picture it here, the perception or emotional reaction may thus need to be re-examined, if there is reason to doubt it, if there appear to be conflicting reasons, or if there is disagreement. This may be unnecessary for the perfectly virtuous person, who directly sees what the right thing to do is, and rashly bypassed in the case of a person lacking in virtue. But for the ordinary, imperfectly virtuous but decent people, this seems the place where the goal of a mean

[185] See Helm 2001 for this distinction and its importance.

between excess and deficiency and the related reference to the practically wise agent can play a role. Aristotle, like Hume and Kant, recognizes that it is "hard work to be excellent"[186]. Everyone is biased in the direction of their own pleasure, and each person has her own natural tendencies toward one excess or another. Almost as if anticipating Kant's metaphor of the 'crooked timber of humanity' (out of which, according to Kant, no straight thing has ever been made), Aristotle suggests that faced with a distorting bias, "[w]e must drag ourselves off in the contrary direction; for if we pull far away from error, as they do in straightening bent wood, we shall reach the intermediate condition."[187] The correct reaction is "defined by reference to reason, that is to say, to the reason by reference to which the prudent person (*phronimos*) would define it."[188] This passage is sometimes read as suggesting that the virtuous person's judgments *make* an alternative correct. But this goes against the general thrust of Aristotle's view that the right reason is determined by what contributes to *eudaimonia*. So it may be better read as a deliberative suggestion – looking up to a *phronimos*, preferably a concrete one, and trying to figure out how she would react to the situation may help to determine what virtue and reason really require in the situation.[189] As in the case of Hume and Kant, this more reflective route to judgment – as well as the fact that the emotions themselves are not brute affective reactions but already incorporate responsiveness to reasons – makes sense of our investing the resulting judgments with authority and taking them to be justifiable, at least to anyone who has been brought up well enough.

---

[186] *NE* 1109a, 29.

[187] *NE* 1109b, 29.

[188] *NE* 1107a. For a defense of the doctrine of the mean, see Hursthouse 2006.

[189] Compare Epictetus: "When you are going to meet with any person, and particularly one of those who are considered to be in a superior condition, place before yourself what Socrates or Zeno would have done in such circumstances, and you will have no difficulty in making a proper use of the occasion." (*Manual*, XXXIII)

Since Aristotle does not require conscious reasoning to play a central role in arriving at moral judgment, his view is relatively easy to fit with the empirical data. First, if judgment issues from moral perception or educated emotions, the reasons that support the judgment will not be immediately or infallibly available to the agent. There may be no principle to refer to; perhaps the only thing to say if asked to justify the judgment is "Don't you see?".[190] On this reading, it may be a mistake to demand verbal explanations, and no surprise if people's access to their actual reasons is limited. Even if we *could* engage in, say, explicit dialectical reasoning about the good life and deliberation about the best means to it, it would be a waste of expensive resources – not to mention time – to do so, whenever we can delegate the job to emotions and perception. Second, that emotions affect judgments is part of the theory, and as we have seen, Aristotle has resources to distinguish when this serves judgment well and when badly. His theory does not fare particularly well, it must said, on the psychopath issue, since what they are lacking is social emotions in particular, and Aristotle has little to say about them. It is not hard to see, however, how the theory should be amended to handle these cases, since the general point that one must have the right sort of emotional reactions to make moral judgments (particularly in thick terms) is at the heart of the theory. Third, Aristotle does distinguish between conventional and non-conventional norms in his discussion of justice[191], but does not address the issue of how we come to make such a distinction. Again, it is not hard to see how the theory could be amended, either in terms of different natural tendencies or different sort of feedback from caretakers (that is, different sort of normative education), or both. Finally, since Aristotle acknowledges, particularly in the *Politics*, that the good life can take different forms depending on social, historical, and natural circumstances, and correspondingly require different virtues, his view predicts that there will be widespread cultural differences, mitigated by universal features of human nature.

---

[190] See McDowell 1979/1998, 63.

[191] *NE* Bk V, sec 7.

This introduction is not a place to argue for any particular view about moral judgment in the process sense. But I do hope to have shown that in contrast to inflated claims made by psychologists and empirically minded philosophers, the various metaethical traditions provide philosophical explanations of the phenomena that remain serious contenders both as descriptive and normative accounts of moral thinking.

Essay 2: 'Moral Judgment and Volitional Incapacity'

In 'Moral Judgment and Volitional Incapacity', I develop a new criticism of expressivist theories of moral judgment in the product sense and outline a cognitivist and rationalist theory that makes sense of both success and failure of moral motivation. Expressivism is the contemporary heir of emotivism and non-cognitivism. It is, in the first instance, a thesis about what it is to think moral thoughts. The distinctive mark of expressivism is that according to it, thinking that something is wrong or someone is admirable does not involve ascribing moral properties, but rather amounts to an attitude toward an object on the basis of its natural properties. Moral utterances – sometimes also confusingly called 'moral judgments' – get their sense from the psychological states they express, and are thus not in the business of describing the world. They can be true only in a deflationary sense, in which calling something true is just endorsing it. If that is the case, to say that it is *true* that murder is wrong is to say nothing over and above saying that murder is wrong.

Why should we believe in expressivism? There are two basic motivations for it. One is that it gives a legitimate role for moral thought and discourse – expressing and coordinating attitudes and actions – that does not involve a commitment to moral facts, which would be 'queer' additions to a naturalistic worldview. The other benefit of expressivism is that it is thought to explain the special motivational role of moral judgments. The metaethical thesis known as *moral judgment internalism* says that thinking that something is my duty (roughly speaking) necessarily motivates me to do it. It comes in many varieties, from very strong (we always do what we think is

morally required) to very weak (if we never have any motivation to do what we think is morally required, we are not really thinking moral thoughts). In any case, internalism presents a challenge to cognitivist theories of moral judgment: why would a mere belief have such an intimate, non-contingent connection to motivation? Expressivists, by contrast, have an easy answer: since moral judgment itself consists in an attitude, such as a complex higher-order desire, emotion, or a planning state, it is not surprising that it gives rise to motivation. While cognitivists may have the edge when it comes to accounting for the fact-like surface features of ethical discourse, expressivists have an advantage when it comes to explaining moral motivation.

This is the picture that I challenge in Essay 2. I start with the observation that not only success but also failures of moral motivation must be made intelligible by an account of moral judgment. To be sure, sophisticated expressivists like Allan Gibbard (1990, 2003), my main target, can account for some failures of motivation, such as ordinary weakness of will. But there seem to be other cases, uncovered by recent philosophy of action, that force us to reject any view that identifies moral judgment with will or decision or planning state, as Gibbard explicitly does. A central class of these cases have been labelled those of *volitional incapacity* by Gary Watson (2003). Volitionally incapacitated agents do not suffer from weakness of will, which entails that there are competing motivational forces, such as the intention to act (the will) and a competing desire, within the agent. Rather, they are unable to take the first step from a first-personal ought-judgment to forming the corresponding intention. They do not, thus, suffer from inner motivational conflict, but a conflict between their judgment and the will (or motivational states in general). If that is the case, their judgment cannot be identified with any volitional or conative state. That means that ought-judgments must be some kind of cognitive states.

Does this mean that we must also reject moral judgment internalism, which is, after all, an independently plausible thesis? Not necessarily. As Michael Smith (1994) has argued, one can be both cognitivist and internalist, provided that it can be shown that

there is a necessary connection between moral belief and motivation in *rational agents*, by virtue of the *content* of the moral belief. I develop an alternative to Smith's own story of the content of moral beliefs on the basis of Robert Brandom's (1994, 2000) *inferentialist* conceptual role semantics. According to inferentialism, the content of a concept is given by the circumstances and consequences of its application. More precisely, the question to ask is under what circumstances is one committed and entitled to apply the concept, and what consequences applying the concept has to one's score of commitments and entitlements. Following and modifying Brandom's account, I argue that ought-talk makes explicit inferential commitments. Crudely put, if I say that I ought to milk the cow, I am anyone in my situation would be entitled to milk the cow (which explains the *rationality* of setting out to milk the cow) and that I am not entitled not to milk the cow, which amounts to saying that anyone is entitled to negatively sanction me for failing to milk the cow (which explains the *bindingness* of the ought-judgment). Since the inferential commitment constituting the ought-judgment consists is rationally binding, a rational agent – understood simply as an agent who is capable of responding to acknowledged commitments to which one takes oneself to be entitled – will form the will (intention or plan) corresponding to the judgment, and act accordingly, unless suffering from weakness of will. Volitional incapacity and weakness of will, on this picture, are disruptions of two different rational-causal mechanisms.

3 The Psychology of Moral Responsibility

Moving from narrowly metaethical questions to broadly metaethical ones, the focus of the inquiry shifts from the person as the *subject* of moral evaluation to the person as the *object* of moral evaluation. What does it take for someone to be a fit target of reactive attitudes like praise, blame, gratitude, and resentment? This question brings us to the debates about free will and autonomy, since the practical and philosophical interest in them derives largely from their role in making it fair to praise and blame people – it does not make much sense to criticize or reward people for something that was not up to them, if they were not in some significant sense in control of what they did. As it turns out, there are many kinds of control that may be relevant. There are also at least three senses of responsibility that must be distinguished.[192] One is *causal responsibility*, which is clearly distinct from any moral notions – a worn-out timing belt may be causally responsible for the engine stalling, for example. Another sense is closely related to *self-disclosure*. Actions that we are responsible for in this sense are deeply attributable to us. They reveal our character or who we really are. A third sense of responsibility is often called *accountability*. We can be held to account, praised or blamed, for actions and attitudes for which we are responsible in this sense.

[192] See especially Watson 1996, but also Scanlon 1998 and Wolf 1990.

3.1 The Metaphysics of Free Will

The first questions in this area are metaphysical. The traditional divide is between *compatibilists* and *incompatibilists*. Compatibilists believe that free will and moral responsibility are compatible with being determined by the past and the laws of nature, provided that this determination is of the right kind. Incompatibilists deny this. In recent decades, this crude division between metaphysical options has been considerably refined. On the compatibilist side, the core thesis of *classical compatibilism* is that determinism is compatible with the ability to have done otherwise, which is taken to be essential to free will and moral responsibility. The simplest version of the argument goes something like this: If the causal chain that leads to action runs through the agent's own beliefs and desires, then had those beliefs and desires been different, the agent would have acted otherwise, and this is all we mean by 'free will'. More sophisticated arguments along this line are offered by Dennett (1984, 2003), Smith (1998), and Pettit (2001). The other main current of compatibilism is *semi-compatibilism*, which holds that moral responsibility does not require the power to do otherwise in the first place, so determinism does not threaten it. Semi-compatibilists typically appeal to so-called 'Frankfurt-style cases' (originating in Frankfurt 1969), in which a 'counterfactual intervener' would have made the agent to do what she did, had she not chosen to do so on her own; in such a case, it is not true that the agent could have done otherwise, but intuitively, she is still responsible. In John Martin Fischer's terms, the agent lacks *regulative control* over which course of events comes about, but she can still have *guidance control*. What matters is the nature of the actual causal sequence that leads to action, not what might have happened. The most sophisticated defence of this kind of view is surely Fischer and Ravizza (1998) (see below).

On the incompatibilist side, the shared starting point is the rejection of compatibilism. The simplest incompatibilist argument, directed against classical compatibilism, is that since determinism entails that there is at any point only one possible future, there are no

alternative possibilities among which we could choose from, so determinism is incompatible with freedom. A different argument that does not rely on alternative possibilities, and may thus work against semi-compatibilism which does not require them, starts from the idea that to have sufficient control of our actions, we must be their ultimate *source*.[193] Relying on some version of what has come to be called 'the principle of transfer of non-responsibility', the argument claims that we cannot be sources of our actions if determinism is true: we are not responsible for the distant past or laws of nature, so if they determine our character, beliefs, and desires, we are not responsible for them either, and consequently not responsible for our actions, even if they result from our own beliefs and desires. Source incompatibilism thus sets the bar even higher than simple incompatibilism: even if we did have alternative possibilities, it would not suffice for free will, unless we were also the ultimate sources of our actions.

Incompatibilists come in optimistic and pessimistic varieties, depending on whether they take us and the actual world to meet the demanding conditions they set on free will and responsibility. Those on the optimistic side of incompatibilism are called *libertarians*. The simplest and most baffling form of it is *agent-causal libertarianism*, according to which there is a special form of causation, agent-causation, which is exempt from the laws of nature, and yet can make a difference to what happens. Though even the philosopher who introduced it to contemporary discussion, Roderick Chisholm, gave up on the idea as incoherent (see Chisholm 1995), it still has defenders. *Event-causal libertarianism* makes do with ordinary sort of causation, but argues that suitably situated random deviations from deterministic laws are necessary for free will. A major challenge for views of this kind has always been explaining why mere chance or luck (which is what a random deviation in fact is) would not rather *undermine* freedom and responsibility. Robert Kane's views are perhaps the most sophisticated response to this problem (see below).

[193] This kind of incompatibilism was labeled 'source incompatibilism' by Michael McKenna (2000).

Not all incompatibilists are optimistic about the existence of just the right kind of indeterminacy in the world. What are traditionally called *hard determinists* believe that both incompatibilism and determinism are true, so that we are not free or morally responsible. Lately, an even more pessimistic view has gained favour: philosophers like Galen Strawson (1994, 2002) and Derk Pereboom (2001) believe that we are not responsible whether determinism is true or not. For these *hard incompatibilists*, as they are sometimes called, the very notion of ultimate responsibility (which is the sort of responsibility they think we need) is incoherent. They draw on a regress argument, according to which to be responsible for our actions we must be responsible for our characters from which they issue, but to be responsible for our characters we must be responsible for creating them, and that either passes the buck back to responsibility for actions (as Aristotle thinks), leading to the same step again, or leads to infinite regress. No one can be *causa sui*, her own cause, since to cause anything, one must already exist.

3.2 The A Priori Psychological Conditions of Moral Responsibility

So much for a very brief overview of the metaphysical options on the table today. What is important for our purposes is that both compatibilists and incompatibilists are quickly led to questions in philosophical moral psychology. The sort of causation or chance that optimists on each side think is necessary and sufficient for responsibility takes place in the mind. For both sides, the key question is the following: what must the psychological process leading to action be like for the agent to be fully morally responsible for the action? An answer to this question will constitute at least part of the answer to another, related question: what does it take for an agent to be autonomous? These are paradigmatic *a priori* questions – it is hard to imagine that anyone would take answers to them to be found in the world.

On the compatibilist side, the best-known answers appeal to the notion of *desire ownership*. The basic idea is that a person is responsible for an action (and autonomous with respect to it) if the desire that leads to it is genuinely the agent's *own*, part of her *real self*, not something that is in some sense *alien* to the agent. Different theories draw the line between alienated and non-alienated desires in different ways. Harry Frankfurt (1971) popularized *structural* views, according to which desire ownership is simply a matter of how the desire in question fits with the total structure of the agent's psychological states. On Frankfurt's own version, a desire is truly an agent's own when the agent desires to have that desire and so in this sense reflectively endorses it. A desire that one wishes not to have, by contrast, is alien – the unwilling addict struggling in vain against his urge to inject drugs is a paradigmatic example of this sort of alienation, which is clearly a case of lacking freedom of the will. Frankfurt's view is compatibilist, since for him the origin of the desires and second-order desires is simply irrelevant to whether they are the agent's own. This has spawned several challenges. According to a different structuralist view, associated with Gary Watson, the problem with Frankfurt's version of reflective endorsement is that higher-order desires as such lack authority – why should they get to delineate what really constitutes the agent's own point of view?

Valuing and Planning

Drawing inspiration from Plato, Watson suggests that it is the agent's *evaluative system* – her beliefs about what is good and worth pursuing – that has the necessary authority. On this view, it is evaluative beliefs that constitute the true self, and alienation is a matter of having desires that conflict with these beliefs. Again, the unwilling addict thinks it is *bad* for him to use drugs, but cannot help himself. This is why his responsibility is reduced. Thorny issues arise here, however, for we do want to blame those weak-willed agents who we think could have resisted the temptation that they fell for, such as the typical guilty adulterer, who thinks she should not cheat but goes on anyway. Compatibilists must find a way to distinguish

between compulsion and criticisable failure of self-control, which seems to require a making sense of ability to do otherwise in a deterministic world – the natural way to put the difference is that the unwilling addict could do no other, while the adulterer could. Perhaps the best effort to date is that of Michael Smith, and it may be worth taking a little time to take a closer look at it. He begins a cognitive analogy. Suppose a philosopher, John, fails to think of a clever response in a conversation. Now take two scenarios. In the first one, John goes home, reads some papers on the issue, and realizes what he should have said; in the second, the right answer comes to his mind as he is on the way home on the basis of what he already knew and had thought. It is very natural to say that while the first John could not have thought of the right answer on the spot (because he had to read up later), the second one could have (because he had all the necessary information; it was just a fluke that he blanked during the conversation). Yet, as Smith points out, if the world is deterministic and we take 'could have' to mean 'could have even if the past and the laws of nature had been identical', we lose the distinction: on this understanding of ability to do otherwise, it is equally impossible for Blanking John to have thought of the response as it is for Ignorant John.[194] Clearly something is wrong, since our theory should capture the difference in their capacities.

To make sense of this, we should think of 'could have' as David Lewis suggests: Blanking John could have thought of the right response, because he (or his counterpart[195]) would have thought of it in a possible world whose history and/or laws diverge minimally from ours, whereas Ignorant John could not, since the possible world

---

[194] Smith 2004, 117–118.

[195] There is no need to go any deeper into the debate on the nature of possible worlds here, but on the Lewisian view, individuals are world-bound – we inhabit exactly one possible world, namely the actual one. The truthmakers for our possible doings are the doings of our counterparts in other possible worlds. See Lewis 1971 and 1986. Those who take possible worlds talk less literally than Lewis are sometimes willing to talk about identity across possibilities; see for example Kripke 1980. For simplicity, I will ignore these distinctions here.

in which he would have thought of it is much more remote.[196] To complicate matters, Smith notes that this does not yet suffice to rule out the case of a fluke – even if John was a very bad philosopher, it could have happened by chance that a fitting response occurred to him. Moreover, the possible world in which Fluky John thinks of the response could be equally close to ours as that in which Blanking John does. That's why in order to get at a *capacity* we need to look at a pattern, a whole raft of counterfactuals: Blanking John's counterparts come up with a response in a host of nearby possible worlds where similar questions are asked, while Fluky John's counterparts do not.[197] Smith argues persuasively that the same model applies to capacities relevant to self-control. Though both the addict and the adulterer act against their evaluative judgment in the actual world, different counterfactuals are true of them. In a host of nearby possible worlds, the adulterer refrains from cheating – we only need to tweak her incentives or imagination a little, and she resists the temptation. That is what the compatibilist of this type means when he or she says that the weak-willed agent has the capacity to control herself. The addict, in contrast, lacks the capacity and thereby full responsibility (leaving aside issues about responsibility for becoming an addict): the worlds in which he refrains from shooting up are very distant from ours, and her counterparts in those have very different beliefs and dispositions. In this way the compatibilist can make at least many of the crucial distinctions needed to make sense of our everyday attributions of responsibility.

Even if evaluativist compatibilists clear this hurdle, they face further challenges. A rather obvious one is that there seem to be attitudes that we think it is wrong for us to have, but which are nonetheless intuitively our own, as Watson himself acknowledges.[198] We seem to be able to imagine a case in which a man falls in love

---

[196] This is actually only the first pass for Smith due to complications arising from 'finkish' dispositions; see Smith 2004, 120–122, Johnston 1993, and Lewis 1997.

[197] Smith 2004, 123–125.

[198] Watson 1987, 150.

with a woman he is sure will betray him and reproaches himself, but will not give up the love, which has come to define who he is. Further, it seems possible that we can commit to one goal while regarding another as equally good or incommensurable with it, in which case our evaluative judgments do not suffice to determine where we stand.[199] And third, at least some evaluative judgments involve an expectation of convergence among rational or reasonable agents (as discussed in sections 2.2 and 2.4), and one might well not have such an expectation concerning some of one's commitments.[200]

Searching for an alternative that both preserves the notion of agential authority missing in second-order desire accounts and the possibility of identifying with what one does not regard as best, Michael Bratman suggests that desires of one's own are those that are endorsed by one's 'self-governing policies' as reason-giving in practical deliberation. This needs some unpacking. Bratman has long emphasized the central role of planning in human agency: we do not just act moment by moment, but organize and structure our goals in advance.[201] Much of our practical reasoning takes place against the background of one long-term plan or another, concretizing the goal or specifying the means. Some plans are policies: they concern what to do in a recurring type of situation, like whether to wear a seat belt while driving.[202] Some policies concern practical reasoning: I might adopt a policy to give more weight to my mother's needs in the future when making decisions, or a higher-order policy to ignore feelings of jealousy that arise when my girlfriend takes a tango class. Bratman labels policies to treat some considerations as "reason-providing in motivationally effective deliberation" *self-governing policies*.[203] Now, plans and policies introduce an element of cross-temporal stability into an agent's life, and Bratman sees this as the key to understanding the sort of psychological connections that constitute personal identity over time in a Lockean conception. We

---

[199] Bratman 2005/2007, 205.

[200] See Bratman 2004/2007, 235–238.

[201] For the original statement, see Bratman 1987.

[202] Bratman 2000/2007, 27.

[203] Bratman 2007, *passim*.

are unified as persons because our beliefs, desires, intentions, and emotions at any given time refer to past and future psychological states and are shaped by a disposition to maintain a sort of coherence over time. The complex planning states that constitute self-governing policies thus play an important role in defining who the agent is. For Bratman, this is how they earn the right to decide which desires are really the agent's own; as he puts it, "[t]hese attitudes have agential authority *at* a time in virtue of their roles in constituting and supporting the interwoven, interlocking structures of agency of that very person *over* time"[204].

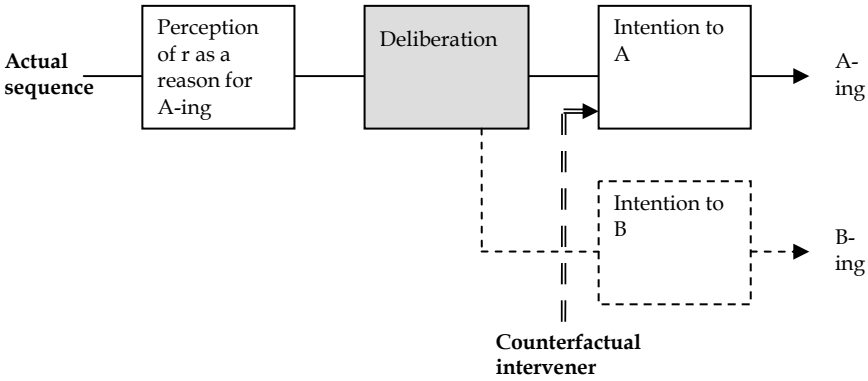
History and Reason

Evaluativist and planning views seem to help with the problem of agential authority, but they are subject to a further objection originally raised against hierarchical views: it seems that certain sorts of *histories* undermine responsibility, even if the action results from the preferred kind of structure. For example, it seems entirely conceivable that someone who drifts into a religious sect can acquire a new set of values or self-governing policies or higher-order desires as a result of external pressure, indoctrination, and manipulation. Often, we are inclined to think that such people are lacking in free will and responsibility, and certainly in authenticity. If there are any such cases, purely structural models of free will cannot be sufficient. There must be in addition some kind of restriction on the sorts of histories that are compatible with freedom and responsibility. One popular answer is that the values or attitudes must result from *reasons-responsive* mechanisms. In this vein, Alfred Mele argues that agents lose autonomy when their values are changed by mechanisms that *bypass* their capacities for controlling their mental lives, centrally capacities to make and revise evaluative judgments on the basis of standards one endorses and modify desires and emotions

[204] Bratman 2004/2007, 245.

accordingly.[205] This is building on the sort of evaluativist model Watson outlines.

Fischer and Ravizza, in turn, build their model on the basis of consideration of Frankfurt-style cases. In Frankfurt's original example, the counterfactual intervener Black stands by to interfere with Jones's brain processes in case he chooses B rather than Black's preferred alternative A; alas, Jones chooses A anyway, so Black never does anything. Identifying making a choice with adopting an intention as a result of deliberation, the situation can be represented with the following simplified diagram:



In this case, Frankfurt argues, Jones "will bear precisely the same responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he do it"[206]. Suppose the action is shooting the president. Jones, deeply unhappy about the warmongering of the administration and aware of the likely consequences, has decided it is time to trim the executive branch. Black, a clever scientist frustrated with the president's anti-scientific

[205] Mele 1995, 166–167. Mele allows that one can autonomously arrange for oneself to be compelled in this sense.

[206] Frankfurt 1969, 836.

religiosity, has covertly installed a remote-controlled monitoring and controlling device in Jones's brain. Were Jones to give a sign that he is about to form the intention to refrain from shooting, Black would press a button and cause Jones to from the intention to shoot anyway.[207] But as it is, Jones does not waver and pulls the trigger. Jones could not have done otherwise, but he is nonetheless intuitively responsible.

Fischer and Ravizza point out that in the actual sequence – the one in which Jones responsibly chooses A – Jones's choice results from a process of practical reasoning, while in the counterfactual sequence, it results from direct stimulation of the brain. The relevant difference, he argues, is that *were* Jones presented with sufficient reason r' to choose B rather than A, the actual, responsibility-entailing mechanism of practical reasoning would in a range of different scenarios lead him to choose B, while the counterfactual mechanism would be entirely insensitive to reasons.[208] Fischer and Ravizza distinguish between two aspects of reasons-responsiveness: reasons-recognition (being able to recognize the reasons there are) and reasons-reactivity (being able to make choices in accordance with reasons that are recognized to be sufficiently good). Since we can be responsible when we fail to make the best available choices, they believe that moderate reasons-responsiveness (being able to respond to a significant range of, but not all, reasons) suffices for guidance control and moral responsibility.[209] However, they add a further requirement that the agent also *own* a reasons-responsive

---

[207] The need for a prior indication creates problems if the world is indeterministic, since in that case there is no reliable sign of the choice to be made until it is actually made – whatever the conclusion of the deliberation, the agent might still choose B. By then, it will be too late for the controller to intervene, but if the controller intervenes before the choice, he is no longer merely counterfactual, and the agent is no longer intuitively responsible (since the choice resulted from external intervention). See Widerker 1995 for this type of argument against Frankfurt and Mele and Robb 1998 for a sophisticated response.

[208] Fischer 2006, 18.

[209] Fischer and Ravizza 1998, 41–46.

mechanism – they believe that one cannot be manipulated into reasons-responsiveness without losing responsibility.[210]

Not all views that take reason-responsiveness to be central focus on history and the right sort of actual sequence. Susan Wolf's 'Reason View' builds on the intuition that responsibility for choices requires having a grasp of their value. Wolf refers to the so-called M'Naghten rules in criminal law, still widely referred to in common law jurisdictions.[211] In 1843, Daniel McNaughton, a Scottish woodturner, shot a civil servant while suffering from serious paranoid delusion. The case required the House of Lords to set a standard for insanity defence. According to it, a defence on this ground requires that it is clearly established that

> at the time of the committing of the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing; or, if he did know it, that he did not know he was doing what was wrong (M'Naghten's Case [1843] UKHL J16[212])

This is not quite precise enough for philosophical purposes. Intuitively, someone who is *incapable* of telling right from wrong cannot fairly be blamed for doing something wrong. A negligent person might also fail to know that what he is doing is wrong, but as long as he is not incapable, he is still culpable – he *should* know. Moreover, one may know that what one is about to do is wrong but be unable to refrain from doing it nonetheless.[213] Consequently, not only evaluative but also motivational capacities to act according to reasons are needed for full responsibility. Wolf argues that these considerations favour an *asymmetrical* view about the need for alternative possibilities. If one does the right thing for the right reasons, "in accordance with the True and the Good"[214], one need

---

[210] Fischer and Ravizza 1998, ch. 8.

[211] Wolf 1987, 381. See also Wallace 1994.

[212] http://www.bailii.org/uk/cases/UKHL/1843/J16.html

[213] In law, this is known as 'irresistible impulse'.

[214] Wolf 1990, 79.

not be able to have done otherwise to be fully morally responsible. Someone who cannot help jumping in the water to save a drowning child because she sees it as the only thing to do in the situation (that is, her psychological makeup makes it impossible for her to do otherwise) deserves no less praise. As Wolf points out, phrases like "I cannot tell a lie" and "he couldn't hurt a fly" are "not exemptions from praiseworthiness but testimonies to it."[215] However, if one does something wrong, then it must be true that one *could* have done the right thing for one to be responsible in the accountability sense. In her example, JoJo, the son of a dictator, born and bred for brutality, becomes a cruel dictator himself. As a result of his twisted and unusual upbringing, he is literally incapable of realizing that his actions are wrong. Even if JoJo wholeheartedly endorses his first-order desires, even if they are in line with his values, even if they fit with his self-governing policies, it does not seem that he is responsible for his actions, Wolf claims.[216]

One thing the Reason View has been criticized for is its neglect of history. Mele brings up a case in which a person who is capable of recognizing the True and the Good deliberately hardens himself and eventually becomes a cold-hearted killer unable to orient himself by the Good. It follows from the Reason View that he is no longer responsible. But for Mele, this is implausible: since being guided by reasons "is a capacity that he voluntarily and successfully sought to eliminate, it is difficult to see why its absence should absolve him of moral responsibility for his behaviour"[217]. From the other direction, Fischer argues that it is implausible that one could be morally responsible if one was manipulated *into* having the right sort of

---

[215] Wolf 1990, 80.

[216] Given the thin understanding of reasons-responsiveness that Mele and Fischer and Ravizza use, JoJo might be responsible on their accounts, too. I cannot discuss this further here. The plausibility of Wolf's example is a different issue – just what kind of upbringing can make a physiologically normal person blind to basic moral distinctions? For worries on this account, see Vogel 1993.

[217] Mele 1995, 163.

connection to the Good.[218] Should Wolf be worried about these cases? Perhaps not. Mele's killer may be responsible for hardening his character (which he did while still able to do otherwise), and thereby indirectly responsible for his later actions. As to becoming tuned to reasons, Fischer's criticism may reflect a kind of prejudice. As McDowell's reading of Aristotle emphasized, it may well be a matter of luck and non-rational processes of moral upbringing that one comes to be aware of the reasons that there really are.[219] Once one has the reasons in view, it does not really matter how one got there. Suppose Wolf's JoJo were given a virtue pill that gave him the ability to see the world aright. Surely if he continued beating up cartoonists and whatever, he would now be fully responsible for it.

Libertarian Moral Psychology

I have spent most of this section discussing various compatibilist theories of the psychological conditions of moral responsibility. There do exist, however, also incompatibilist (libertarian) versions. I will finish with a quick look at Robert Kane's theory.

For Kane, the most important condition for free will is Ultimate Responsibility (UR); insofar as alternative possibilities matter, it is because they matter to UR. To be ultimately responsible for an action, an agent "must be responsible for anything that is a sufficient reason, cause, or motive for the action's occurring."[220] If a choice results from an agent's character and motives, for example, she must be responsible for them to be responsible for the choice. If responsibility for character requires having made choices in the past that have formed the character, the agent must be responsible for those choices in turn. This quickly threatens to lead to the sort of regress that pessimistic source incompatibilists like Galen Strawson (2001) delight in. Kane agrees with Strawson to the extent that if

[218] Fischer 2006, 34n53.
[219] McDowell 1995/1998. See above, section 2.2.3.
[220] Kane 2005, 121.

determinism is true, there will be sufficient conditions for whatever we do in the long-gone past, so that ultimate responsibility is impossible. But what if determinism is false? Then it could be, in principle, that at some points in our lives we are able to make choices for which there are not sufficient conditions in the past. The big question for libertarians is just what kinds of effect of indeterminacy count as *our choices* rather than random swerves of atoms.

Kane begins his positive answer by limiting the scope of undetermined choice needed for UR. As compatibilists like to point out, the fact that Luther "could do no other" when he stood condemned does not mean he was not responsible for standing up. Kane acknowledges that having one's will set in one way is compatible with UR, but claims that in that case the agent must be responsible for his will being set that way. This means that the real debate concerns relatively few "will-setting" or "self-forming" actions. At some point in the past, the ultimately responsible agent must have been able to act in more than one way, and do so "voluntarily, intentionally, and rationally"[221], not just by accident or mistake. This is the connection between UR and alternative possibilities. But what kind of indeterminism makes possible voluntary and rational self-forming actions? Kane appeals to two physical theories, quantum indeterminacy, which makes it plausible that the exact timing of the firing of individual neurons in the brain might be indeterminate, and chaos theory, according to which small changes can have large effects in suitable conditions.[222] His hypothesis is that moments of personal conflict create just the sort of conditions in which this sort of indeterminacies might occur and their effects be magnified. As he puts it, "[t]he uncertainty and inner tension that we feel at such soul-searching moments would thereby be reflected in the indeterminacy of our neural processes themselves."[223]

Kane's central example is a businesswoman faced with a choice of making it in time to a very important meeting or stopping to help

[221] Kane 2005, 128.
[222] Kane 2005, 133–135.
[223] Kane 2002, 417.

a stranger. This is a potentially formative choice between selfishness and morality. Kane imagines that in this kind of situation there are two competing neural networks in the agent's brain, one constituting her desire to make it to the meeting and the other her desire to help.[224] Supposing this competition brings about chaotic indeterminacy, there is no way in advance to tell which way the agent goes (which neural pathway "wins"). Whichever choice she will make, it is not determined by the past and the laws of nature. At the same time, Kane argues, whichever choice she will make, it will be voluntary rather than accidental (since it will be what she tries to do) and it will come about for the agent's own reasons. After all, it is the neural processes that constitute the agent's trying to decide that create the indeterminacy in the first place. They are her own pretty much in the same sense as compatibilists understand ownership: "what makes these efforts, deliberations, reasons, and intentions *hers* … is that they are embedded in a larger motivational system realized in her brain in terms of which she defines herself as a practical reasoner capable of responding to and acting on such reasons."[225] Though the choice is not directly controlled by the agent's reasons and motives, it is not a matter of brute luck either. The agent's own character and motives explain why there is conflict and effort to solve it without explaining the outcome.[226] An indeterminacy of this sort during self-forming actions, Kane believes, suffices to stop the regress of ultimate responsibility and so ground full moral responsibility.

I will not here discuss the philosophical challenges to this view. Even if Kane's suggestion works philosophically, it is an empirical question whether the brain works the way it must for us to be free and responsible.

---

[224] Kane 1996, 126; Kane 2002, 419.

[225] Kane 2002, 423–424.

[226] Kane 1996, 127.

3.3 Empirical Questions about Moral Responsibility

As I noted, questions about the nature of psychological processes needed for moral responsibility are paradigmatically *a priori*, and until recently there has not been much empirically informed work done in the area. Recent years have, however, seen an explosion of interest in the issue. Part of this is because the research programme of experimental philosophy promises a scientific method for answering questions about folk concepts and intuitions, in this case concerning freedom and responsibility. Eddy Nahmias, Thomas Nadelhoffer, and their colleagues have run surveys that call into question the alleged intuitiveness of incompatibilism, at least when it comes to concrete cases.[227] As I argue in Essay 1, I do not believe that philosophical claims about concepts are empirically testable in this way, so I will leave these studies aside here. However, even if we can answer the conceptual and metaphysical questions *a priori*, there is a further question that is of much interest to us: whatever it takes to be free and morally responsible, do (any or most) *human beings* have the required capacities? What does it take for human beings to come to have such capacities? These are undeniably *a posteriori* empirical questions. They arise for both incompatibilists and compatibilists. As far as I know, there are no empirical studies either confirming or disconfirming the empirical assumptions that Kane, the most sophisticated event-causal libertarian, makes, nor is obtaining such results within the means of contemporary science, so I will not discuss the challenge to incompatibilism here. But there do exist data that allegedly call into question whether human beings are morally responsible in the compatibilist sense. This gives rise to a position

[227] See Nahmias et al. 2005. Woolfolk, Doris, and Darley (2006) found support for the view that agents who identify with their actions are held responsible even if their actions are completely determined. In an interesting study, Josh Knobe and Shaun Nichols (forthcoming) found that when the theses were formulated abstractly, people tended to be incompatibilists, but when they had to allocate responsibility for particular cases in a deterministic world, they tended to be compatibilists!

Eddy Nahmias has labelled 'neurotic compatibilism'[228]. Neurotic compatibilists are convinced that determinism and moral responsibility are compatible, provided that the causal chain that leads to action is of the right kind, but worry that human beings might lack the necessary psychological mechanisms to produce such causal chains.

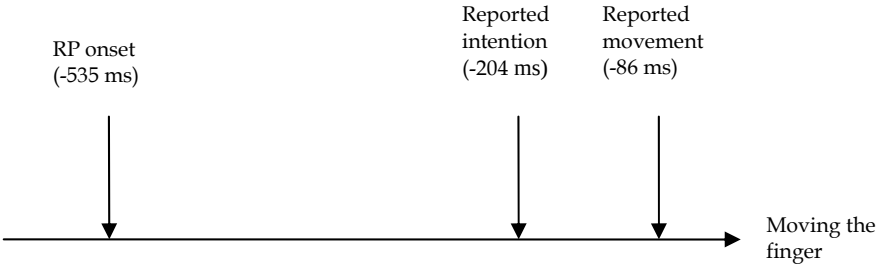
Two kinds of empirically-based worries have been particularly prominent recently. First, neuroscientists have come up with results that suggest to some that our decisions are already made by the time we become conscious of them. Second, and along the same lines, social psychological studies on automaticity suggest that potentially large behavioural differences arise from minor, non-rational, subconscious and uncontrollable environmental cues. What is the philosophical relevance of these results?

The Neurophysiological Challenge

Early brain research in the 1960s (such as Kornhuber and Deecke 1965) showed that a measurable peak of electric activity in the brain occurs reliably up to 800 ms before intentional motor activity, such as moving one's fingers at will. This peak was termed 'readiness potential' (RP). Since the late 1970s, the neuroscientist Benjamin Libet has conducted a series of experiments to measure the exact timing of RP in relation to conscious intention to act. In the most famous study (Libet et al. 1983), the participants were equipped with devices to record electric activity in the brain, told to lean back in a chair, and flex their fingers or wrist at any time they felt like doing so (but at least forty times during the experiment), and some moreover to "let the urge to act appear on its own at any time without any preplanning or concentration on when to act"[229]. All they had to do in addition was to observe the position of a fast-moving dot on a clock face in a computer screen at the time when they first

[228] Nahmias (forthcoming).
[229] Libet et al. 1983, 625.

experienced the urge or intention to act or the time the experienced the actual movement. Though there were minor differences depending on the methods of reporting and measuring, the results were clear: "[N]euronal processes that precede self-initiated voluntary action, as reflected in the readiness potential, generally begin substantially *before* the reported appearance of conscious intention to perform that specific act."[230] The following figure shows the average times recorded (modified from Wegner 2002, 53):



Libet interprets these results as showing that "the brain evidently 'decides' to initiate, or, at the least, prepare to initiate the act at a time before there is any reportable subjective awareness that such a decision has taken place"[231]. In later work, he suggests that we can, however, consciously 'veto' an action before the "actual motor outflow" – after all, the awareness does occur some time prior to the action itself, even if readiness potential precedes it.[232] In subjects instructed to intend to flex at a pre-set time and cancel that intention just before then, an early RP is followed by "flattening or reversing" of the potential around 150–250 ms prior to the pre-set time.[233] Nevertheless, when action does occur, it is initiated by the brain before conscious awareness.

[230] Libet et al. 1983, 635.
[231] Libet et al. 1983, 640.
[232] Libet 1985, 537–538.
[233] Libet 1985, 538.

Why, then, do we experience our conscious thoughts as causing our actions, if they could not possibly do so? That is, why do we have the illusion of conscious will? Perhaps the most comprehensive answer to this is Daniel Wegner's (2002) theory of apparent mental causation. Wegner surveys an impressive array of phenomena in which people either think that their conscious thoughts are causing their actions or other events when that cannot be the case (such as phantom limbs[234], direct brain stimulation[235], stopping a mouse pointer on the screen when another person is in fact in control[236]) or that their thoughts are *not* causing their actions or other events when that it is fact the case (such as automatic writing, dowsing (well-spotting)[237], ideomotor action[238]). For Wegner, such phenomena and experiments, including Libet's results, show that conscious willing

---

[234] People who have amputated limbs often experience willingly moving them when there is in fact no action (see Wegner 2002, 40–45.

[235] In José Delgado's studies, electric stimulation of part of a brain caused the subject's head to move to one side or another; when asked, however, the subjects confabulated reasons for turning their head (such "I am looking for my slippers"). See Wegner 2002, 46–47.

[236] Wegner and Wheatley (1999, 487–489) told subjects move a pointer around randomly on a screen with pictures of household objects on it, and stop whenever they felt like it. A confederate was moving the same mouse, and unbeknownst to the subject given instructions to stop at specific times. During the experiment, the subject heard on intervals names of objects on the screen to prime him or her to think about a specific object at a specific time. When the subject had been primed to think about an object (such as a swan) and the confederate forced the pointer to stop on it, the subjects reported that *they* had stopped the pointer themselves. This was not the case if they did not think of the object beforehand. (More precisely, the subjects reported, on average, a high degree of intentionality for forced stops if they were primed just beforehand; if the thought occurred 30 seconds before the stop, they did not experience consciously stopping the pointer at the mentioned object.)

[237] It turns out that dowsers are in fact responding to observable cues about the presence of water (differences in vegetation and so on) and do no better than chance without them (Wegner 2002, 116–120).

[238] See Wegner 2002, 120–230 and the discussion of automaticity below.

does not cause actions. Instead, unconscious brain processes cause *both* actions *and* causally unrelated conscious intentions. Appearance of causality, as in other cases, results from the perceived priority, consistency, and exclusivity of thought with respect to action.[239] As Wegner summarizes it, "[f]or the perception of apparent mental causation, the thought should occur before the action, be consistent with the action, and not be accompanied by other potential causes."[240] Since this often happens, we tend to think we have conscious control, even when we do not.

The Automaticity Challenge

Social psychologists have long argued that social perception involves automatic, non-conscious processes that give rise to evaluatively laden categorizations. Studies on stereotyping and trait attributions have amply shown that we tend to be unaware of environmental cues that lead us to form beliefs about racial minorities and attractive people, for example. John A. Bargh and his colleagues have extended the study of this sort of automaticity to the effects of unconscious perceptions on behaviour. To take just one example, Bargh, Chen, and Burrows (1996) primed subjects in three randomly assigned groups with a sentence-scrambling test which was said to measure linguistic ability. One group had to unscramble sentences to do with politeness (like "they her respect see usually"), another rudeness, and third neutral. Their next task was to go down a hallway to report to the experimenter for the second part of the study. They found the experimenter having an interminable conversation with another 'participant' (who was in fact a confederate, another experimenter) and in no way acknowledging the presence of the participant. The object of the study was simply to see whether the participant would interrupt the conversation within ten minutes. 37% of the subjects who had unscrambled neutral words interrupted the conversation,

[239] Wegner and Wheatley 1999, 482–487. Compare Hume's *Treatise* 1.3.2.
[240] Wegner 2002, 69.

but the figure rose to 63% for those primed with rude words and reduced to 17% for those primed with polite ones.[241] When asked later, "no participant showed any awareness or suspicion as to the scrambled-sentence test's possible influence on their interruption behaviour"[242]. Similar lack of awareness of actual causes of behaviour has been found in a large number of other social-psychological studies.[243]

What explains this? According to Bargh's 'auto-motive model', automatic perceptions can directly give rise to pursuit of goals in the same way as conscious decisions – there is a "direct and automatic route … from the external environment to action tendencies, via perception"[244]. Quite simply, just as environmental stimuli can non-consciously activate a stereotype, they can non-consciously activate a goal representation, and this can be experimentally shown.[245] If this

---

[241] Bargh, Chen, and Burrows 1996, 235. In another experiment in the same study, people primed with stereotypically elderly-related words like "Florida", "sentimental", "conservative", and "wrinkle" were measured walking more slowly afterwards than those in the control group!

[242] Bargh, Chen, and Burrows 1996, 234.

[243] In one classic study, Nisbett and Schachter (1966, cited in Nisbett and Wilson 1977, 237) gave subjects placebo pills that were said to produce symptoms that were in fact those caused by electric shocks. The hypothesis was that they would tolerate stronger shocks if they attributed the symptoms to the pill. When shocks were subsequently given to the subjects, the placebo group tolerated on average four times more amperage than control subjects without placebo pills. But when they were asked afterwards whether taking the pill had had any effect on them, they denied it, explaining their greater tolerance to shocks with answers like "Well, I used to built radios and stuff when I was 13 or 14, and maybe I got used to electric shock"!

[244] Bargh and Chartrand 1999, 465.

[245] The idea is this: it has been shown before that our consciously held goals make a difference to what we perceive, remember, how much of an effort we make, and so on. If non-conscious priming has similar effects, we can conclude that it gives rise to non-consciously held goals. In one experiment (Bargh et al. 2001), Bargh and his colleagues first primed subjects with an unscrambling task involving achievement-related words, and then

happens, conscious decision-making becomes merely epiphenomenal. As Bargh and Chartrand put it, "it may be, especially for evaluations and judgments of novel people and objects, that what we think we are doing while consciously deliberating in actuality has no effect on the outcome of the judgment, as it has already been made through relatively immediate, automatic means."[246] Some research suggests that this may be the case not only for the sort of relatively trivial judgments tested in laboratory conditions. For example, John Doris brings up research based on public records that suggests that the so-called 'name-letter effect' (people prefer letters that appear in their own names) can influence major life decisions, like where to live and which profession to choose: women named Virginia or Georgia were 36% more likely than others to move to states sharing the same name, and men named Geoffrey or George were 42% more likely than others to be geoscientists![247]

Here, as in the case of affect-driven moral judgments, confabulation rears its ugly head. People do seem to have a need to come up with a story to make their choice look rational, whether or not they had any reasons. In a classic study, Nisbett and Wilson (1977) had subjects choose from four identical pairs of nylon stockings and indicate which was of best quality.[248] It turned out that there was a considerable positional effect: subjects were four times more likely to choose stockings placed on the right. However, the reasons subjects gave for their choice made no reference to position,

---

had them perform a task of building as many words as possible from Scrabble letters. These subjects came up with 8 more words in five minutes than a non-primed control group – a similar result to studies in which people are consciously construing a task in terms of achievement. As Bargh and Ferguson put it, in this and similar experiments "the achievement-primed participants consistently showed classic properties of being in a motivational state, despite not having consciously chosen or guided their behavior towards this goal." (Bargh and Ferguson 2000, 936)

[246] Bargh and Chartrand 1999, 475.

[247] Doris (forthcoming). The research that he cites is Pelham et al. (2002).

[248] Nisbett and Wilson 1977, 243–244.

talking instead about differences in knit, sheerness, and weave. When the experimenters explicitly asked about the effect of position, they denied that it made any difference! They hypothesize that when asked for an explanation, people consult cultural and personal "a priori causal theories" – basically, assumptions about what stimuli cause what response – that appear to make sense of their (or others') reactions, whether or not these schemas pick out the actual causes.[249]

Implications for Freedom and Responsibility

On the face of it, at least, the scientific and psychological data outlined above pose a challenge to philosophical conceptions of freedom and responsibility. If our conscious decisions and plans are epiphenomenal, or if it is irrelevant to what we in fact do which considerations we take to be reasons, and we are instead driven by uncontrollable brain processes and subconscious environmental cues, our practices of holding each other responsible rest on a massive error. It is therefore imperative for philosophers working in this area to examine what the data in fact show.

To begin with the neurophysiological studies, Al Mele draws attention to the carelessness with which Libet and many of his followers identify the onset of RP with intention or willing, while also talking interchangeably about 'urges' or 'wants'.[250] Clearly, from the perspective of philosophy of action, urges and intentions play a very different role in the generation of action. Urges, wants, and desires may serve as inputs or stimulants for practical reasoning, whereas intentions are on its output side and subject to very different rational constraints – for example, it is irrational if not inconceivable to both intend to do something and intend not to do it, while it is not unusual to both want to do something (like eat a piece of chocolate) and want to not do it (because we are all getting too fat

[249] Nisbett and Wilson 1977, 248–249.

[250] Mele (forthcoming); Mele 2006, ch. 2.

anyway).[251] There is a difference among intentions, too. Some concern future actions (I will play football next Sunday), some standing policies (I will prepare my job applications in time), and some what to do right now (I will press this button now). In Mele's terms, the first two are varieties of *prior* intentions, while the latter are called *proximal* intentions, since they are the proximal causes of intentional action.

Given these standard distinctions in the philosophy of action, how should we interpret Libet's results? It seems reasonable that the subjects had formed the *prior* intention to flex their wrists at least 40 times during the experiment whenever it occurred to them. So, they were consciously on the lookout for any sort of urge to flex during that specified time. Mele suggests, very plausibly, that the onset of RP at -550 ms represents just such an urge; it is, after all, a sort of bodily readiness to take action, just what one would be looking for in order to complete the specified task. (It is not an 'urge' in the sense of a sudden desire for something pleasurable, for example.) In that case, the reported awareness at -200 or so ms could well precede the *proximal* intention to flex now, and the subject would form (or not form, in the case of the veto studies) the proximal intention as a result of consciously experiencing the urge. On the basis of Libet's data and independent reaction time studies, Mele places the proximal intention itself at -90 to -50 ms prior to the bodily movement.[252] On this interpretation, then, there is no unconscious willing or deciding going on. The only real decision involved is to follow the instructions given, and consequently to form a proximal intention (a kind of 'conscious will') to act whenever one feels like it, or, in the veto studies, just record the time of the urge and refrain from forming the proximal intention. To be sure, the urge or readiness to act itself is still initiated unconsciously on this picture, but nobody ever claimed that we directly consciously control the

---

[251] Consequently, Libet's (1985) instruction for veto-subjects to both intend to flex at a pre-set time and intend *not* to flex at that time is incoherent. They can intend to prepare to flex and stop short of it, which is what they indeed seem to be doing.

[252] Mele (forthcoming), sections III and V.

emergence of such things, so that is no challenge to our commonsense picture of the will.[253]

As to Wegner's data, there is no need to deny that the extraordinary phenomena that he discusses exist. All that defenders of free will need to show is that they are indeed *extraordinary*. Eddy Nahmias draws a parallel with visual illusions: they are predictable products of a generally reliable system faced with unusual input.[254] He also emphasizes how odd it would be to think in the first place that we form conscious intentions before each and every movement – most of the time there is at best conscious monitoring going on (for example, I correct my balance if I slip while walking, but I do not rehearse each step in my mind before taking it). Certainly Wegner's results are not sufficient to support the conclusion that "the real causal mechanisms underlying behaviour are *never* present to consciousness"[255].

What about automaticity? Are we just happily rationalizing decisions made in response to to subconscious situational cues that have nothing to do with what we take to be our reasons? Consider what happens if I am cooking Thai red curry and it comes to a boil soon after I have put the coconut milk in. If I am at the same time having a conversation with a friend, I may turn down the heat not, merely without deliberating, but also without even noticing that I did so. If the food turns out well, I am no less to praise for such automatic responses to situational cues than I am for more deliberate choices like how much coconut milk to use. They are no less sensitive to reasons and no less manifestations of my skill.

---

[253] There is still an indirect control of the emergence of urges going on – most of us do not regularly feel the urge to flex our wrists!

[254] Nahmias (forthcoming).

[255] Wegner and Wheatley 1999, 490. The claim would, of course, be trivially true if by 'real causal mechanisms' was meant the neurophysiological states and their relationships that realize or constitute our psychological states and their relationships. But that is not what Wegner and Wheatley try to say; rather, they mean that my belief that a car is approaching (or, presumably, the neurophysiological state that realizes or constitutes it) cannot be the cause of my jumping off the road.

Automaticity as such seems thus not to threaten any element of our commonsense picture of agency. But what if automatic processes are not in harmony with our more or less consciously held values? The first thing to note is that even if they are beyond our direct control, they may be indirectly controllable. Pizarro and Bloom (2003) point to the effects of consciously framing an issue in a certain way, selective exposure to right sort of environments (for example, implicit racism is reduced through exposure to positive black exemplars[256]), and redirecting attention.

Indirect control, naturally, presupposes some awareness of the automatic processes, so it cannot be the answer to all challenges from automaticity. What should we make of the sort of priming effects that Bargh and colleagues highlight? Not only are the subjects unaware of them, but they also deny that they made any difference to their behaviour when asked. Looking more carefully at the studies and their results, it does not seem that they are cases of people being driven by forces beyond their control. Rather, the test scenarios are carefully constructed so that the subjects have roughly equal reasons to do one thing or another, for example either interrupt the irritating conversation or remain patient. Moreover, what is at stake is nothing very dramatic, so there is no particular need for the subjects to reason carefully about their choice. Under these conditions, it is not so very shocking that associations activated by priming tip the balance one way or the other. They seem to be on par with the effect of a bad or good cup of coffee, or listening to an aggressive talk radio or NPR on the way to work.[257]

---

[256] Dasgupta and Greenwald 2001. They conclude that "the present research provides new evidence suggesting that automatic preference and prejudice may indeed be malleable" (Dasgupta and Greenwald 2001, 806).

[257] What about the study by Pelham et al. (2002) on the effect of one's name on choice of profession or place to live? Surely these are dramatic effects? Indeed they are. But one's name is not exactly a situational factor. It is surely plausible that someone named Georgia would have a conscious interest and perhaps a fondness for the state of Georgia – surely it would catch one's attention more than others, at least. Would it be surprising if when one deliberated between two roughly equal choices, say Georgia and

There exists, in fact, some experimental data supporting the tiebreaker interpretation. Macrae and Johnston (1998) primed people with helping-related stimuli and conducted the actual experiment in an elevator on the way out, when the subjects thought the whole thing was already over. In the first scenario, a confederate dropped a number of ordinary pens on the floor; in the second, the pens were leaking and messy-looking. As other automaticity studies would lead one to expect, helping-primed subjects were more likely than the control group to help in the first scenario. But in the second scenario, when there was a stronger reason not to help (getting one's clothes dirty), the difference vanished. Moreover, when participants had a competing conscious goal, this had the same result. These data suggest that the motivational effects of non-conscious priming are much less dramatic than automaticity enthusiasts think.

In short, existing empirical data do not seem to threaten the existence of free will and moral responsibility, since they do not undermine the existence and influence of the sort of rational and volitional capacities that are necessary for compatibilist freedom, even if the scope of these capacities may be narrower than some philosophers have thought. We can therefore turn the question around: what kind of empirical circumstances are conducive to the acquisition and use of rational and other capacities required by freedom and responsibility? This leads to different empirical and conceptual questions, which I address in some detail in 'The Social Dimension of Autonomy'.

Essay 3: 'The Social Dimension of Autonomy'

I begin 'The Social Dimension of Autonomy' with a brief overview of some traditional conceptions of autonomy. They are closely related to views about free will and moral responsibility, and this is no

South Carolina, this would tip the balance? It would be considerably more dramatic if people named Georgia had little other reason to move there but did so anyway, but this seems unlikely, and is certainly not shown by the sort of archival study that Pelham and colleagues did.

accident: as I take it, autonomy is a theoretical concept we use for the sort of self-governance that grounds ascriptions of full moral responsibility. Autonomy, for short, is authentic self-determination. To be autonomous, one must both have the sort of capacities required for authentic self-determination and exercise them – without the exercise, one is at best *potentially* autonomous. This gives rise to three questions:

1) What are the *psychological capacities* needed for autonomy?
2) What does it take to *acquire* the autonomy-relevant capacities?
3) What does it take to *exercise* the autonomy-relevant capacities?

In the spirit of reasons-responsiveness views about moral responsibility, I argue that autonomy requires the sort of normative competence that allows one to recognize what reasons one has and the capacity to be motivated by one's perception of reasons. I assume that Wolf and Fischer and Ravizza are correct that having such capacities is compatible with determinism. The real focus of the paper, however, is on two other questions, particularly the last one. On the issue of acquisition, I qualifiedly endorse the communitarian view that one must belong to a community to acquire the ability to stand back from one's desires and assess one's options in the language of qualitative distinctions. Communitarians like Sandel go too far, however, when they suggest that one's 'constitutive ends' must permanently remain outside assessment; I see no reason why they should not be subject to Neurathian scrutiny (see section 2.2). This aspect of the social dimension of autonomy is by now familiar and, with the sort of qualifications I make, relatively uncontroversial.

The question of what it takes to *exercise* the autonomy-relevant capacities – what it is to live an autonomous life – is less familiar, and often unasked. Since my interest is in the social conditions of autonomy, I in effect divide it in two subquestions:

3.1) What are the social conditions for having the psychological capacities and attitudes required for exercising autonomy?
3.2) What are the social conditions for leading an autonomous life?

Axel Honneth's attempt to naturalize Hegel's theory of recognition, I argue, aims to answer the first of these questions. There are two stages to the argument: identifying the psychological capacities or attitudes conceptually necessary for exercising autonomy, and then identifying the social psychological mechanisms and relationships that are nomologically necessary for the development and maintenance of those capacities and attitudes. Thus, Honneth argues that if we lacked *basic self-confidence*, we would be hindered from exercising our autonomy-relevant capacities, since we would not give our desires and needs the weight they deserve in deliberation. Drawing on developmental psychology, he claims that *love* and caring are the forms of recognition by others that are needed for the development of basic self-confidence. (I criticize the paucity of Honneth's empirical evidence here and elsewhere, but accept the basic thrust of the argument.) If we lacked *self-respect*, understood as a positive attitude toward one's ability to make rational decisions, our autonomy would be again undermined – we would defer to others, debase ourselves. Honneth ties self-respect to having the sort of sense of entitlement that comes with having *rights*, a different form of recognition by others, not just by individuals but also by institutions. Finally, if we lacked *self-esteem*, understood as valuing ourselves under particular role-descriptions and identities, we would be hindered from undertaking particular projects involving those identities. Honneth claims, without much in the way of hard evidence, but still plausibly, that self-esteem in this sense results from *social esteem* manifested in the distribution of both symbolic and material rewards.

Honneth's work is a valuable contribution to the literature on autonomy and responsibility, and I offer a friendly amendment to it by reformulating some of its theses in the language of normative competence views. However, I argue that Honneth fails to see that there is a further social dimension of autonomy consisting of the social conditions for exercising autonomy-capacities in the real world, that is, for living an autonomous life. Self-confidence or no self-confidence, one cannot act on reasons deriving from one's emotional needs if one is shut out of personal relationships by

cultural patterns of value. No matter how much self-respect one has, one cannot shape one's life on the basis of independent judgments unless one's social status includes a guarantee against arbitrary intervention by others. Even a self-respecting slave is still a slave, and as such not responsible for choices with respect to which she cannot exercise her autonomy-capacities. And finally, self-esteem does not suffice for exercising one's autonomy with respect to particular projects; one must also have genuine access to practices within which these projects and the identities they involve acquire significance. In short, the subjective experience of intersubjective recognition must be met with the objective fact of intersubjective recognition for an agent to be able to exercise her autonomy, and so for her to be fully morally responsible. There are non-psychological necessary conditions for exercising autonomy.

These arguments have consequences for any view in political philosophy that takes autonomy to be an important value. Liberal theories, in particular, take respecting and promoting autonomy to be central for justifying and constraining policies. The interesting question is whether acknowledging the internally complex social dimension of autonomy forces us to reconsider the basic liberal tenets of individual rights, state neutrality, and equality of opportunity. I finish the paper by sketching some areas where the social conception of autonomy exerts pressure on the liberal views.[258]

[258] The volume in which the paper will be published will include Honneth's response, but as of this writing, I have not seen it.

4 Normative Ethics and What We Are Like

In the previous sections, I have examined the role of moral psychology in second-order questions about our ethical practices. When it comes to first-order, normative ethical questions, the role of psychological facts is different. As I suggested in section 1, the distinction between the philosophical and the non-philosophical seems to coincide with that between the normative and the descriptive.[259] However, what is normatively required of us depends on descriptive psychological facts in two ways. First, insofar as ought implies can, the demands of morality on us are constrained by our cognitive and motivational limitations – tying in with the issue of moral responsibility, we cannot fairly be blamed for not doing something that it is impossible for us to do.[260] More controversially,

[259] This is not universally agreed upon. Appiah (forthcoming) explicitly argues that empirical research has a bearing on normative questions not only by way of helping derive concrete recommendations, but also directly informing us of what well-being consists in, for example. Prinz (forthcoming) argues that if naturalism is true, *all* meaningful questions, normative questions included, are amenable to empirical study. I do not have the space to examine these arguments and their flaws here.

[260] There is a sense in which we can never do other than we actually do if determinism is true. That is why various compatibilists about moral responsibility have rejected that ought implies can. Fischer's (2006, 24–25) example is a woman, Sally, who has an arrangement with a lifeguard that she will raise her hand to alert him if a child is in trouble. She sees a child drowning, but does not raise her hand, and the child drowns. Unbeknownst to her, she is temporarily paralyzed in such a way that she would have been unable to raise her hand in any case. Fischer argues that Sally is blameworthy nevertheless, and that she acted wrongly, and thus that she ought to have raised her hand. If this is the case, ought does not imply can.

some have argued that even if something is not strictly impossible for us, morality should not demand us all to be *saints*, who are, after all, individuals of exceptional moral capacities.[261]

Second, depending on the normative theory in question, various sorts of psychological facts must be taken into account in deriving concrete rules or recommendations from abstract principles. In the simplest case, what the right rules or actions are according to hedonistic utilitarianism depends on which alternative results in the greatest amount of pleasure for sentient beings. This obviously requires knowing what kind of things give pleasure and pain to people and animals. Philosophical interest of this sort of facts, however, is not limited to hedonistic theories of value. Since virtually every normative theory, even Kantian deontology, gives some place to happiness, the recently burgeoning field of 'positive psychology', empirical study of the causes and measurement of felt well-being, holds much interest for ethicists.[262] On the non-consequentialist side, the key value is respecting and promoting personal freedom and autonomy. I have already discussed its empirical and conceptual conditions in Essay 3, and draw out further implications to the structure of normative arguments in Essay 4.

4.1. Psychological Realism in Normative Ethics

In *Varieties of Moral Personality*, Owen Flanagan articulates a thesis he calls the Principle of Minimal Psychological Realism:

Arpaly (2007, ch. 2) argues by parity with norms for belief, which is commonly not taken to be under volitional control.

[261] The classic discussion is Wolf 1982, who argues that moral perfection "does not constitute a model of personal well-being which it would be particularly rational or good or desirable for a human being to strive" (Wolf 1982, 419).

[262] For an overview from a philosophical perspective, see Tiberius 2006.

> Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behaviour prescribed are possible, or are perceived to be possible, for creatures like us. (Flanagan 1991, 32)

Theories that fail to meet the constraint articulated by the principle arguably fail at a central task of ethical theory, namely giving us some guidance in decision-making and evaluating. Certainly, many moral philosophers believe that moral theory is not just an attempt to describe and explain how things are in the way that biology or physics is. It also has a practical import. Aristotle saw this clearly:

> The purpose of this inquiry is not, as it is in other inquiries, the attainment of theoretical knowledge: we are not conducting this inquiry in order to know what virtue is, but in order to become good, else there would be no advantage in studying it. (*NE* 1103b)

From a more theoretical perspective, minimal psychological realism is supported by the connection between the ability to follow the demands of morality and fairness of holding one responsible for failures, as noted above. There are thus both practical and theoretical reasons for normative ethicists to ensure that they do not set up ideals that would be either cognitively or motivationally too taxing for us.

On the cognitive side, the constraint of psychological realism is widely accepted – just about any act-consequentialist acknowledges that as a decision procedure, it would require "an impossible amount of attention to one's action options"[263]. We simply cannot calculate and compare the outcomes of all possible actions. The standard response is to defend act consequentialism as a criterion of rightness instead, and allow for one heuristic or another to do the job of guiding decision.[264] Motivational constraints are more controversial. One reason for this is that it is not so clear what we can

[263] Flanagan 1991, 34.

[264] See for example Hare 1981 on the levels of moral thinking and Railton 1984.

and cannot do in this respect. Flanagan distinguishes between natural and social limits on psychological traits.[265] Sexual desire, for example, is natural, but its expressions in action are heavily influenced by culture. Consequently, the sort of things that are not too demanding for members of one culture may be such for those who have deeply internalized the practices of another. They may not be strictly speaking impossible, but too costly to be worth it. This is, of course, controversial, since it involves modifying "ought implies can" into something like "ought implies can without excessive cost to personal well-being or self-deception". Many are tempted to rather conclude with Kant that if we are too weak to meet the demands of morality, so much worse for us: what is wrong is still wrong.

I need not take a position on the motivational demandingness issue here. Instead, I want to discuss empirical evidence for two psychological claims that would relatively uncontroversially require us to modify our normative theories if they were true: psychological egoism and the thesis that there are no such traits as virtues.

The Reality of Altruism

I have already discussed the issue of the motivational effectiveness of moral reasoning and moral sentiments above, so I will limit myself here to the more general question of the possibility – and actuality – of altruism, which provides a good illustration of the different roles of a priori reflection and experimental study. The form of altruism at issue is *psychological altruism*, which is a matter of having non-instrumental motives to benefit others – that is, motives that do not derive from some concern for one's own good.[266] It must be distinguished from *behavioural altruism*, which is the rather trivial

[265] Flanagan 1991, 41–46.

[266] For a definition of psychological altruism that clearly distinguishes it from other things (sometimes misleadingly) called altruism, see Joyce 2005, 13–16.

claim that we sometimes act in ways that benefit others at a cost to ourselves[267], and from *evolutionary altruism*, the thesis that sometimes genes that dispose their vehicles (individuals) to sacrifice their reproductive fitness for others get selected for. Opposed to psychological altruism is psychological egoism, the thesis that all our motives are ultimately self-regarding.[268] The truth or falsity of altruism is naturally of significance for minimally psychologically realist moral theories, since if it is the case that we are *incapable* of genuinely other-regarding motives, morality cannot demand them from us, supposing that ought implies can. This would require a radical revision of at least Kantian and virtue theories, as well as some consequentialist theories of agent-evaluation.[269]

Since psychological egoism and altruism are hypotheses about actual human motivations, they are straightforwardly an empirical matter.[270] Testing them empirically is nevertheless challenging, since there is always potentially a gap between external behaviour and ultimate motives, as well as between our introspective view of our motives and our real motives. Getting around this requires considerable ingenuity, and is not easily accomplished. Perhaps the

---

[267] There are forms of behavioural altruism, however, whose existence is contentious. Ernst Fehr and his colleagues in the field of experimental microeconomics have conducted a series of experiments to study altruistic punishment, which is a form of behavioural altruism that seems to play a particularly important role in social coordination. See Fehr and Fischbacher (2003) for an overview of this research, which corrects some simplistic behavioural assumptions that rational choice theorists are sometimes liable to make.

[268] As Stich, Doris, and Roedder (MS) remind us, it is not simple to decide which desires are self-regarding and which are not. This is clearly a conceptual question that must be settled before any empirical investigation.

[269] Insofar as consequentialists assess *actions* only in light of their consequences, the possible falsity of altruism does not affect them.

[270] Philosophers have, to be sure, not always thought so. Mill 1863, for example, tries to prove a priori that altruism is "physically and metaphysically" impossible because of the connection between pleasure and desire. Blackburn 1998 gently corrects him with Butler's point about the dependence of many pleasures on pre-existing desires.

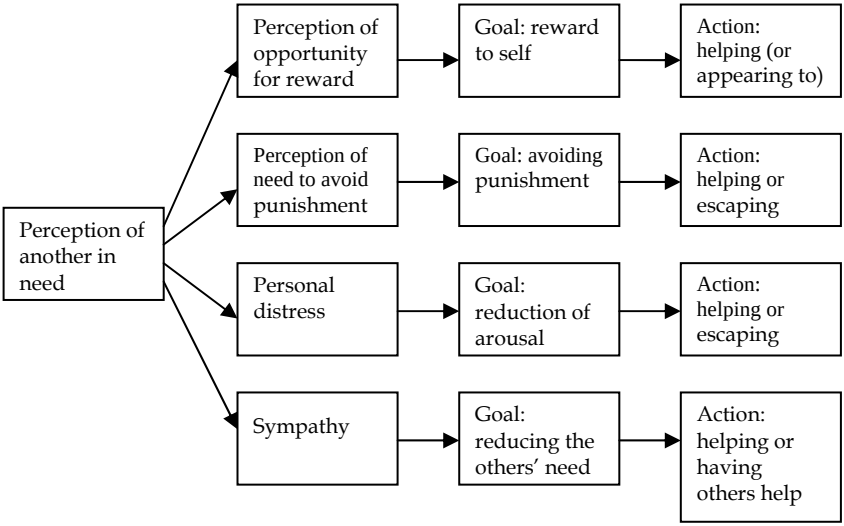
cleverest experiments to date have been conducted by Daniel Batson and colleagues. Batson notes correctly that altruistic motivation for helping is compatible with receiving psychological and non-psychological rewards (such as pleasure, fame, and money) for it, as long as the rewards are a *by-product* (a perhaps anticipated but unintended consequence) rather than the ultimate goal of helping behaviour.[271] When motivation is egoistic, helping the other is only a *means* to some benefit for the self. To settle the question about psychological egoism we must thus be able to distinguish between goals, means, and unintended consequences of actions on the basis of observable behaviour.

How, in general, do we do this? Well, goal-seeking behaviour presumably terminates when the goal is reached or believed to be unachievable, assuming the agent is minimally rational. In contrast, if something is a by-product, bringing it about or believing it impossible makes no difference to the behaviour in question. Suppose Joan's boss is a big classical music aficionado who goes to all concerts of the local symphony orchestra. One day she announces she is going to go to see a Brahms violin concerto. Joan knows that if the boss sees her there, he will be pleased and think favourably of her in the next round of promotions. How can we discover whether Joan is going to the concert for the music (in which case pleasing the boss is a by-product of her action) or to please the boss (in which case going to the concert is a means to pleasing the boss)? It is hard to be certain, but if finding out that the boss will not be there after all leads Joan to drop her plan, or if finding an easier way to gain promotion has the same result, we can reasonably infer that going to the concert was for her at least in part a means for gaining the boss's favour, rather than something she wanted to do for its own sake. If she does go anyway, that does not, however, yet show that her ultimate goal in going is listening to music – there could be some further end to which going to the concert is a means. But if no plausible candidate is on offer, we have grounds to conclude that she wants to listen to Brahms for its own sake.

[271] See Batson 1991, 64–67; Sober and Wilson 1998, 217–222.

In general, if both A and B are known consequences of x's action, we can reasonably infer that bringing about A is not a means to B for x if x brings A about while believing it cannot lead to B or while believing there is an easier or more efficient way C to gain B.[272] This schema is important for testing psychological egoism, since the egoist's claim is that benefiting others (paradigmatically relieving the suffering of others) is never the ultimate goal, but always the means to some benefit to oneself. Thus we can test various egoist hypotheses by removing the postulated benefits to self or adding an easier way to get them, and then observing whether the subject goes on nonetheless to provide benefit to others. This method is necessarily inconclusive, since there could, in principle, always be some other benefit to self that the action gives rise to. But given general assumptions about human psychology, the number of plausible egoistic hypotheses is very limited. Batson and his colleagues have focused on three most common ones. The alternatives can be diagrammed as follows (simplified and modified from Batson 1991, 76):

[272] More precisely, we can infer that bringing about A is not *mere* means to B. As Aristotle already noted, it is possible to want something for its own sake *and* for the sake of something else; this is the relationship of virtues to *eudaimonia*.



Each story begins with the perception of someone else in need – they are, after all, alternative psychological hypotheses about the motivation to benefit others. The three first alternatives are egoistic, since in each case, helping another is only a means to some benefit to self. If rewards and punishments are understood as external, the first two hypotheses are very implausible – people certainly seem to benefit others without expecting an external reward for doing so or any kind of punishment for failing to do so. But there are also internal rewards and punishments, like empathic joy, guilt, and shame. These either cannot be had, or are most likely to be had, when we empathize with another person. Empathy, as such, can thus be a source of egoistic motivation.

Batson labels the hypothesis that empathy leads to helping another as a way of getting an internal reward the *empathy-specific rewards* hypothesis, and the view that empathy leads to helping another as a means to avoid an internal punishment like guilt the

empathy-specific punishments hypothesis.[273] The third hypothesis differs from the second in that what is avoided by helping is simply the unpleasant feeling caused by observing or thinking another's suffering, not the unpleasant feeling resulting from one's own failure to respond to it. This is known as *aversive-arousal reduction* hypothesis. For these egoistic hypotheses, empathy must be understood broadly to cover feelings aroused either by the feelings of others or the objective situation of others when it is such as to give rise to feelings in them. The first disjunct covers sharing the feelings that the other has (like feeling sad that John lost his job because John feels sad that he lost his job), while the second, which Sober and Wilson call 'sympathy', involves a feeling for the other that she does not necessarily have herself (like feeling sad for Elisa because her brother has died, even though she does not even know it yet).[274] The fourth alternative Batson labels the *empathy-altruism hypothesis* (EA). According to it, empathy (which is here narrowly understood as "feeling sympathetic, compassionate, warm, softhearted, tender, and the like"[275]) can motivate us to have benefiting another as our ultimate goal. In empirical testing, it functions as the baseline against which egoistic alternatives are tested. I will next briefly summarize some of the experimental results that suggest EA fares better than its egoistic competitors.

According to the aversive-arousal reduction hypothesis (AAR), we are motivated to help another in need, because we feel bad when witnessing someone else in need, and helping is a means to get rid of this bad feeling. The bad feeling is a kind of empathic arousal, since it results from the perception of another's need. AAR gives rise to a testable prediction. If the real reason for helping is getting rid of empathic aversive arousal, one will take an alternative means to that end (for short, escape) if one is available. More precisely, one can be expected to help at a high rate in an experimental situation if there is no escape, particularly if empathetic arousal is high. However, if there is an easy way to escape, one can be expected to take it whether

[273] For Batson's terms, see Batson 1991, chapter 7.

[274] Sober and Wilson 1998, 232–235.

[275] Batson 1991, 86.

empathetic arousal is high or low. By contrast, if EA is true, the ease of escape should make little difference to helping behaviour, since helping is not a means to reduce one's own arousal.

To test these predictions, one needs to be able to manipulate both the subjects' level of empathy and ease of escape. Batson and colleagues came up with the following sort of experimental setup to do this.[276] Participants are told that the object of the study is task performance in aversive conditions and its effects on the attitudes of others. Aversion is to be created by giving electric shocks to the 'worker'. At the beginning of the experiment, participants draw lots for the roles of worker (in fact, a confederate) and observer; the draw is rigged so that the subject always receives the role of the observer, watching the worker, Elaine, perform ten two-minute number-recalling tasks via television (in fact a pre-recorded videotape). After the first few trials, Elaine begins to look very distressed, and an assistant, Martha, goes to ask her what is wrong. Elaine explains that as a child she was thrown off a horse onto an electric fence, and was told that she might react strongly even to mild shocks in the future. Martha offers her a chance to quit, but she says she knows the experiment is important and wants to finish what she started. Then Martha hits upon an idea: maybe she could ask the observer to trade places with Elaine, and Elaine reluctantly agrees. A moment later, the experimenter comes to the observation room and tells the subject that Elaine is "finding the aversive conditions pretty uncomfortable" and that Martha was wondering whether the observer would like to trade places with Elaine, emphasizing that she has no obligation to do so, having drawn the observer lot.

The participants are in four different conditions: easy or difficult escape and high or low empathy. In *easy escape*, they are told before the experiment that they need to observe only two out of the ten trials, and after Elaine's distress that if they decide not to trade places, all they need to do is answer a few questions about their impression of Elaine and they may go. In *difficult escape*, they are told in advance that they have to watch all ten trials, and are reminded

[276] Batson 1991, 113–117.

again of this commitment when they decide whether to trade places. Thus, they can either go and forget about Elaine (easy escape) or watch her suffer (difficult escape). To create the empathy conditions, Batson and colleagues make use of previous studies indicating that people are likely to empathize with people who they think are similar to them.[277] To manipulate this, the subjects, who have several weeks earlier filled in a personality and values questionnaire, are handed before the experiment a copy of the same questionnaire allegedly filled in by Elaine. In *high empathy* condition, Elaine's answers are very similar to the subjects' own answers, in *low empathy*, very dissimilar.

The results can be summed up in the following table[278] indicating how many percent of the subjects in each condition agreed to trade places and take shocks for Elaine:

	Low empathy (dissimilar victim)	High empathy (similar victim)
Easy escape (told to leave)	18%	91%
Difficult escape (must stay and watch)	64%	82%

Recall that it follows from AAR that people would prefer an easy escape to costly helping (here receiving electric shocks), since helping is only a means to reducing one's own aversive arousal. So the proportion of helpers should be smallest in the easy escape/high empathy condition. But the results are the opposite: people with high empathy are most likely to help when they have an easy

[277] In another similar experiment, empathy was manipulated by placebo pills. In studies that were based on the subjects hearing a fake radio broadcast and being offered a chance to help the person in talked about, empathy was manipulated by asking participants in advance either to give an objective description of the elements of the story (low empathy) or think about what it is like to be the person talked about (high empathy).

[278] Based on Batson 1991, 116.

alternative means of reducing their own discomfort. This fits with EA, which thus receives corroboration from the experiment.

Of course, this does not suffice to reject psychological egoism or even AAR – perhaps the manipulations fail in some way (for example, escape does not reduce aversive arousal or similarity does not increase empathy). Batson and his colleagues have therefore, first of all, varied the AAR experiments in several ways, such as manipulating empathy by placebo pills or perspective-taking instead of similarity, but the results have been the same – if people empathize with others, they choose to help rather take the easy escape.[279] One frequent criticism has been that perhaps the escape is not so easy: after all, one might reasonably feel guilty for walking out on someone receiving electric shocks. This criticism is in effect the empathy-specific punishments hypothesis (ESP). It predicts that people are more likely to help others when they would feel guilty otherwise. To test it, one must therefore create conditions that are otherwise similar but differ in the likelihood that guilt is aroused. This is far from trivial. One way Batson and colleagues tried to do this is varying the availability of justification for not helping – after all, it is easy to rationalize guilt away if there is some plausible justification for inaction: "Even those who reflexively slap themselves with guilt and self-recrimination whenever they do wrong are likely to be sensitive to situational cues in determining when they have done wrong"[280]. They reasoned that people will be more likely to think it is all right not to help if they believe that others are not helping either, and more likely to think helping is the thing to do if they think that most others are doing it. As it turned out, providing a rationale for avoiding guilt did not significantly reduce helping behaviour (only 10% fewer participants offered help in the high justification/low-guilt condition); again, subjects behaved in accordance with EA rather than the egoist competitor ESP. Other manipulations of guilt included offering the opportunity to attribute refusal to help to features of the helping task itself – if

[279] See Batson 1991, 118–127.

[280] Batson 1991, 135.

what one must do to help is difficult, one should be less likely to feel guilt for failing to do it.

A different approach is to make use of the so-called emotional Stroop task.[281] It is based on the observation that people are slower to name the colours of words that are related to their emotional state or preoccupation than those of neutral words – for example, an alcoholic is slower to name the colour of the word "whiskey" than the colour of the word "window". A broadly accepted explanation is that this is a result of an attentional bias that interferes with the task, even though one is meant to ignore the semantic content of the words. Making use of this paradigm, Batson and colleagues had subjects in a helping task name the colours of both punishment-relevant words ("guilt", "duty", "should", "shame") and victim-relevant words ("hope", "child", "needy", "friend").[282] EA would predict that empathic helpers would be at some level thinking about guilt rather than the needs and state of the person to be helped, so they should be slower at the colour-naming task for punishment-relevant words than for neutral or victim-relevant ones. Once again, this turned out not to be the case, indicating that people did not have avoiding punishment in mind while helping.

I lack the space to discuss all the studies testing egoistic hypotheses and their variants here. Suffice it to say that they have not fared well in psychological testing. This does not yet, of course, show that egoism is false. For one thing, the experimental setups are far from perfect.[283] But bearing in mind the uneven argumentative burden – the egoist must show there is no altruistic motivation, while the altruist does not deny that there is egoistic motivation – we would need a very strong *a priori* reason to believe in egoism to embrace it. Some egoists have thought that they would find it in

---

[281] See Williams, Mathews, and MacLeod (1996). The original "Stroop effect" was that people are slower to name the colours of words if the words have an incongruent meaning (like the word "green" written in blue ink).

[282] Batson et al. 1988.

[283] For example, Stich, Doris, and Roedder (MS) offer a number of criticisms of Batson's work from a philosophical perspective. The problems they raise do not seem particularly serious, however.

evolutionary theory. After all, how could altruism survive in nature red in tooth and claw? But this line of argument turns out to rest on a confusion of levels. It is genes, not individuals, that get selected for, and it is not at all unusual in nature that the 'interests' of the two fail to coincide. Most obviously, genes that dispose a parent to sacrifice itself for its offspring may well spread in a population when the offspring of non-sacrificers gets eaten. There can be no doubt that costly helping or behavioural altruism exists in spades in nature.[284] Since the proximal mechanisms that govern intelligent behaviour in human beings are psychological states like beliefs and desires, there is good reason to believe that natural selection would favour dispositions to be non-instrumentally concerned for at least those who share one's genes.[285] This, of course, would be a very limited sort of altruism, but there are good reasons to believe that other sorts of prosocial behaviour (and so the motivational structures that promote it) also serve the interest in survival and reproduction.[286]

---

[284] The existence of 'evolutionary altruism' *is* open to question, but also, fortunately, irrelevant to the philosophical and psychological questions.

[285] Sober and Wilson point out that in principle, two kinds of mechanisms for taking care of offspring might have evolved: parents could have an ultimate desire for the well-being of their offspring, or they could get a lot of pleasure from the prospering of their offspring and be hedonists (Sober and Wilson 1998, 312–313). The latter mechanism would be egoistic (since the well-being of the children would be a means to the parents' pleasure). Sober and Wilson argue that the egoistic mechanism would be more unreliable than direct concern for children (because direct concern requires only a belief that children are in risk of being harmed to lead to action, while the egoistic mechanism requires also an additional belief that harm to children would be unpleasant), so the latter, psychologically altruistic alternative is likely to have been selected for (Sober and Wilson 1998, 316–319).

[286] As Richard Joyce puts it, "If kin selection gave our distant ancestors the psychological and physiological structures needed for regulating helpful behavior toward family members, then those structures became available for use in new tasks – most obviously, helpful behavior toward individuals outside one's family – if the pressures of natural selection pushed in that direction." (Joyce 2006, 22) Joyce goes on to list a number of sources of such

We can, then, be fairly confident that psychological egoism is false, since nobody has come up with the sort of strong evidence that would be required for us to reject the appearance (an Aristotelian *endoxon*, surely) that at least some people sometimes have non-derivative concerns for the good of others. Normative ethical theories that require some degree of altruistic motivation have not been shown to be psychologically unrealistic.

Is Virtue Possible?

The distinctive feature of virtue ethics, as opposed to its main competitors, deontology and consequentialism, is its focus on *agents* rather than individual actions. This is not to say that deontologists and consequentialists do not have anything to say about character or that virtue ethics does not have anything to say about individual actions, just that there is a particular emphasis on developing character traits and becoming a good person in virtue theorists. Within normative ethics, there has long been a healthy debate about the need for a specific virtue theory and the best form for it. In recent years, however, virtue ethicists have had to face a somewhat surprising external challenge: do the sort of character traits they think we should have exist at all? To answer this question, we need to first understand what character traits are meant to be.

If everybody behaved the same (similar) way in same (similar) situations, we would have no use for concepts of character and personality.[287] We could still predict and explain their behaviour and

---

pressures, such as the benefits of joint action, direct and indirect reciprocity, and group selection.

[287] 'Sameness' is a bit tricky here. Obviously, no two situations are the same in the sense of being numerically identical, nor are they perfectly similar in the sense of being qualitatively identical. Nonetheless, it seems perfectly unproblematic to classify situations into repeatable types on the basis of contextually salient features. Thus, in one context, you and I can be in the same situation if we both discover at the office that we've left the keys home; in another context, such discovery could mean that our situations are

even patterns of behaviour, but only on the basis of situations themselves and patterns in them. If everybody behaved in a completely random way, we would have no more use for character concepts; nothing except particular circumstances would predict or explain people's behaviour. However, neither of the antecedents is true: people do behave differently in relevantly similar situations, and not in a random way either. The mind of an adult (or even a child) is not a blank slate when it comes to dispositions to perceive and respond to practically relevant changes in the environment. The set of such dispositions constitutes what we call the personality of a person. Character traits are plausibly a subset of personality traits, perhaps distinguished by their prima facie evaluative significance.[288] Both personality and character traits are, at a familiar level of abstraction, shareable with others – you and I can both be fidgety Bible-thumpers (sharing personality traits) or stingy cowards (sharing character traits). Personalities and characters are (as a matter of fact) unique on two dimensions, as it were: horizontally (they have a *different combination* of in principle shareable traits) and vertically (they possess the same traits in *different ways* – your honesty may be different from my honesty). It is conceivable, though empirically unlikely, that two people would have identical personalities. They would still behave differently, however, as long as they faced different situations.

What kind of dispositions are character traits? The recent literature on the empirical adequacy of virtue ethics highlights a number of distinctive features that character traits are traditionally taken to have.[289] They are meant to be *robust*, that is, resistant to situational pressures to the contrary. *Ceteris paribus*, an honest person will not lie even when she could do so with impunity and to a great personal advantage. They are *cross-situationally consistent*, that is, they manifest themselves in a variety of trait-relevant situations across different contexts (home, workplace, and so on). An honest

---

very different – your husband is waiting for you anyway, while nobody has a spare key to the reinforced concrete door to my cozy bunker.

[288] This is how Goldie (2004) draws the distinction.

[289] For the following, see especially Doris 1999.

person will tell the truth in a host of different situations that call for telling the truth and, conversely, refrain from lying or misleading in all (or most) of those situations. Nor will she steal or cheat. Character traits are also *stable* over time, not changing day by day. And finally, as talk of situations that 'call for' a particular kind of response suggests, they are *responsive to reasons*. An honest person will recognize when there is reason to tell the truth and act for that reason (as well as taking it into account in deliberation, if that is needed). If she finds that being silent would mislead her conversational partner, she sees this as a strong non-instrumental reason to speak up. Typically, this does not involve thinking in virtue terms ("That would be the honest thing to do, so I will do it"), but rather simply seeing the relevant features as favoring or disfavoring certain courses of action ("It would be good for me if he believed that, but it just isn't true, so forget about it").

Recently, John Doris (1999, 2002), Gilbert Harman (1999b) and others have argued that there are no robust, cross-situationally consistent, stable, and reasons-responsive dispositions that would play a role in the explanation of behavior, and thus no character traits in the 'global' sense that virtue ethicists need.[290] At most, there are very local dispositions of this sort, like being "dime-finding, dropper-paper compassionate"[291]. The basic thesis of this *situationist* alternative is that "[b]ehavioral variation across a population owes more to situational differences than dispositional differences among persons."[292] Thus, situationists do not deny that some people tell the truth more often than others, for example. The claim is simply that what *explains* this is that these people are placed in certain kinds of

---

[290] I am leaving out of discussion one of the features of character traits that Doris discusses, namely evaluative consistency, the thesis that if a person has one positive or negative character trait (like generosity), she is also likely to have others of same valence (like compassion) (Doris 1999, 506). This is obviously related to the unity of virtue thesis, which not all virtue ethicists endorse.

[291] Doris 1999, 514. For context, see the discussion of Isen and Levin 1972 below.

[292] Doris 2002, 24.

situations (in which truth-telling is easy or advantageous or whatever) more often than others, and *not* something about their character. Doris's basic argument for this is a modus tollens on the basis of a hypothesis about the nature of character traits and the empirical data (Doris and Stich 2006):

1. If behaviour is typically ordered by robust traits, systematic observation will reveal pervasive behavioural consistency.
2. Systematic observation does *not* reveal pervasive behavioural consistency.
3. Therefore, behaviour is *not* typically ordered by robust traits.

The first premise is a little too simple, insofar as character traits manifest themselves first in perceptions of reasons rather than behaviour, but *if* the demands of particular situations are consistent and *if* agents are moved to act by their perceptions of reasons – two major qualifications – we can still expect behavioural consistency.[293] As to the second premise, social psychologists have produced a large number of studies purporting to show 'the power of the situation'. I will next briefly describe three well-known studies in this vein and then take a closer look at whether they support Doris and Harman's contention.

One class of studies concerns the effects of situations on helping. In Isen and Levin's (1972) study, the experimenters left a dime in a public telephone in a shopping mall, so that the next person using the phone (the unwitting experimental subject), would have the nice surprise of a free call.[294] The control group were people using the same telephone without finding a dime. When the subject left the

[293] Kamtekar (2004, 474) is therefore not charitable enough to situationists when she claims that they "treat all the traits on the model of aggression: people who possess a given trait are expected, to the extent that they possess the trait, to behave spontaneously and unreflectively in ways that manifest it on every occasion."

[294] For the following, see Isen and Levin 1972, 386–387.

phone booth, a female confederate walking ahead dropped a manila folder full of papers, apparently accidentally. The dependent measure was whether the subject would help the confederate pick up the papers or not. 14 out of 16 people who found a dime did stop to help, but only 1 out of 25 who did not find a dime stopped. As Isen and Levin interpret the results, what explains the subjects' behaviour is good mood occasioned by the unexpected and convenient find.

Latané and Rodin (1969), in turn, were interested in the effect of the presence of other people (the 'bystander effect') on helping behaviour. In their experiment, college students were asked to fill in a questionnaire on games by a female 'marketing researcher'. While they were working on it, the 'marketing researcher' went to her office in the next room, and after a few minutes was heard to first climb on a chair to reach for a stack of papers and then fall with loud crash, saying "Oh my God, my foot… I… I… I can't move it […] I… can't get this… thing… off me", and moaning loudly.[295] (This was in fact a tape recording, which most of the subjects did not realize.) The variable here was whether subjects would go and help. Subjects were tested in four conditions: filling the questionnaire alone, with a passive confederate (who did not go and help), with another test subject, and together with a friend. Afterwards, they were asked why they acted as they did. Again, the results were striking: 70% of the subjects who were alone helped, while only 7% of those with a passive confederate did anything.[296] Yet when asked, most subjects tested with a confederate reported that the presence of another person had "very little" influence on their behaviour. Latané and Rodin offer two possible explanations for these results. First, the 'social influence' explanation is that faced with an ambiguous situation (What are those noises? Am I meant to do something?), people look to other people's reactions for guidance. When the other

[295] The experiment is described in Latané and Rodin 1969, 191–192.

[296] Latané and Rodin 1969, 193. When two subjects unknown to each other were paired, at least one helped in 40% of the cases, and at least one of two friends helped in 70% of the cases (suggesting that the presence of a friend neither hindered nor increased helping) (ibid., 195–196).

person (the confederate) is unconcerned, they conclude that the situation is not so bad.[297] This does not exclude a second explanation, which is simply that when more than one person is present, responsibility for inaction gets diffused – there is less of a pressure for *me* to do something about the situation.[298]

The final series of studies I want to look at is Stanley Milgram's justly famous obedience experiments. In the original experiment (Milgram 1963), forty men of various socioeconomic groups were paid to participate in an experiment on 'punishment and learning'.[299] Each subject was paired with another 'volunteer' (in fact a confederate), and assigned the role of a 'teacher' by a rigged drawing of lots. The teacher's ostensible task was to give multiple choice tasks involving memorizing word pairs to the learner, and punish the learner with an electric shock every time he made a mistake (or failed to respond), crucially increasing the level of the shock with each mistake. He was shown how the learner was strapped to a chair and electrodes applied to his wrist. The teacher also received a sample shock of 45 volts to convince him of the authenticity of the device he was to operate. The device itself had 30 lever switches marked from 15 to 450 volts, with additional verbal designations from 'Slight Shock' to 'Danger: Severe Shock'. The teacher was told that none of the shocks result in permanent tissue damage, though they may be painful.

During the run of the experiment, the 'learner', situated out of sight in a separate room and responding by way of a signal box, gave wrong answers about 3/4^{ths} of the time. In the original experiment, the learner did not communicate otherwise, except by banging the wall at the 300 volt level and ceasing to give any sign after 315 volts. In later experiments, the learner made other noises and comments.[300] For example, at the Strong Shock level (135-180

---

[297] Latané and Rodin 1969, 198–199. They also hypothesize that since Americans like to appear poised and unconcerned in public, people are mutually led astray by the apparent lack of concern of each other.

[298] Latané and Rodin 1969, 199.

[299] For a description of the setup, see Milgram 1963, 373–374.

[300] See Doris 2002, 41–42.

volts), he said "Ugh!!! Experimenter! That's all. Get me out of here. I told you I had heart trouble.". At Intense Shock (255-300 volts), the reactions included agonized screaming and "I absolutely refuse to answer any more". Extreme Intensity Shock (315-360 volts) resulted in an "intense and prolonged agonized scream", "My heart's bothering me", and hysterical repetition of "Let me out of here", followed by silence at 345 volts. In these studies, if the teacher did not shock the learner after a wrong response, the experimenter, who was with the teacher, prodded him along in a firm but not impolite tone according to a predetermined series from "Please continue" to "You have no choice, you *must* go on".

The results of the first study came as a surprise to everyone involved. Here are some of them:

- no subject stopped before 300 volts, labeled "Intense Shock" ("At this level … the victim kicks on the wall and no longer provides answers to the teacher's multiple-choice questions.")
- 14 subjects stopped between 300 and 375 volts (the latter labeled "Danger: Severe Shock")
- 26 subjects went on to the end of the scale (450 volts, labeled "XXX")

These results have been replicated in a number of follow-up studies. On average, two thirds of subjects go to the end of the scale.[301] And the results cannot be explained away by the subjects' seeing through the experimental setup. Quite the contrary; when asked after the experiment to rate the strength of the last shocks they gave on a 14-point scale from "not at all painful" to "extremely painful", the subjects mean estimate was 13.42 (mode 14). Their manner of behaviour during the experiment matched this belief: "Subjects were observed to sweat, tremble, stutter, bite their lips, groan, and dig their fingernails into their flesh."[302]

[301] Doris 2002, 45.

[302] Milgram 1963, 375.

These kinds of results, Doris and Harman claim, call into question the existence of character traits in the sense that virtue theories must assume. For this to be so, virtue theory must predict that people would not act in the observed ways in the situations. This requires that the following two assumptions hold in each case:

a) Virtue demands behaviour x, not the observed behaviour y
b) The number of participants who would be considered virtuous is significantly higher than the number of participants who actually behave in the way x

Assumption b is required, since, as Doris acknowledges, if virtue is very rare, the results of the studies are precisely what virtue theory would predict – only a few people do the kind thing when conditions are not conducive to it. Let us assume, then, for the time being that b is true in each of the cases. What about a? To begin with, it is not obvious that kindness *demands* one to stop and pick up the papers of clumsy people while you are busy shopping or going home yourself. It would be kind, to be sure, and there is some reason to do it, as well as some reason not to. Failing to help does not mean you are callous or selfish. In fact, the case looks like many of automaticity studies: there is a near balance of reasons for action, and it is tipped in one way or another by a situational factor, in this case a lucky mood triggered by the discovery of a coin. Just minding your own business and walking past seems here compatible with possessing an ordinary degree of kindness. A really kind or helpful person – the sort of person we make a point of calling kind or helpful – might give more weight to this sort of considerations, and we would expect her to be more likely to help in the experimental situation. But that is consistent with the data. Virtue is a matter of degree, and for most virtues, possessing them to a high degree is indeed rare and something to aspire to for most of us.[303]

[303] Cf. Sreenivasan 2002, 57.

At the other extreme with regard to the demands of virtue are Milgram's studies. Here there is no equivocation: however we understand virtue, it demands one to stop participating in the experiment way before XXX. And it seems that the subjects recognize that, too. Here is how Milgram himself describes them:

> It is clear from the remarks and outward behavior of many participants that in punishing the victim they are often acting against their own values. Subjects often expressed deep disapproval of shocking a man in the face of his objections, and others denounced it as stupid and senseless. Yet the majority complied with experimental commands. (Milgram 1963, 376)

There is thus little doubt that the participants perceived the suffering of the learner as calling for them to stop. These people are not sadists, but ordinary, decent citizens. Yet they go on, giving in to fairly mild pressure exerted by the experimenter. They have some excuses available – up to a point the learner does give answers in spite of the shocks, indicating a willingness to go on with the experiment, and of course there is some level of trust in the institution involved; surely a university researcher would not ask them to do something really dangerous to another person. But these really are excuses: the evidence available to them strongly points to serious pain and fatal risk. So what is going on? The best explanation seems to be that the situational pressures inhibit the move from perception of reasons and ought-judgments based on them to action. As Allan Gibbard notes, this is a kind of weakness of will. He suggests that the participants are in the grip of norms of politeness and cooperation with the experimenter, and thus give in against their best judgment.[304] From a virtue theory perspective, the participants – and, we may infer from the studies, most of us – are lacking in executive virtues. The moral failure at play is not in the first instance lack of compassion but lack of courage.

What the Milgram studies, as well as, to an extent, the Latané and Rodin study (which is not such a clear case, since the presence of

[304] Gibbard 1990, 59.

others and observation of their reactions both reduces helping and the strength of reasons to help) and many others like Zimbardo's Stanford Prison Experiment, point to is that virtue is a lot more *fragile* than we might have thought. Even when we recognize reasons, we do not always act on them if there are social pressures to the contrary. This inaction, in turn, may well lead us to rationalize away our original perception. This is an important lesson, and one that is relevant to both private planning and public policy-making.[305] Does it mean that there are no character traits or virtues? Hardly so. The experimental results show at best that the virtues of many of us are less robust, broad, and reasons-responsive than an optimist about human nature might have thought. It does not tell against somewhat the more localized and fragile dispositions whose existence is attested to by the lifetime of better-than-chance predictive success that most of us have enjoyed, particularly with people we have observed and interacted with in many different contexts.[306] The ideal articulated by virtue ethics, particularly as an ideal to aspire to, has not been shown to fall foul of the constraint of minimal psychological realism.

---

[305] Zimbardo (2007), for example, argues that studies like his predict the sort of abuse that occurred (and occurs) in Abu Ghraib and other American military prisons, and that the scandals could and should thus have been avoided by putting the social psychological knowledge to use.

[306] Sreenivasan emphasizes that our tendency to rush into character judgments on the basis of meager evidence (such as a few observations of an individual's behaviour in a particular context) and ignore the effect of the situation in which the person is (the 'fundamental attribution error' in the language of social psychology) does not as such count against the existence of character traits: "If we suppose that a trait has been attributed without warrant, it will come as no surprise to learn that the predictions it licenses are frequently confounded" (Sreenivasan 2002, 54). Poorly grounded predictions will indeed fail: if I think Jack is honest just on the basis of having seen him give a dropped ten-dollar note back to its owner, it is my own stupidity if he walks away with a million-dollar loan I gave him. Virtue theory predicts predictive success only on the basis of thorough acquaintance and, post-Milgram, under normal background conditions (cf. Sreenivasan 2002, 66).

Essay 4: 'Reason, Recognition, and Internal Critique'

The question at the heart of 'Reason, Recognition, and Internal Critique' is a very old one: how do you rationally persuade someone to accept a moral norm she does not at present accept? One answer comes from a minority philosophical tradition going back at least to Socrates and Aristotle and resurfacing with Hegel and Marx: by showing that the person or group in question is in fact already committed to accepting it. This, in brief, is what internal critique amounts to. In the final (and oldest) paper of my dissertation, I examine the advantages and different varieties of internal critique, as well as the role that psychological facts having to do with the importance of recognition can play in it.

I begin with a look at the logical space of possible sources of standards of normative criticism. (In the paper, I present the issue at the social level, but *mutatis mutandis*, the same points apply to individuals.) One could criticize another society's practices on that basis of one's own norms as such, objective norms like natural law, or the society's own norms. The first two are forms of external criticism. Appeal to the mere fact that *we* find, for example, universal healthcare to be a necessary component of a decent society is highly unlikely to persuade someone who disagrees, nor should it. In contrast, if there are objective reasons (and I am less sceptical about them now than I was when I wrote the paper), they do, obviously, carry normative weight with everyone. But the other party to the debate will probably have a different view about objective reasons, and it is not easy to show that one is in an epistemically superior situation. If it is possible to show that the target society's own fundamental normative commitments already require, say, providing taxpayer-funded healthcare to everyone, concerns about the status and accessibility of objective normative truths are bypassed. In addition, the criticism is more likely to be effective in practice, since some kind of motivation to comply with the norms already exists. Internal critique is therefore preferable to external criticism both in justificatory and pragmatic terms.

Internal critique, too, can take several forms. I distinguish between *simple* internal critique, which draws on the explicit commitments of a society or group and aims to show that they contradict some policy or practice, and *reconstructive* internal critique, which begins by making explicit commitments that are implicit in a society or group's practice, and then aims to show that they are not consistent with a particular policy or practice. Implicit norms manifest themselves in unarticulated emotions and informal sanctions, for example, and may form an unstated basis for customs and norms. Within reconstructive internal critique, I further distinguish weak and strong forms. Weak reconstructive critique draws on contingent implicit norms of a society, whereas strong reconstructive critique articulates normative commitments that are unavoidable, that any social group must undertake to some degree to function and reproduce. The critical theories of Habermas and Honneth are both forms of strong reconstructive internal critique, and as such more ambitious that simple critique of ideology.

As already discussed in 'The Social Dimension of Autonomy', Honneth argues that social psychological research supports the quasi-Hegelian view that taking up the sort of attitudes toward ourselves that enable autonomous agency and forming a personal identity requires, as a matter of empirical fact, that others express corresponding attitudes toward us. (The Hegelian twist is that for these attitudes of others to matter to us, we must, in turn, recognize them – respect from someone I do not respect does not help my self-respect.) For him, this is an anthropological fact about becoming and being a person. This *need* for recognition gives rise to a normative *demand* for recognition. Pre-philosophically, to be sure, we are not aware that it is recognition we are looking for, but that goal and its normative status for us can be read in our emotional reactions to failures of recognition. These negative moral emotions manifest the expectation we have to be treated as persons, and, when systematic, motivate action for social change. (There are presumably positive emotions in response to achieving recognition.) I find Honneth's basic idea here plausible, even exciting, but point out that we must be more careful when looking at moral emotions as indicators of lack of recognition: sometimes people have them because of false beliefs

when there is no failure of recognition, and sometimes lack them because of ideology even though there is misrecognition.

If Honneth and the Hegelian tradition in general are right, every society must by 'anthropological necessity' already be implicitly committed to some norms of mutual recognition that we can reconstruct from their practices and emotional reactions. Internal critique can then proceed on the basis of pointing out the contradiction between a particular practice or explicit norm in force in the society and the norms of recognition. But, I argue in the paper, there are two challenges even if we accept all this. First, there is the question of why norms of recognition should be given priority when they conflict with other norms of the society (I call this the Priority Challenge). Second, there is a further question about what reason the society or group in question (or those in power within it) has to extend the application of recognition to those from whom they do not need it (the Application Challenge). On Honneth's behalf, I suggest that a critic can appeal to the functional importance of recognition to show that the norms related to it are more fundamental than most other norms. As to the Application Challenge, I argue that the best response to it will appeal to further moral considerations. Honneth himself appeals to the value of 'self-realization', which, however, threatens to limit the scope of internal critique to those for whom self-realization is a fundamental value. I favour coherentist arguments instead. Armed with knowledge about the empirical importance of recognition, we can try to convince the members of the criticized society that were they to reach wide reflective equilibrium, they would find that they have no normative basis for withholding recognition from outsiders and outcasts. This kind of internal critique can lead to a new normative self-understanding by locating and amplifying forces of change within the criticized society.

Though the argument of 'Reason, Recognition, and Internal Critique' is formulated in terms of dialectical argumentation – from one perspective, it is an attempt to show how to defeat relativism without assuming superior epistemic access to moral values – the same psychological facts can be employed in more straightforward normative theorizing. Just like utilitarians must know what as a

matter of fact promotes happiness to derive concrete moral principles and judgments from the abstract goal of maximizing general welfare, Kantians and other liberals must know what as a matter of fact promotes and respects the autonomy of persons to derive concrete principles and judgments from the abstract goal of promoting everyone's autonomy to the extent that it does not conflict with the autonomy of others. The normative remarks at the end of 'The Social Dimension of Autonomy' adopt this simpler strategy.[307]

[307] In his response to my paper (published in the same issue of *Inquiry*), Honneth seems to adopt a similar strategy, and so making what he calls self-realization the value on which the interest in recognition depends.

Conclusion

At the beginning of this introduction, I laid out what I consider to be the three main questions of philosophical moral psychology:

1. What are the necessary psychological conditions for making moral judgments – that is, what is the nature of moral thinking?
2. What are the necessary psychological conditions for being morally responsible and thus fit to be praised and blamed?
3. What are the implications of facts about human psychology for normative ethical theory?

The distinctively philosophical approach to the first two questions has been to begin with an *a priori* investigation of the concepts and conditions of possibility involved. Recently, experimental philosophers have challenged this, and conducted empirical studies to discover *a posteriori* what ordinary people take to be necessary for moral judgment or moral responsibility. One response to this would be to reject the importance of folk concepts, but I agree with the experimentalists that this is a costly move. Philosophers need to stick closely to what ordinary people mean by moral thinking or responsibility if their work is to have any relevance. In 'The Rise and Fall of Experimental Philosophy', I argue that survey studies cannot, for reasons of principle, achieve their goal, and traditional methods of reflection and dialogue can. This deflects a major challenge to philosophical moral psychology.

When we move to substantial issues, however, further challenges to philosophical reflection arise. There is a constant theme to much recent work in empirical moral psychology, whether it concerns the

process of moral judgment, the kind of control required for moral responsibility, or the existence of moral virtues. Study after study claims to show that the true causes of our judgments and actions are hidden from us, that our self-understanding as rational, responsible agents is an illusion. Since much of philosophical moral psychology consists in articulating that self-understanding, these results threaten to reduce it to irrelevance. However, a careful survey of these studies shows that their claims are inflated for two different reasons. First, neither our commonsense self-understanding nor philosophical views are nearly as naïve as the psychologists like to paint them. As I have tried to show in this introduction, major philosophical traditions acknowledge the importance of emotional and automatic processes in moral judging and other decision-making alongside conscious reasoning. Second, the conclusions that can be legitimately drawn from the experimental data are far less dramatic than empirical researchers themselves suggest. As far as the studies show, non-rational situational cues and unacknowledged emotions play the biggest role at the margins and when nothing very important is at stake. To put it from a different perspective, the empirical accounts that purport to undermine our self-understanding fail to explain the felt authority of morality and the fact that our choices in general are intelligible in light of consciously held values and goals.

It may be appropriate to finish this introduction by highlighting the answers that I give to the basic question of philosophical moral psychology in the papers that comprise the substantial part of this dissertation:

1. What are the necessary psychological conditions for making moral judgments – that is, what is the nature of moral thinking?

To think that one morally ought to do something is to locate oneself in the space of reasons, understood as a web of commitments and entitlements. It is to take anyone in a relevantly similar situation to be entitled to do the same (unless pre-empted by prior commitments) and to take anyone to be entitled to sanction oneself negatively unless

one acts in the way in question. The first-personal moral judgment is true if one's deontic status really is as described. It motivates an agent to act accordingly insofar as she is rational in the sense of responding to the acknowledgment of commitments and entitlements in the appropriate way.

2. What are the necessary psychological conditions for being morally responsible and thus fit to be praised and blamed?

Autonomous, fully morally responsible agents must be at least moderately capable of recognizing desire-independent reasons for action, giving them appropriate weight in deliberation, and being motivated accordingly. To exercise these capacities, they must have self-confidence, self-respect, and self-esteem, which appears as a matter of empirical fact to require standing in relationships of recognition with other agents.

3. What are the implications of facts about human psychology for normative ethical theory?

Especially in dialectical moral argumentation, psychological facts can serve as a lever to coax the opponent into modifying her position. For example, since human beings can as a matter of fact become autonomous and morally responsible agents only through standing in relations of mutual recognition to other agents and institutions, any society or normative theory for which autonomy is a central value must also grant the importance of love, rights, and social esteem, and treat measures that promote them as increasing rather than compromising autonomy.

None of these answers pretends to be a comprehensive response to the question. A single article can contribute only so much to a

collaborative project that philosophers have been engaged in for several thousand years. I only hope to have nudged these debates forward a little.

References

Anscombe, Elizabeth (1958a), *Intention*. Harvard University Press, Cambridge, Mass.

Appiah, Kwame Anthony (forthcoming), *Experiments in Ethics*. Harvard University Press, Cambridge, Mass.

Aquinas, St. Thomas, *Summa Theologica*. Tr. Fathers of the English Dominican Province. Online at http://www.newadvent.org/summa.

Aristotle (1999), *Nicomachean Ethics*. Second edition, ed. and tr. Terence Irwin. Hackett, Indianapolis.

Arpaly, Nomy (2003), *Unprincipled Virtue*. Oxford University Press, Oxford.

Arpaly, Nomy (2007), *Of Merit, Meaning, and Human Bondage*. Princeton University Press, Princeton.

Audi, Robert (1989), *Practical Reasoning*. Routledge, London.

Audi, Robert (2004), *The Good in the Right*. Princeton University Press, Princeton.

Bargh, John (1994), 'The Four Horsemen of Automaticity: Awareness, Efficiency, Intention, and Control in Social Cognition'. In J. R. S. Wyer & T. K. Srull (Eds.), *Handbook of Social Cognition*, 2nd edition. Erlbaum, Hillsdale, 1–40.

Bargh, John A, Chen, Mark, and Burrows, Lara (1996), 'Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action'. *Journal of Personality and Social Psychology* 71 (2), 230–244.

Bargh, John A. and Ferguson, Melissa J. (2000). 'Beyond Behaviorism: On the Automaticity of Higher Mental Processes'. *Psychological Bulletin* 126, 925-945.

Bargh, John A., Gollwitzer, Peter M., Lee-Chai, Annette, Barndollar, Kimberly, and Trötschel, Roman (2001), 'The Automated Will:

Nonconscious Activation and Pursuit of Behavioral Goals'. *Journal of Personality and Social Psychology* 81 (6), 1014–1027.

Batson, C. Daniel, Dyck, Janine L., Brandt, J. Randall, Batson Judy G., Powell Anne L., McMaster M. Rosalie, and Griffitt Cari (1988), 'Five Studies Testing Two New Egoistic Alternatives to the Empathy-Altruism Hypothesis'. *Journal of Personality and Social Psychology* 55, 52– 77.

Batson, C. Daniel (1991), *The Altruism Question: Toward a Social-Psychological Answer*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Bealer, George (2003), 'Modal Epistemology and the Rationalist Renaissance'. In Tamar Szabó Gendler and John Hawthorne (eds.), *Conceivability and Possibility*. Oxford University Press, Oxford, 71–126.

Blackburn, Simon (1998), *Ruling Passions: A Theory of Practical Reason*. Clarendon Press, Oxford.

Blair, James (1995), 'A Cognitive Developmental Approach to Morality: Investigating the Psychopath'. *Cognition* 57, 1–29.

Blair, James (2006), 'The Emergence of Psychopathy: Implications for the Neurophysiological Approach to Developmental Disorders'. *Cognition* 101, 414–442.

Boyd, Richard (1988), 'How to be a Moral Realist'. In Geoffrey Sayre-McCord (ed.), *Essays on Moral Realism*. Cornell University Press, Ithaca, 181–228.

Bratman, Michael (1987), *Intention, Plans, and Practical Reason*.

Bratman, Michael (2000/2007), 'Reflection, Planning, and Temporally Extended Agency'. In Bratman 2007.

Bratman, Michael (2004/2007), 'Three Theories of Self-Governance'. In Bratman 2007.

Bratman, Michael (2005/2007), 'Planning Agency, Autonomous Agency'. In Bratman 2007.

Bratman, Michael (2007), *Structures of Agency. Essays.* Oxford University Press, Oxford.

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

Broadie, Sarah (1991), *Ethics With Aristotle*. Oxford University Press, Oxford.

Burnyeat, Myles (1980), 'Aristotle on Learning to Be Good'. In Amelie Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*. University of California Press, Berkeley, 69–91.

Chisholm, Roderick (1995), 'Agents, Causes, and Events: The Problem of Free Will'. In Timothy O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. Oxford University Press, New York, 95-100.

Chomsky (1981), 'Principles and Parameters in Syntactic Theory'. In Norbert Hornstein and David Lightfoot (eds.), *Explanation in Linguistics: The Logical Problem of Language Acquisition*. Longman, London, 1981.

Churchland, Paul (1996), 'The Neural Representation of the Social World'. In May, Friedman, and Clark (eds.) 1996.

Clark, Andy (1996), 'Connectionism, Moral Cognition, and Collaborative Problem Solving'. In May, Friedman, and Clark (eds.) 1996.

Cosmides, Leda and Tooby, John (2006), 'Evolutionary Psychology, Moral Heuristics, and the Law'. In G. Gigerenzer and Christoph Engel (eds.), *Heuristics and the Law*. MIT Press, Cambridge, Mass. 2006.

Cushman, Fiery, Young, Liane, and Hauser, Marc (2006), 'The Role of Conscious Reasoning and Intuition in Moral Judgment'. *Psychological Science* 17 (12), 1082–1089.

Damasio, Antonio (1994), *Descartes' Error. Emotion, Reason and the Human Brain*. Grosset/Putnam, New York.

Dancy, Jonathan (1993), *Moral Reasons*. Blackwell, Oxford.

Dancy, Jonathan (1999), 'Can the Particularist Learn the Difference between Right and Wrong?' In K. Brinkmann (ed.), *The Proceedings of the Twentieth World Congress of Philosophy, vol. 1: Ethics*. Bowling Green State University Philosophy Documentation Center, 59–72.

Dancy, Jonathan (2000), *Practical Reality*. Oxford University Press, Oxford.

Daniels, Norman (1979), 'Wide Reflective Equilibrium and Theory Acceptance in Ethics'. *Journal of Philosophy* 76, 256–282.

Dasgupta, Nilanjana and Greenwald, Anthony G. (2001), 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals'. *Journal of Personality and Social Psychology* 81 (5), 800–814.

Doris, John (1999), 'Persons, Situations, and Virtue Ethics'. *Noûs* 32 (4), 504–530.

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

Doris, John and Stich, Stephen (2006), 'Moral Psychology: Empirical Approaches'. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Online at http://plato.stanford.edu/entries/moral-psych-emp/.

Dreyfus, Hubert (1990), 'What is Moral Maturity? A Phenomenological Account of the Development of Ethical Expertise'. Online at http://socrates.berkeley.edu/~hdreyfus/html/papers.html.

Fehr, Ernst and Fischbacher, Urs (2003), 'The Nature of Human Altruism'. *Nature* 425, 785–791.

Fischer, John and Ravizza, Mark (1998), *Responsibility and Control*. Cambridge University Press, Cambridge.

Fischer, John Martin (2006), *My Way: Essays on Moral Responsibility*. Oxford University Press, New York.

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

Frankfurt, Harry (1969), 'Alternate Possibilities and Moral Responsibility'. *Journal of Philosophy* 66 (23), 829–839.

Frankfurt, Harry (1971), 'Freedom of the Will and the Concept of a Person'. *Journal of Philosophy* 68 (1), 5–20.

Gabennesch, Howard (1990), 'The Perception of Social Conventionality by Children and Adults'. *Child Development* 61, 2047–2059.

Gibbard, Allan (1990), *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Harvard University Press, Cambridge, Mass.

Greene, Joshua (forthcoming a), 'The Secret Joke of Kant's Soul'. To appear in Walter Sinnott (ed.) *Moral Psychology*. Vol. 3. MIT Press, Cambridge, Mass.

Greene, Joshua (forthcoming b), 'Reply to Mikhail and Timmons'. To appear in Walter Sinnott (ed.) *Moral Psychology*. Vol. 3. MIT Press, Cambridge, Mass.

Greene, Joshua and Haidt, Jonathan (2002), 'How (and Where) Does Moral Judgment Work?' *Trends in Cognitive Sciences* 6 (12), 517–523.

Greene, Joshua D., Nystrom, Leigh E., Engell, Andrew D., Darley, John M., Cohen, Jonathan D. (2004), 'The Neural Bases of Cognitive Conflict and Control in Moral Judgment'. *Neuron* 44, 389–400.

Greene, Joshua D., Sommerville, R.B., Nystrom, Leigh E., Darley, John M., & Cohen, Jonathan D. (2001), 'An fMRI Investigation of Emotional Engagement in Moral Judgment'. *Science* 293, 2105–2108.

Grotius, Hugo (1625/2005), *The Rights of War and Peace*. Ed. Richard Tuck from the edition by Jean Barbeyrac. Liberty Fund, Indianapolis.

Goldie, Peter (2004), *On Personality*. Routledge, London.

Gottlieb, Paula (2006), 'The Practical Syllogism'. In Kraut (ed.) 2006, 218–233.

Habermas, Jürgen (1983), *Moralbewusstsein und kommunikatives Handeln*. Suhrkamp, Frankfurt am Main.

Habermas, Jürgen (1996), *Die Einbeziehung des Anderen. Studien zur politischen Theorie*. Suhrkamp, Frankfurt am Main.

Haidt, Jonathan, Koller, Silvia H., and Dias, Maria G. (1993), 'Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?' *Journal of Personality and Social Psychology* 65, 613–628.

Haidt, Jonathan (2001), 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment'. *Psychological Review* 108 (4), 814–834.

Haidt, Jonathan (2003), 'The Emotional Dog Does Learn New Tricks: A Reply to Pizarro and Bloom'. *Psychological Review*, 110, 197-198.

Haidt, Jonathan and Björklund, Fredrik (forthcoming), 'Social Intuitionists Answer Six Questions About Moral Psychology'. To appear in Walter Sinnott (ed.) *Moral Psychology*. Vol. 3. MIT Press, Cambridge, Mass.

Hare, Richard Mervyn (1952), *The Language of Morals*. Clarendon Press, Oxford.

Hare, Richard Mervyn (1981), *Moral Thinking: Its Levels, Methods, and Point*. Oxford University Press, Oxford.

Harman, Gilbert (1975), 'Moral Relativism Defended', *Philosophical Review* 84, 3–22.

Harman, Gilbert (1999a), *Reasoning, Meaning, and Mind*. Clarendon Press, Oxford.

Harman, Gilbert (1999b), 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error.' *Proceedings of the Aristotelian Society* 99, 315-331.

Harman, Gilbert and Thomson, Judith Jarvis (1996), *Moral Relativism and Moral Objectivity*. Blackwell, London.

Hauser, Mark (2006), *Moral Minds*. Harper Collins, New York.

Hauser, Mark, Young, Liane, and Cushman, Fiery (forthcoming), 'Reviving Rawls' Linguistic Analogy'. In Walter Sinnott-Armstrong (ed.), *Moral Psychology and Biology*. Oxford University Press, Oxford.

Held, Virginia (1996), 'Whose Agenda? Ethics versus Cognitive Science'. In May, Friedman, and Clark (eds.) 1996, 69–87.

Helm, Bennett (2001), *Emotional Reason*. Cambridge University Press, Cambridge.

Hoffman, M. L. (2000), *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press, Cambridge.

Hubin, Donald (2001), 'The Groundless Normativity of Instrumental Rationality', *Journal of Philosophy* 98, 445–468.

Hume, David (1739-1740/2000), *A Treatise of Human Nature*. David F. Norton and Mary Norton (eds.). Oxford University Press, Oxford.

Hume, David (1751/1948), *Enquiry Concerning the Principles of Morals*. In Henry D. Aiken (ed.), *Moral and Political Philosophy*. Hafner Press, New York, 171–291.

Hursthouse, Rosalind (2006), 'The Central Doctrine of the Mean'. In Kraut (ed.) 2006.

Irwin, Terence (1980), 'The Metaphysical and Psychological Basis of Aristotle's Ethics'. In Amelie O. Rorty (ed.), *Essays on Aristotle's Ethics*. University of California Press, Berkeley, 35–53.

Irwin, Terence (1999), 'Notes'. In Aristotle 1999, 172 – 314.

Isen, Alice M. and Levin, Paula F. (1972), 'Effect of Feeling Good on Helping: Cookies and Kindness'. *Journal of Personality and Social Psychology* 21 (3), 384–388.

Johnson, Mark L. (1996), 'How Moral Psychology Changes Moral Theory'. In May, Friedman, and Clark (eds.) 1996, 45–68.

Johnston, Mark (1993), 'Objectivity Refigured: Pragmatism Without Verificationism'. In John Haldane and Crispin Wright (eds.), *Reality, Representation, and Projection*. Oxford University Press, Oxford, 85–130.

Joyce, Richard (2006), *The Evolution of Morality*. MIT Press, Cambridge, Mass.

Kagan, J. and Lamb, S. (eds.) (1987), *The Emergence of Morality in Young Children*. University of Chicago Press, Chicago.

Kane, Robert (2002a) (ed.), *The Oxford Handbook of Free Will*. Oxford University Press, Oxford.

Kane, Robert (2002b), 'Some Neglected Pathways in the Free Will Labyrinth'. In Kane (ed.) 2002a, 406–437.

Kane, Robert (2005), *A Contemporary Introduction to Free Will*. Oxford University Press, Oxford.

Kant, Immanuel (1785/1898), *Fundamental Principles of the Metaphysic of Morals*. In Kant 1898, 1–86.

Kant, Immanuel (1788/1898), *Critique of Practical Reason*. In Kant 1898.

Kant, Immanuel (1898), *Kant's Critique of Practical Reason and Other Works on the Theory of Ethics*. Tr. Thomas Kingsmill Abbott. Kongmans, Green, and Co., London.

Kant, Immanuel (1785/1996), *Groundwork of the Metaphysics of Morals*. In Kant 1996, 37–108. (*Groundwork*)

Kant, Immanuel (1788/1996), *Critique of Practical Reason*. In Kant 1996, 133–272.

Kant, Immanuel (1996), *Practical Philosophy*. Tr. Mary J. Gregor. Cambridge University Press, Cambridge.

Kamtekar, Rachana (2004), 'Situationism and Virtue Ethics on the Content of Our Character'. *Ethics* 114, 458–491.

Kauppinen, Antti (2006), 'Lovers of the Good: Response to Knobe and Roedder'. Online at http://garnet.acns.fsu.edu/~tan02/OPC%20Week%20Three/Commentary%20on%20Knobe.pdf. Retrieved April 10, 2007.

Kauppinen, Antti (in preparation), 'Kind Words and Cruel Facts'.

Kelly, Daniel, Stich, Stephen, Haley, Kevin J., Eng, Serena J., and Fessler, Daniel M. T. (forthcoming), 'Harm, Affect, and the Moral/Conventional Distinction'. *Mind and Language*.

Knobe, Joshua and Nichols, Shaun (forthcoming), 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions'. *Nous*.

Knobe, Joshua and Roedder, Erica (2006), 'The Concept of Valuing: Experimental Studies'. Online at http://garnet.acns.fsu.edu/~tan02/OPC%20Week%20Three/knobe.pdf Retrieved April 10, 2007.

Koenigs, Michael, Young, Liane, Adolphs, Ralph, Tranel, Daniel, Cushman, Fiery, Hauser, Marc, and Damasio, Antonio (2007), 'Damage to Prefrontal Cortex Increases Utilitarian Moral Judgments'. Forthcoming in *Nature*.

Korsgaard, Christine (1996), *Sources of Normativity*. Cambridge University Press, Cambridge.

Kohlberg, Lawrence (1981), *Essays on Moral Development, Vol. 2: The Psychology of Moral Development*. Harper&Row, San Fransisco.

Korsgaard, Christine (1997), 'The Normativity of Instrumental Reason'. In Garrett and Gullity (eds.), *Ethics and Practical Reason*. Clarendon Press, Oxford, 215–254.

Kornhuber, H.H. and Deecke, L. (1965), 'Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale'. *Pflügers Archiv für Gesamte Psychologie* 284, 1–17.

Kraut, Richard (2006), 'How to Justify Ethical Propositions: Aristotle's Method'. In Kraut (ed.) 2006, 76–115.

Kraut, Richard (ed.) (2006), *The Blackwell Guide to Aristotle's Nicomachean Ethics*. Blackwell, London.

Kripke, Saul (1980), *Naming and Necessity*. Harvard University Press, Cambridge, Mass.

Kripke, Saul (1981), *Wittgenstein on Rules and Private Language*. Blackwell, Oxford.

Latané, Bibb and Rodin, Judith (1969), 'A Lady in Distress: Inhibiting Effects of Friends and Strangers on Bystander Intervention'. *Journal of Experimental Social Psychology* 5, 189–202.

Libet, Benjamin, Gleason, Curtis A., Wright, Elwood W. and Pearl, Dennis K. (1983), 'Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential)'. *Brain* 106, 623–642.

Libet, Benjamin (1985), 'Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action'. *Behavioral and Brain Sciences* 8, 529–539.

Little, Margaret (1997), 'Virtue as Knowledge: Objections from the Philosophy of Mind', *Noûs* 31 (1), 59–79.

Lewis, David (1973), *Counterfactuals*. Blackwell, Oxford.

Lewis, David (1981), 'Are We Free to Break the Laws?' *Theoria* 47, 113–121.

Lewis, David (1986), *On the Plurality of Worlds*. Blackwell, Oxford.

Lewis, David (1997), 'Finkish Dispositions'. *Philosophical Quarterly* 47, 143–158.

Lewis, David (1999), 'Elusive Knowledge'. Reprinted in his *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge, 418–445.

Mackie, John (1977), *Ethics: Inventing Right and Wrong*. Penguin, Harmondsworth.

Macrae, C. N and Johnston, L. (1998), 'Help, I Need Somebody: Automatic Action and Inaction'. *Social Cognition* 16, 400–417.

May, Larry, Friedman, Marilyn, and Clark, Andy (eds.) (1996), *Mind and Morals. Essays on Cognitive Science and Ethics*. MIT Press, Cambridge, Mass.

McDowell, John (1979/1998), 'Virtue and Reason'. Reprinted in McDowell 1998a, 50–73.

McDowell, John (1980/1998), 'The Role of *Eudaimonia* in Aristotle's Ethics'. Reprinted in McDowell 1998a, 3–22.

McDowell, John (1981/1998), 'Non-Cognitivism and Rule-Following'. Reprinted in McDowell 1998a, 198–218.

McDowell, John (1995/1998), 'Might There Be External Reasons?'. Reprinted in McDowell 1998a, 95–111.

McDowell, John (1996/1998), 'Two Sorts of Naturalism'. Reprinted in McDowell 1998a, 167–197.

McDowell, John (1998a), *Mind, Value, and Reality*. Harvard University Press, Cambridge, Mass.

McDowell, John (1998b), 'Some Issues in Aristotle's Moral Psychology'. In McDowell 1998, 23–49.

McKenna, Michael (2000), 'Source Incompatibilism, Ultimacy, and Transfer NR'. *American Philosophical Quarterly*.

Mikhail, John (2000), *Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice.'* PhD Dissertation, Cornell University.

Mikhail, John (forthcoming), 'Universal Moral Grammar: Theory, Evidence, and Future'. *Trends in Cognitive Sciences*.

Mill, John Stuart (1863), *Utilitarianism*. Online at http://etext.library.adelaide.edu.au/m/mill/john_stuart/m645u/.

Mele, Alfred (1995), *Autonomous Agents*. Oxford University Press, Oxford.

Mele, Alfred and Robb, David (1998), 'Rescuing Frankfurt-Style Cases'. *Philosophical Review* 107, 97–112.

Mele, Alfred (2006), *Free Will and Luck*. Oxford University Press, Oxford.

Mele, Alfred (forthcoming), 'Free Will: Action Theory Meets Neuroscience'. In C. Lumer (ed.) *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy*. Ashgate, London.

Moore, George Edward (1939), 'Proof of an External World', *Proceedings of the British Academy* 25, 273–300.

Nado, Jennifer, Stich, Stephen P., and Kelly, Daniel (forthcoming), 'Moral Psychology'. In John Symons and Paco Calvo (eds.), *Routledge Companion to the Philosophy of Psychology*. Routledge, London.

Nahmias, Eddy (forthcoming), 'The Psychology of Free Will'. To appear in Jesse Prinz (ed.), *The Oxford Handbook on the Philosophy of Psychology*. Oxford University Press, Oxford.

Nahmias, Eddy, Morris, Stephen, Nadelhoffer, Thomas, and Turner, Jason (2005), 'Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility'. *Philosophical Psychology* 18, 561–584.

Nichols, Shaun (2004), *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press, Oxford.

Nisbett, R. E. and Wilson, T. D. (1977). 'Telling More Than We Can Know: Verbal Reports on Mental Processes'. *Psychological Review* 84, 231–259.

Nucci, Larry (1986), 'Children's Conceptions of Morality, Social Conventions and Religious Prescription'. In C. Harding, (ed.), *Moral Dilemmas: Philosophical and Psychological Reconsiderations of the Development of Moral Reasoning*. Precedent Press, Chicago.

Nucci, Larry and Turiel, Elliot (1993), 'God's Word, Religious Rules, and Their Relation to Christian and Jewish Children's Concepts of Morality'. *Child Development* 64, 1475– 1491.

Nucci, Larry and Weber, Elsa K. (1995), 'Social Interactions in the Home and the Development of Young Children's Conceptions of the Personal'. *Child Development* 66, 1438–1452.

Nussbaum, Martha C. (1990), 'The Discernment of Perception: An Aristotelian Conception of Private and Public Rationality'. In *Love's Knowledge: Essays on Philosophy and Literature*. Oxford University Press, Oxford, 54–105.

Nussbaum, Martha C. (1995), 'Aristotle on Human Nature and the Foundations of Ethics'. In James Edward John Altham and Ross Harrison (eds.), *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Cambridge University Press, Cambridge 1995, 86–131.

Pelham, Brett W., Mirenberg, Matthew C., and Jones, John K. 2002, 'Why Susie Sells Seashells by the Seashore: Implicit Egotism and Major Life Decisions.' *Journal of Personality and Social Psychology* 82: 469-487.

Pereboom, Derk (2001), *Living Without Free Will*. Cambridge University Press, Cambridge.

Pettit, Philip and Smith, Michael (1993), 'Practical Unreason'. *Mind* 102, 53–79.

Prinz, Jesse (2006a), 'The Emotional Basis of Moral Judgments'. *Philosophical Explorations* 9 (1), 29–43.

Prinz, Jesse (2006b), "Is the Mind Really Modular?" In R. Stainton (ed.), *Contemporary Debates in Cognitive Science*. Blackwell, Oxford.

Prinz, Jesse (forthcoming), *The Emotional Construction of Morality*. Oxford University Press, Oxford.

Quine, Willard van Orman (1951), 'Two Dogmas of Empiricism'. *The Philosophical Review* 60, 20–43.

Quine, Willard van Orman (1960), *Word and Object*. MIT Press, Cambridge, Mass.

Railton, Peter (1984), 'Alienation, Consequentialism, and the Demands of Morality'. *Philosophy and Public Affairs* 13, 134–171.

Rawls, John (1971), *A Theory of Justice*. Harvard University Press, Cambridge, Mass.

Reeve, C. D. C. (1992), *Practices of Reason*. Oxford University Press, Oxford.

Reeve, C. D. C. (2006), 'Aristotle on the Virtues of Thought'. In Kraut (ed.) 2006, 198–217.

Richardson Lear, Gabriel (2006), 'Aristotle on Moral Virtue and the Fine'. In Kraut (ed.) 2006, 116–136.

Ross, William David (1930/2002), *The Right and the Good*. Ed. Philip Stratton-Lake. Oxford University Press, Oxford.

Sayre-McCord, Geoffrey (1994), 'Why Hume's 'General Point of View' Isn't Ideal – and Shouldn't Be'. *Social Philosophy and Policy* 11 (1), 202–228.

Scanlon, Thomas (1998), *What We Owe to Each Other*. Harvard University Press, Cambridge, Mass.

Schueler, G. F. (1995), *Desire: Its Role in Practical Reason and the Explanation of Action*. MIT Press, Cambridge, Mass.

Searle, John (1983), *Intentionality. An Essay in the Philosophy of Mind*. Cambridge University Press, Cambridge.

Shweder, Richard A. (1990), 'In Defense of Moral Realism: Reply to Gabennesch'. *Child Development* 61, 2060–2067.

Shweder, Richard A., Mahapatra, M., and Miller, J. G (1987), 'Culture and Moral Development'. In Kagan and Lamb (eds.) 1987, 1–82.

Simpson, Evan (1999), 'Between Internalism and Externalism in Ethics'. *The Philosophical Quarterly* 49 (195), 201-214.

Small, Deborah A. and Loewenstein, George (2003), 'Helping *a* Victim or Helping *the* Victim'. *Journal of Risk and Uncertainty* 26, 5-16.

Smetana, Judith G. (1981), 'Preschool Children's Conceptions of Moral and Social Rules', *Child Development* 52, 1333–1336.

Smetana, Judith. G. (1989), 'Toddlers' Social Interactions in the Context of Moral and Conventional Transgressions in the Home'. *Developmental Psychology* 25, 499–508.

Smetana, Judith G. (1993), 'Understanding Social Rules'. In

Smith, Adam (1759/1976), *The Theory of Moral Sentiments*. Ed. D. D. Raphael and A. L. Macfie. Oxford University Press, Oxford.

Smith, Michael (1994), *The Moral Problem*. Blackwell, Oxford.

Smith, Michael (1987), 'The Humean Theory of Motivation'. *Mind* 96, 36–61.

Smith, Michael (1995), 'Internal Reasons'. *Philosophy and Phenomenological Research* 55 (1), 109–131.

Smith, Michael (2004), 'Rational Capacities'. In *Ethics and the A Priori: Selected Essays on Moral Psychology and Meta-Ethics*. Cambridge University Press, Cambridge.

Sober, Elliott and Wilson, David Sloan (1998), *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press, Cambridge, Mass.

Sreenivasan, Gopal (2002), 'Errors About Errors: Virtue Theory and Trait Attribution'. *Mind* 111, 47–68.

Stevenson, Charles L. (1945), *Ethics and Language*. Yale University Press, New Haven.

Stich, Stephen, Doris, John, and Roedder, Erica (MS), 'Egoism vs. Altruism (Death by Altruism)'.

Stratton-Lake, Philip (2000), *Kant, Duty and Moral Worth*. Routledge, London.

Strawson, Galen (1994), 'The Impossibility of Moral Responsibility', *Philosophical Studies* 75, 5–24.

Strawson, Galen (2002), 'The Bounds of Freedom'. In R. Kane (ed.), *The Oxford Handbook on Free Will*. Oxford University Press, Oxford, 441–460.

Sturgeon, Nicholas (1985), 'Moral Explanations'. Reprinted in Geoffrey Sayre-McCord (ed.), *Essays on Moral Realism*. Cornell University Press, Ithaca, 229–255.

Tenenbaum, Sergio (forthcoming), *Appearances of the Good: An Essay on the Nature of Practical Reason*. Cambridge University Press, Cambridge.

Tiberius, Valerie (2006), 'Well-Being: Psychological Research for Philosophers', *Philosophy Compass* 1, 493–505.

Tisak, Marie S. and Turiel, Elliot (1984), 'Children's Conceptions of Moral and Prudential Rules'. *Child Development* 55, 1030–1039.

Turiel, Elliot (1983), *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press, Cambridge.

Turiel, Elliot (2002), *The Culture of Morality. Social Development, Context, and Conflict*. Cambridge University Press, Cambridge.

Turiel, Elliot, Killen, M., and Helwig, C.C. (1987), 'Morality: Its Structure, Functions, and Vagaries'. In Kagan and Lamb (eds.) 1987, 166–245.

Valdesolo, Piercarlo and DeSteno, David (2006), 'Manipulations of Emotional Context Shape Moral Judgment'. *Psychological Science* 17 (6), 476–477.

Velleman, David (2006), 'An Introduction to Kantian Ethics'. In *Self to Self: Selected Essays*. Cambridge University Press, Cambridge, 16–44.

Vogel, Lawrence (1993), 'Understanding and Blaming: Problems in the Attribution of Moral Responsibility'. *Philosophy and Phenomenological Research* 53 (1), 129–142.

Wallace, R. Jay (1990), 'How to Argue About Practical Reason', *Mind* 99, 355–385.

Wallace, R. Jay (1994), *Responsibility and the Moral Sentiments*. Harvard University Press, Cambridge, Mass.

Wallace, R. Jay (2003), 'Explanation, Deliberation, and Reasons'. *Philosophy and Phenomenological Research* 67, 429–435.

Wallace, R. Jay (2005), 'Moral Psychology'. In Frank Jackson and Michael Smith (eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford University Press, Oxford.

Watson, Gary (1975), 'Free Agency'. *Journal of Philosophy* 72 (8), 205–220.

Watson, Gary (1987), 'Free Action and Free Will'. *Mind* 96, 145–172.

Watson, Gary (1996), 'Two Faces of Responsibility'. *Philosophical Topics* 24 (2), 227–248.

Watson, Gary (ed.) (2003), *Free Will*. Second Edition. Oxford University Press, Oxford.

Wedgwood, Ralph (2002), 'Practical Reason and Desire'

Wegner, Daniel and Wheatley, Thalia (1999), 'Apparent Mental Causation: Sources of the Experience of Will'. *American Psychologist* 54 (7), 480–492.

Wegner, Daniel (2002), *The Illusion of Conscious Will*. MIT Press, Cambridge, Mass.

Wheatley, T and Haidt, Jonathan (2005), 'Hypnotically Induced Disgust Makes Moral Judgments More Severe'. *Psychological Science* 16, 780–784.

Williams, Bernard (1981a), *Moral Luck: Philosophical Papers 1973–1980*. Cambridge University Press, Cambridge.

Williams, Bernard (1981b), 'Internal and External Reasons'. In Williams 1981a, 101–113.

Williams, Bernard (1981c), 'Persons, Character, and Morality'. In Williams 1981a, 1–19.

Williams, Bernard (1995), 'Internal Reasons and the Obscurity of Blame'. In *Making Sense of Humanity and Other Philosophical Papers 1982-1993*. Cambridge University Press, Cambridge, 35–45.

Williams, J. M. G., Mathews, A., and MacLeod, C. (1996), 'The Emotional Stroop Task and Psychopathology'. *Psychological Bulletin* 120, 3–24.

Wilson, T. D. (2002), *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Belknap, New York.

Widerker, David (1995), 'Libertarianism and Frankfurt's Attach on the Principle of Alternative Possibilities'. *Philosophical Review* 104, 247–261.

Wittgenstein, Ludwig (1922/1981), *Tractatus Logico-Philosophicus*. Ed. David F. Pears, Routledge, London.

Wolf, Susan (1982), 'Moral Saints'. *Journal of Philosophy* 79 (8), 419–439.

Wolf, Susan (1987), 'Sanity and the Metaphysics of Responsibility'. Reprinted in Watson (ed.) 2003, 372–387.

Wolf, Susan (1990), *Freedom Within Reason*. Oxford University Press, Oxford.

Woolfolk, Robert L., Doris, John M. and Darley, John M. (2006), 'Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility.' *Cognition* 100 (2), 283–301.

Zajonc, R. B. (1980). 'Feeling and Thinking: Preferences Need No Inferences'. *American Psychologist* 35, 151–175.

Zangwill, Nick (1995), 'Moral Supervenience'. *Midwest Studies in Philosophy* XX.

Zimbardo, Philip (2007), The Lucifer Effect. Understanding How Good People Turn Bad. Random House, New York.