Between Facts and Norms: Ethics and Empirical Moral Psychology

Hanno Sauer

A Cold, Hard Look

For most of its history, philosophical moral psychology has been in bad shape. People were asking the right questions, but their methods were questionable: rampant speculation was revised in light of pure guesswork; guesswork had to be amended on the account of arbitrary superstition; superstition was corrected by flimsy moralizing, and the whole thing was rounded off by a healthy dose of wishful thinking. Philosophical theories of human nature had to state how human beings ought to be, rather than how they actually are.

It is not a good idea, generally speaking, to speculate about the nature of the moral mind without systematically investigating how the mind works. Why philosophers failed to appreciate this rather obvious truth is something I can only speculate about myself. The—arguably false—idea that the mind is transparent to itself, and can thus be studied without external aid, may have played a role. We now know that this type of self-transparency is an illusion and that expecting the mind to give honest answers when examined by introspection alone is hopelessly naive.

Perhaps I exaggerate, and it wasn't quite as bad. To find out how moral agents think and act, some philosophers like Aristotle, Hume, or Kant did consult the best science of their time. Then again, this did not necessarily amount to much. Others—Nietzsche comes to mind (Knobe and Leiter 2007)—were in fact pioneers and gave the field of empirically informed moral psychology, most of which was yet to emerge at the time, new directions to pursue, and new questions to address. Yet all too often, philosophers "have been content to invent their psychology […] from scratch" (Darwall et al. 1992, 189). A "cold, hard look at what is known about human nature" (Flanagan 1991, 15) seems to me to be the best cure for this affliction.

H. Sauer (✉)
Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands
e-mail: h.c.sauer@uu.nl

B.G. Voyer, T. Tarantola (eds.), *Moral Psychology*,
DOI 10.1007/978-3-319-61849-4_2

The main tension between philosophical and empirical accounts of human moral judgment and agency comes down to the fact that, at the end of the day, philosophers are interested in moral psychology for one thing, and one thing only (I exaggerate again). They want to know what facts about the *psychological* foundations of morality can teach us about the foundations of morality, *period*: how facts about human nature bear on right and wrong, good and bad, and just and unjust. This tension is further aggravated by the fact that many philosophers deem this to be a hopeless endeavor that is doomed to fail from the outset. The problem, these philosophers argue, is that there is no way (no legitimate and informative one, at any rate) to get from an *is* to an *ought*. Rumor has it that facts are different from values. Descriptive statements, it is said, do not entail prescriptive propositions. Empirical information, the story goes, has no normative significance. Nature allegedly has no moral import.

In what follows, I will refer to this problem as *the gap*. In the first section of this chapter, I will briefly explain what the gap is, why it is said to exist, and to what extent it is supposed to pose an obstacle to empirically informed theorizing about ethics.

In the second section, I will take a look at some of the most interesting recent developments in empirical moral psychology and explain what their normative implications are supposed to be. My selection of topics will be somewhat arbitrary, and the discussion I provide by no means is comprehensive. I am not attempting to give an overview of the whole field of contemporary moral psychology. This has already been done elsewhere, by people more qualified to do this than myself (see Doris and Stich 2005; Rini 2015; Kumar forthcoming; Alfano and Loeb 2014; Alfano 2016; Appiah 2008; Tiberius 2014, and the remainder of this book). Instead, I choose a more focused approach and look at the whole field from the perspective of what I take to be the main issue of philosophical interest: my aim is to illustrate how empirical moral psychology might be brought to bear on issues of normative significance—what the virtues are, what makes for a good life, whether free will exists, what role luck plays in morality, what constitutes an action, what it means to be a person, how people arrive at moral judgments, whether these judgments are relative, and whether we are at all competent to make them. My discussion will be arranged around four clusters: normative theory, moral agency, moral and nonmoral judgment, and moral intuition.

In the final section, I will extract some lessons from this discussion. Are the skeptics right? When it comes to figuring out what demands morality makes on us, does empirical information remain thoroughly irrelevant? Or are there grounds for optimism? Do empirically informed ethics have a future after all? I will argue that the normative significance of empirical studies of human moral cognition and behavior, though always indirect, comes in essentially three forms: by debunking the processes on the basis of which we make moral judgments and develop moral concepts; by undermining the empirical presuppositions of some normative theories, vindicating those of others; and by providing tools for the reflective improvement of moral judgment and agency by bringing to light the sometimes egregious mistakes that escape our powers of introspection and the empirically unaided mind.

The Gap

In philosophy, skepticism about the relevance of empirical facts for so-called normative questions—questions about right and wrong, permissible and forbidden, and virtue and vice—can draw on two *loci classici*. One can be found in the third part of David Hume's *Treatise of Human Nature*, where he complains that

> "[i]n every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not" (Hume 1739/2000, III.I.I)

Hume argued that this transition was as widespread as it was illegitimate; for in his view, and the view of many others, there is no logically valid way to derive a proposition with normative content (it is not ok to lie; drone surveillance is reprehensible; chastity is a virtue; we have a duty to help others, and when doing so, it involves little cost to ourselves) from a set of premises with purely descriptive, factual content (people lie all the time; drones are really useful; your father wants you to be chaste; helping others will make people like you). An inference is logically valid just in case the truth of its premises guarantees the truth of its conclusion. No such inference, Hume thought, could ever take you from an *is* to an *ought*.

The second go-to place for friends and foes of *the gap* is G. E. Moore's (1903) *Principia Ethica*. Here, Moore coined the term "naturalistic fallacy" (Moore 1903) to refer to attempts to identify the property of being *good* with any natural property, such as being *useful*, or *maximizing pleasure*, or being *economically efficient*, or being *sanctioned by the state*. Moore's point was that *good* and *bad* cannot be defined in natural terms, because if they could, then whenever we had found some action or event instantiating the natural property picked out by our definition (given that said definition is correct), the question whether the action or event is also good would necessarily be *closed* to anyone but the conceptually confused. Centaurs, and only centaurs, are creatures with an anthropic upper and hippic lower half; if I manage to show you such a thing, the question whether it is also a centaur is *closed*. Now Moore argued that for every proposed natural definition of the good—say "the good which maximizes pleasure"—it always remains possible to ask whether something instantiating the natural property specified in the definiendum is also good. "It maximizes pleasure, but is it also good," or "it is loved by the Gods, but is it also good," or "it is useful for society, but is it also good," and so on. These questions all make sense, and the property of being good cannot be conceptually reduced to other natural properties. This is Moore's famous "open question argument."

The naturalistic fallacy is not strictly speaking a fallacy, and as we have seen, the term was originally supposed to refer not to *the gap*, but an entirely different, semantic point. Then again, people love to accuse one another of fallacious reasoning, and the term is catchy, so "naturalistic fallacy" stuck around and is now widely

used for illicit attempts to bridge *the gap*. Examples for naturalistic fallacies are ridiculously easy to find and are especially common in debates on evolutionary psychology, sexual morality, and most other topics in applied ethics. I will not cite any sources here, as the research would have been too depressing. But I *can* give a few examples of the kind of reasoning I have in mind and which we are all too well acquainted with: evolution favors the selfish and competitive, so that is how we, too, ought to act; homosexuality is unnatural and should thus be banned; humans are the only animals with the power to reason, so the rational life is best for humans; people have always killed animals for food, and women were always discriminated against, so clearly there is nothing wrong with those things. Never mind whether these inferences get the facts right or not—because even if they did, they would fail to establish their conclusion on account of *the gap*.

On the other hand, it seems hard to see how empirical facts could *always* remain *thoroughly* irrelevant to normative inquiry. Whether or not abortion is permissible, and under what conditions, it will surely depend on what kind of being a fetus is and whether it can feel pain or has interests and conscious experiences. Likewise, my indignation toward the man I believe my wife cheated on me with, and which I am about to punch in the face, will readily switch its target once I have found out that *this* man isn't the culprit, but the pathetic scoundrel standing next to him. What should be done about climate change, or whether anything should be done at all, cannot be assessed without factual knowledge. And whether or not you should perform that tracheotomy to save your suffocating friend will depend on how likely it is that you will succeed. In all these cases, empirical facts have bearing on issues of normative significance, if only via the nonmoral facts upon which moral facts are grounded.

Moreover, many normative moral theories seem to make rather straightforward assumptions about what kinds of agents we are, assumptions which are far from empirically innocent. For instance, some Kantians argue that moral norms are prescriptive rules whose authority does not depend on whether or not one is already motivated to conform to them: these rules are supposed to be motivating *independently* of an agent's desires and goals, simply by virtue of the fact that they specify what it means to be an agent (Korsgaard 1996; Velleman 2011). But what if this paints an unrealistic picture of how motivation works and of what constitutes an agent? Virtue ethicists often claim that a good person is a person with a coherent set of laudable character traits (Hursthouse 1999; Foot 2001). Does this account rely on an erroneous idea of how people function and how well their personalities are integrated? Some consequentialists hold that the right action—the one we ought to choose—is the unique action that has the best consequences. But what if figuring out which action is beyond human deliberative powers (Mason 2013)? In all these cases, normative theories make empirical presuppositions.

The question, then, is this: despite the fact that no ought ever follows from an is and despite the fact that the concept of the good cannot be identified with any empirical property, how should we understand the *normative relevance of empirical facts* in light of the *empirical presuppositions of normative theories*?

Normative Theory

(i) *Consequentialism and Deontology*. Contemporary normative ethics is organized around a distinction that manages at the same time to be one of the least well liked and yet one of the most popular in all of philosophy: the distinction between *consequentialism* and *deontology*. Consequentialist moral theories hold that the rightness or wrongness of an action is determined *only* by its (actual or expected) consequences. Deontological moral theories deny this. Some deontologists hold that intentions matter for the moral evaluation of an action as well, others argue that there are certain side constraints (such as individual rights) on the maximization of the good, and that it can make a moral difference whether one actively does something or merely allows it to happen or whether someone uses someone else as a mere means to an end rather than an end in his/herself. There is plenty of evidence that on an intuitive level, people take deontological considerations to be morally relevant (Young et al. 2007). Often, their judgments conform to deontological rules such as the doctrine of double effect (according to which harming someone can be permissible when it is an unintended but foreseen side effect, rather than when the harm is directly intended, Mikhail 2007, Kamm 2007), even though such slightly more sophisticated principles may remain ineffable.

What about *the gap*? Can empirical data shed light on which theory is correct? One way to model the difference between consequentialism and deontology is to look at sacrificial dilemmas involving urgent trade-offs between harming an individual person and promoting the greater good and to see which conflicting actions consequentialism and deontology classify as right and wrong, respectively, when doing what's best overall clashes with certain intuitively plausible moral rules. Moral emergencies (Appiah 2008, 96ff.) of this sort form the basis of what is perhaps the single most thriving and controversial research program in normatively oriented empirical moral psychology: Joshua Greene's *dual process* model of moral cognition (Greene 2014). According to this model, cognitive science can show that one of the two normative theories is superior to the other. Consequentialism, the evidence is purported to show, engages more rational parts of the brain and more sophisticated types of processing than deontology, which is associated with more emotional parts of the brain and more crude forms of cognition (Greene 2001, 2004). When people judge it impermissible, for instance, to kill one person to save five others (thereby endorsing the deontological option), they arrive at this judgment via a more emotional and less calculating route. Deontological moral theory, then, amounts to little more than post hoc rationalizations of those brute, alarm-like responses (Greene 2008; see chapter "Cognitive and Neural Sciences: Investigating the Moral System" for a more thorough discussion of the neuroscience of moral judgment).

The dual process model's main normative upshot is supposed to be a vindication of consequentialist and a debunking of deontological intuitions on the basis of empirical evidence regarding the cognitive processes that produce these two types of moral intuitions. But it remains unclear whether the way people arrive at their

consequentialist responses deserves to be described as consequentialist reasoning at all, rather than an ordinary weighing of competing considerations for and against a proposed action (Kahane 2012). Even worse, the consequentialist judgments that some people end up endorsing do not seem to be based on an impartial concern for the greater good, but on much more sinister dispositions (Kahane et al. 2015). Perhaps most importantly, the connection between consequentialist judgments and controlled, System II processing on the one hand, and deontological judgments and automatic, System I processing on the other hand (Evans 2008; Stanovich 2011; Kahneman 2011), seems to be due to the fact that in Greene's original studies, the consequentialist option always *happened to be* the counterintuitive one. When this confound is removed and counterintuitive deontological options are included, the pattern is reversed (Kahane et al. 2012; cf. Greene et al. 2014).

Dual process theory continues to be haunted by *the gap*. Empirical data on which type of process, or which brain region, is involved in the production of a moral judgment tells us very little about whether or not this judgment is justified or not—unless we *already* know which processes are unreliable and which aren't, which we arguably do not. Now the dual process model's two best shots are an *argument from morally irrelevant factors* and an *argument from obsoleteness*. Firstly, it could be shown that regardless of whether people arrive at them through emotion or reasoning, deontological intuitions pick up on *morally irrelevant factors*, such as whether an act of harming someone has been brought about in a distal or proximal way. Such sensitivity to morally extraneous features is often sufficient to indict a particular type of judgment as unreliable. Secondly, one could argue that some moral intuitions are generated on the basis of processes which are unlikely to deliver correct results under conditions they have neither evolved nor have been culturally shaped in (Singer 2005). For instance, moral cognition may be good at dealing with how to secure cooperation in stable small-scale communities. Dynamic, large-scale societies comprised of strangers and organized on the basis of complex economic and political institutions may constitute a hostile environment for the cognitive processes our ancestors bequeathed to us. Since a similar story may be true of the processes responsible for deontological intuitions and the conditions we currently inhabit, this, too, could help undermine the credibility of those intuitions via those processes (Nichols 2014).

The problem with these arguments, however, is that it is far from clear which role empirical evidence has to play in them at all, and whether most or all of the normative heavy lifting isn't done by armchair theorizing about what does and what doesn't count as morally relevant—which is to say, by moral philosophy (Berker 2009; Sauer 2012b). As for the second point, it has to be emphasized that the primitive cognitive processes of modern social conditions with their dynamic, large, anonymous societies and complex technological challenges do not exclusively deliver deontological intuitions. Conversely, the cognitive processes that are required to successfully navigate such conditions are not exclusively consequentialist in nature. As far as the consequentialism/deontology distinction is concerned, dual process theory is thus neither here nor there. What remains of its steep ambitions may simply be that some moral judgments are produced by

automatic and some by controlled cognitive processes, together with the claim that under certain conditions, the former are less likely to produce correct responses than the latter.

(ii) *Moral Relativism*. But why speculate about the correct normative theory, when it is far from clear whether moral problems have correct solutions at all? Isn't it clear that people have widely diverging and irreconcilable views about what morality requires? One does not need empirical research to confirm that people disagree about morality. The so-called argument from disagreement (Brink 1984; Mackie 1977) is supposed to use this fact of life to make the case for moral relativism, the view that there is no single true morality and that moral norms and values are only ever valid relative to some individual, social, or cultural context.

The problem with this argument is that there is a rather obvious objection to it. Disagreement does not entail relativity (Enoch 2009): people disagree about all kinds of things, but this doesn't mean that there are no facts of the matter about which one side is right and the other wrong. Non-relativists like to point out that what is needed for the argument from disagreement to get off the ground is a case of intractable *fundamental* moral disagreement—disagreement that would persist even under ideal conditions of full information and flawless reasoning on the part of those disagreeing. Non-fundamental disagreement, the kind that is purportedly not damaging to moral universalists, is disagreement for which a so-called defusing explanation can be given. Such disagreement can be due, among other things, to disagreement about the underlying facts, special pleading, or irrationality. Special pleading occurs when people refuse to apply a value consistently, trying to make an exception for themselves (e.g., endorsing the death penalty except when oneself is to be executed); irrationality can occur when people fail to appreciate what their values entail (e.g., wanting to reduce the suffering of sentient beings, but not adjusting one's diet in light of this goal).

What can empirical data contribute to this debate? Recently, Doris and Plakias (2008) have tried to revive the argument from disagreement by bringing evidence from cultural psychology to bear on the issue of whether it is possible to identify a case of fundamental moral disagreement for which no defusing explanation seems to be available. For instance, Doris and Plakias draw heavily on Nisbett and Cohen's (1996) "culture of honor" explanation for differences in attitudes toward violence between people from the American North and South. Evidence from criminal statistics, legal decisions, lab experiments, and field studies all point in the direction that Southerners are both more prone to violence and more tolerant of it. Nisbett and Cohen attribute this tendency, which is restricted to violence in response to threats, insults, and other violations of *honor*, to the reputational demands of herding economies. In contrast to economies based on farming or trade, a herding economy is a high-stakes environment in which a person's entire assets could be stolen, which made it necessary for individuals to convey that they would be willing to respond violently to threats. Others (Fraser and Hauser 2010) have argued that some cultures (e.g., rural Mayans) do not see a morally relevant difference between acts and omissions, which is another promising candidate for a fundamental moral disagreement.

Does this type of argument succeed in bridging *the gap*? Doris and Plakias argue that none of the aforementioned defusing explanations plausibly account for differences in Southern and Northern attitudes toward violence. If true, this would support their case for moral relativism. However, there are reasons for doubt. To a large extent, cross-cultural agreement about certain general prima facie duties is compatible with seemingly dramatic disagreement about all-things-considered obligations (Meyers 2013). Many disagreements concern *how* wrong or right something is and do not involve one party thinking that something is completely wrong which the other thinks is completely innocuous. That Southerners behave more violently and are more likely to condone violence does not mean that they take it to be more permissible (Leiter 2007). Moreover, most disagreements vanish under close scrutiny: when they are subjected to the sort of inquiry moral universalists favor, moral disputes tend to disappear (hint: less rural/more formally educated Mayans *do* see a difference between doing and allowing). The disagreements Doris and Plakias base their argument on can be located at the level of unreflective System I responses, where they inflict hardly any damage on non-relativists (Fitzpatrick 2014). If Southerners were informed about Nisbett and Wilsons's "culture of honor" explanation *itself*, and thus about the fact that the original economic rationale for their attitudes no longer obtains, they may well be inclined to change those attitudes (Sneddon 2009). This sort of genealogical defeater is demonstrably effective (Paxton et al. 2012). The issue of moral relativism thus can be addressed empirically, at least as long as its defenders and opponents are willing to make clear predictions on how much convergence or divergence in people's moral views, and of what sort, to expect if their respective positions are true.

Moral Agency

(iii) *Character, Situation, and Virtue*. The fourth main player in normative ethics besides consequentialism, deontology, and moral relativism—*virtue ethics*—does not merely incur, as it were by accident, empirical presuppositions regarding what kinds of agents we are. Rather, its normative criteria are straightforwardly built upon an account of agency, thereby rendering it particularly hostage to empirical fortune. The rightness of an action does not, on this account, lie in the extent to which it satisfies some principled criterion of rightness. The right action, virtue ethicists argue, is the one the virtuous person would perform under the circumstances. The virtuous person is a person of good character, that is, an agent who possesses an assortment of praiseworthy traits such as honesty, courage, persistence, tranquility, magnanimity, and other quaint things.

It has long seemed fair to empirically minded philosophers (Harman 1999; Doris 2002) to ask whether this account of human agency is at all realistic. Perhaps unsurprisingly, they have been keen to show that it is not (Ross and Nisbett 1991; Doris 2009). The evidence—ranging from landmark experiments, such as Milgram's obedience studies, Zimbardo's prison experiment, and various studies on helping

behavior (Isen and Levin 1972; Darley and Batson 1973), to real-life atrocities such as the massacre of My Lai, the Rwandan genocide, or the violent and humiliating abuse of prisoners at Abu Ghraib (Doris and Murphy 2007)—consistently suggests that cross-situationally stable character traits of the kind postulated by virtue ethicists are nowhere to be found. The influence of frequently subtle and seemingly insubstantial situational features towers over that of internal dispositions.

However, even in this seemingly open and shut case in favor of situationism, the gap is not bridged without resistance. Some virtue ethicists have argued that character traits need to be construed differently (Kristjánsson 2012; Webber 2013), sought elsewhere (Merritt 2009), or that there is contrary evidence pointing toward the existence of virtues (Vranas 2005). Others chose to insist on the fact that the acquisition of virtues was always supposed to be a rare ideal, so that evidence for the rarity of virtuous agency cuts no ice (Miller 2003). Then again, few are comfortable defending unattainable ideals, and rightly so.

Among the more radical friends of situationism, some have suggested that we should abandon futile character education in favor of effective situation management (Harman 2009). Others have advocated a different form of moral technology that relies on the factitiousness of virtue: the nonexistence of global traits gives us no reason to abandon trait talk, which can function as a self-fulfilling prophecy. This suggests that we stop attributing only undesirable traits (Alfano 2013). Finally, some have argued that virtue ethics fails even if traits are real (Prinz 2009), because its normative authority rests upon an account of universal human nature that is debunked by cultural psychology.

(iv) *Freedom of the Will*. Virtue ethics is perhaps the clearest example of a normative theory that can be assessed in light of empirical facts. Other aspects of moral agency, such as freedom of the will, are harder to pin down; after all, many philosophers believe that free will just isn't the kind of thing that can be studied empirically.

The contemporary debate on the nature and existence of freedom of the will, perhaps one of the most mature in all of philosophy, cannot be adequately summarized here. Instead, I wish to mention two types of empirically supported challenges to free will and moral responsibility and to see what may follow from them normatively. One has to do with the *timing* of choice, the other with whether we have reason to believe conscious intentions ever really cause actions at all.

The first challenge, and arguably the more famous one, aims to show that people's conscious intentions do not initiate their actions (Libet 1985). In a series of experiments, Benjamin Libet could show that people's decision to execute a simple motor action is preceded, in the range of an average 350 ms, by a readiness potential (measured via EEG) initiating the action before people become aware of it. Other studies (Soon et al. 2008) report that it is possible to predict, with above chance accuracy, which of two simple actions an individual will perform up to 10 s before a subject's conscious decision. This makes it hard to see how conscious intentions could be responsible for action initiation.

According to the second challenge, a range of phenomena such as illusions of control, where people have the feeling of agency without any actual causal impact; episodes of confabulation, where people make up reasons for their actions that

couldn't possibly have played a motivating role; or certain pathological conditions such as utilization behavior or alien hand syndrome and, in general, the pervasive automaticity of human behavior (Bargh and Chartrand 1999) support the view that mental causation and the experience of conscious willing are illusory (Wegner 2002). In particular, people can have a sense of agency when their agency couldn't possibly have made a difference and are more than happy to come up with reasons for their actions that couldn't possibly have played a role in why they did what they did.

Both challenges are taken to suggest that our actions are determined by unconscious processes beyond our conscious awareness and control. I wish to remain agnostic about whether or not these challenges to free will are ultimately successful. But let me emphasize that the evidence also suggests that, at the very least, people retain a form of veto control over their actions (Schultze-Kraft et al. 2016). An unfree will may not be so hard to swallow if we at least have a free unwill.

Moreover, the Libet experiment (a) only concerns intentions *when* to perform a certain preselected action, and says nothing about decisions regarding *what* to do (however, see Haggard and Eimer 1999); (b) only investigates *proximal*, but crucially depends on the causal efficacy of *distal* intentions to follow the instructions of the experiment (Schlosser 2012a); and (c) presents only insignificant options which subjects have no good reasons to choose either way (Schlosser 2012b, 2014).

The normative problem of free will has two main aspects. One has to do with the consequences of people believing or disbelieving in free will. The other is about how we, individually and socially, should respond if free will turned out to be an illusion or to be much less free than we intuitively suppose. Firstly, people who have been primed to believe in determinism (which many, though importantly not all, hold to be incompatible with free will) are more likely to cheat on a subsequent task. Other studies suggest that disbelief in free will increases aggressiveness and reduces helping (Baumeister et al. 2009). On the other hand, a belief in free will need not have only desirable consequences, as it can make people more punitive and judgmental (Clark et al. 2014).

Secondly, and in line with the last point, the close tie between free will and moral responsibility entails that the nonexistence of free will has important ramifications for our social practice of punishment. To be sure, free will skepticism would leave three of the four functions of punishment—deterrence, protection, and rehabilitation—untouched, at least in principle. If free will does not exist, however, it may well turn out that all forms of *retributive* punishment are severely wrong (Zimmerman 2011). At the very least, it would open our punitive practices up for a sober empirical assessment in light of their consequences; drastically less harsh punishments, and perhaps even positive incentives to refrain from crime, are likely to be the upshot (Levy 2015). Retributive punishment has many undesirable consequences both for the punished and for society, which has to pay for expensive incarceration and deal with people who leave prison traumatized, stigmatized, and unemployable. When practices of punishment are assessed in light of their consequences rather than what wrongdoers allegedly deserve, these costs could be avoided.

Moral and Nonmoral Judgment

(v) *Personal Identity*. If situationists and free will skeptics are right, we are patchy puppets. Now what? Entities who are candidates for possessing free will or character traits are called *persons*. Persons, in turn, are the primary bearers of moral status: the coveted privilege of belonging to the circle of beings who enjoy special moral consideration in the form of rights and the dreaded burden of being the addressee of corresponding duties.

What does it take to be a person with an identity that remains stable over time? Either physical (the "stuff" people are made of) or psychological (people's "soul") continuity has been emphasized as the feature that decides what makes a person persist as one and the same (Martin and Barresi 2003). However, there is now a wealth of evidence suggesting that this is not how people think about personal identity.

Many concepts previously thought to be nonevaluative in character are actually downstream from people's moral assessments (the most famous perhaps being the concept of intentionality; more on this below). Personal identity is one such concept. For instance, people think that changes to a person's moral traits matter the most for whether a person stays the same or not (Strohminger and Nichols 2014). Moral judgments also influence how people think about what constitutes a person's true self, rather than more superficial aspects of their personality. First of all, people think that a person's core self is *fundamentally good* (Newman et al. 2015). This means that whether they take, say, an individual's inner dispositions or her explicit beliefs to constitute this core will depend on *their own* moral judgments: conservatives are more likely to think that a person's explicit beliefs form her true self when these beliefs display an aversion to homosexuality, but less likely to think so when those beliefs are pro-gay, and the other way around for a person's feelings of attraction. This leads to what is now sometimes referred to as the *Phineas Gage effect* (named after Phineas Gage, a nineteenth century railroad worker who allegedly underwent a drastic change of character after sustaining brain injury, Tobia 2015): changes for the better are seen as moves *toward* and changes for the worse as moves *away from* a person's true identity.

What is the normative relevance of this type of evidence? Of the many pressing moral issues for which personal identity is very important—how should we treat people's past wishes? what is the moral relevance of people who do not yet exist?—let me mention only one. A standard objection to utilitarianism has it that it licenses illicit trade-offs between people when aggregate welfare is maximized. As long as many can enjoy a life of leisure, it is palatable for a few to toil and drudge. But this, many think, ignores the essential separateness of persons: *inter*personal trade-offs, where a cost to one person is supposedly compensated by a larger benefit to another, should not be assimilated to *intra*personal trade-offs, where a cost incurred *now* can be outweighed by a later benefit to the same person. But if our intuitions about personal identity—the basic moral unit, as it were—are themselves shaped by moral intuitions, then our judgments about whom we are inclined to treat as a person at all, how to draw the lines between persons, and about the extent to which such lines carry moral weight may be deeply called into question.

(vi) *Intentionality*. Personal identity is only one of the domains where our thinking is influenced by moral considerations. In fact, some have suggested that the influence of moral judgments on the application of seemingly nonmoral concepts is pervasive: we are moralizers through and through (Pettit and Knobe 2009).

The most famous example is perhaps the concept of intentionality. Numerous studies confirm the basic asymmetric pattern: people are more likely to attribute intentionality for bad side effects than for good ones (Knobe 2003). When asked about whether the chairman of a company intentionally brought about a side effect to the environment, people are more likely to answer affirmatively when said side effect is bad rather than good. But why is this, when we tend to think that we need to establish intentionality first, to judge the morality of those intentional actions later?

And intentionality isn't the only concept people attribute asymmetrically when something of normative significance is at stake. Far from it, plenty of studies—on the doing/allowing distinction, the means/end distinction, knowledge, causality, free will, happiness, and many more (Cushman et al. 2008; Cova and Naar 2012; Beebe and Buckwalter 2010; Nichols and Knobe 2007; Phillips et al. 2011; Pettit and Knobe 2009; Knobe and Fraser 2008)—show that a host of other cognitive domains are susceptible to the same striking effect.

Knobe's surprising claim has long been that this influence of moral considerations on seemingly nonmoral issues is not a contaminating one where an otherwise value-neutral process is derailed, distorted, and illegitimately biased by people's moral beliefs (Knobe 2010; Sauer and Bates 2013). Rather, he has argued that moral judgments kick in at a deeper level, for instance, when setting the defaults against which intentionality and other psychological categories are assessed. In the case of the environment, the default is to be somewhat in favor of helping it; not caring about helping it at all, as the chairman is described in the original vignette, thus falls under this threshold. With respect to harming the environment, the default is to be against it; so in this case, not caring about harming it at all surpasses this threshold—hence the attribution of intentionality. Others have proposed that the aforementioned asymmetries are driven by judgments about norms more generally (Robinson et al. 2015; Holton 2010) or about people's so-called deep selves (Sripada and Konrath 2011).

Whatever the scope and substance of the correct explanation (Sauer 2014), the normative implications of the effect are potentially far-reaching and deeply revisionary. Outcomes which were brought about intentionally may not be worse than merely foreseen ones—worse outcomes would simply count as more intentional. Virtually all cases where intentionality is supposed to make a moral difference are affected by the asymmetry. Finally, the asymmetry may make it exquisitely difficult for jury members to accurately establish intentionality when immoral acts such as murder or rape are at issue (Nadelhoffer 2006). The very concepts we base our moral judgments upon may be suffused with morality from the outset. This would require us to reshape not just the way we think about a good deal of our practices, but those practices themselves.

(vii) *Moral Luck*. Other asymmetries are just as puzzling. A father whose children drown in the tub seems dramatically more blameworthy than one whose kids do not, even when both have exerted the same amount of care (or negligence) and one merely had good, the other bad luck. A drunk driver who happens to hit and injure someone is seen as a bad person, but millions of drunk drivers who simply had more luck are cut quite a bit of slack.

Moral luck is the degree to which luck affects the moral status of an action or person. The *problem* of moral luck, then, is how to reconcile the intuitive difference between lucky and unlucky fathers and drivers with the idea that people cannot be blame or praiseworthy for things beyond their control. Brute outcomes should make no moral difference.

Normatively speaking, the issue comes down to whether we should think moral luck is real, or whether it is a mistake to let luck play any role in our moral assessment of people and their actions. Some have argued that moral luck is the result of hindsight bias: after the fact, people think that an outcome was more likely to happen simply because it did happen, which biases their moral verdict. Others have favored various forms of epistemic reductionism (Schinkel 2009); moral luck intuitions could be explained by the fact that what we are after when we make moral judgment is people's intentions, but that we use outcomes as *evidence* for people's intentions. Alternatively, these intuitions may be based on knowledge attributions; unlucky drivers and fathers hold false beliefs about the future outcomes of their actions, which may make us view them as more morally blameworthy (Young et al. 2010).

How do these explanations bear on *the gap*? Recently, people have turned to an evolutionary perspective for answers. Here, the idea is that blame and punishment serve an adaptive function: they are supposed to provide a learning environment that favors cooperation and pro-social dispositions at the expense of free-riding and antisocial tendencies. Now, the empirical evidence suggests that only rigid punishment based on outcomes rather than intentions or the goal of deterrence can do this (Cushman 2008, 2013, 2015). Perpetrators can deceive others about their intentions, which always remain somewhat opaque; moreover, they can strategically disincentivize punishment by indicating that they are unwilling to learn, thereby ruling out deterrence as a possible rationale for punishing. Only outcome-based punishment escapes these two problems. Sensitivity to resultant luck thus makes evolutionary sense.

This suggests that moral luck is justified for consequentialist reasons which used to obtain in our environment of evolutionary adaptedness (Kumar 2017). Interestingly, some people have used similar evolutionary arguments to make the opposite point: in assigning blame, it used to make sense to rely on proxies for potential wrongdoers' mental states which are hard, if not impossible, to access directly (Levy 2016). However, this also means that whenever we have more direct and reliable evidence regarding people's mental states, these more informed judgments should trump those which are based on less trustworthy proxies.

Moral Intuition

(viii) *Rationalism and Sentimentalism*. Should we think of the influence of moral judgments on seemingly nonmoral concepts as a pernicious one? Obviously, this does not merely depend on the relevance of moral judgments for those other cognitive domains, but also on whether moral judgments themselves have a sound basis.

For an astonishingly long time, philosophers have thought that the question whether moral judgments can be trusted or not could be substituted for the question whether these judgments were based on emotion or reason. Some sentimentalists, such as Hume, thought moral judgments had to be grounded in the former. Reason, his argument went, was in the business of determining facts; moral judgments, on the other hand, were capable of motivating people to act. But, Hume also argued, only feelings and desires have such motivational force; and since feelings and desires do not have the right "direction of fit" (Smith 1987), they are not in the business of determining facts. Hence, moral judgments could not be based on reason. Others, such as Kant, argued that this could not be true, since moral judgments were supposed to have unconditional authority, which emotion could not deliver. They thus went looking for a purely rational justification of moral requirements that was cleansed of all emotional impurity.

I say "astonishingly long time" because on closer inspection, the idea that reason and emotion are somehow opposed forces has little to commend it and tends to evaporate rather quickly. And yet for the most part, empirically informed philosophers have not just sided with the sentimentalist tradition (Nichols 2004; Prinz 2006, 2007), but continued to dress up their sentimentalism—the claim that moral judgments are based on emotion—as an alternative to rationalism.

As far as the empirical evidence is concerned, this meant showing that emotions do not merely accompany moral judgments, but properly constitute them. One way to do this is to show that reasoning doesn't produce moral judgments. Emotionally charged intuitions take primacy, which reason merely rationalizes after the fact. When people's reasoning is debunked, they tend not to give up their moral intuitions, but enter a state of "moral dumbfounding" (Haidt 2001). It is true in general that people only have poor introspective access into what drives their moral judgments (Uhlmann et al. 2009; Hall et al. 2012). Moreover, emotions seem to be both necessary and sufficient for moral judgment (Prinz 2006). Evidence from psychopathic individuals suggests that impaired emotion leads to impaired moral judgment (Blair 1995). Emotion manipulation studies seem to demonstrate that changing people's emotions changes their moral beliefs as well (Schnall et al. 2008; Wheatley and Haidt 2005; Valdesolo and DeSteno 2006). Then again, more recent studies suggest that psychopaths, though suffering from diminished empathy, guilt, and remorse, are indeed able to draw the distinction between moral and conventional norms (Aharoni et al. 2012). The aforementioned emotion manipulation studies, in turn, are problematic in that they focus on very specific subgroups of the population (e.g., highly hypnotizable subjects), find statistically significant effects

only for some vignettes, and, perhaps most importantly, fail to alter the polarity of people's moral beliefs (e.g., from "X is right" to "X is wrong"; May 2014). But even if it had been shown that moral judgments are thoroughly saturated with emotion, it remains unclear why this would have any implications for how trustworthy they are (Sauer 2012a).

(ix) *Evolutionary Debunking*. What other grounds, besides an obsolete commitment to the incompatibility of emotion and reason and the shaky evidence adduced to support it, are there for believing that moral intuition may be a poor guide to the moral truth?

Evolution—of course. I have already mentioned one example for how evolutionary considerations can be used to undermine a subset of moral intuitions: the Greene/Singer strategy of debunking deontological intuitions as alarm-like responses to morally irrelevant factors such as up-close-and-personal harm that were selected for in an environment we no longer inhabit (see section (i) above).

But so-called evolutionary debunking arguments (Kahane 2011) can be generalized to cover all moral judgments. The basic strategy is this: many, if not all, of our moral judgments can in some way be traced back to a few basic evaluative dispositions. We want to avoid pain, punish evildoers, sympathize with vivid suffering, care about our kin, like to reciprocate favors, and dislike cheaters. It is overwhelmingly plausible that evolution has something to do with why we hold these values and not their opposites, or something else entirely (such as "the fact that something is purple is a reason to scream at it," Street 2006, 133). Now, suppose there are certain objective moral facts: facts about right and wrong, or about what we have most moral reason to do. How likely is it that we are in a position to know these facts when relying on our basic evaluative dispositions?

Spectacularly unlikely, some have argued (Joyce 2006). In fact, it would be pure serendipity for our moral beliefs to hit upon the moral truth by accident, given that the mechanism that shaped the dispositions we rely upon in making those judgments bore no connection whatsoever to their truth. Evolutionary pressures select for traits which are adaptive; but unlike in the nonmoral case, where false beliefs can get you killed, moral beliefs don't have to be true to allow you (and a fortiori your genes) to survive. Unless we have something else to go on—which we do not—this insight thoroughly undermines our moral intuitions.

I cannot summarize the rich literature on this topic here, so let me just hint at some possible responses, always keeping an eye on *the gap*. Evolutionary debunking arguments pose a reliability challenge—the processes that produce our moral judgments do not aim at truth, but at increasing the frequency of our genes in a given population. Now, some have argued that the evolutionary challenge can be met (Huemer 2005; Fitzpatrick 2014): our capacity to make moral judgments may be the upshot of a more general capacity, such as reason or intelligence, for which there *is* an evolutionary rationale. Some have tried to show that the challenge overgeneralizes in various unwelcome ways. After all, what reason is there to believe that evolution has given us the capacity to recognize mind-independent mathematical truths (Clarke-Doane 2012)? Some have

suggested that the challenge can be redirected. According to evolutionary debunkers, moral judgments are produced by off-track processes. But what if there is no track to be *on* at all? If moral judgments do not aim at discerning any mind-independent moral truths to begin with, then the threat of moral skepticism is disarmed (Street 2006). Finally, some have argued that there is a class of moral beliefs that remains immune to debunking, because it cannot be explained on evolutionary grounds (de Lazari-Radek and Singer 2012). An attitude of universal and impartial benevolence, for instance, seems to confer no fitness benefits. The debate on evolutionary debunking shows, at any rate, how tightly connected normative and so-called *metaethical* questions regarding the nature of moral values and value judgments are.

(x) *The Reliability of Intuition*. Distal causes such as evolution are not the only ones to cast doubt on the trustworthiness of our moral intuitions. Proximal ones, such as the susceptibility of those intuitions to irrelevant features of the situation, seem to provide more direct and less speculative grounds for skepticism toward the reliability of moral cognition.

For instance, people's moral beliefs appear to be subject to order effects (Liao et al 2012; Schwitzgebel and Cushman 2012). For instance, subjects are more likely to judge it permissible to push a person to her death to save five others when the respective scenario was presented before a similar one in which a runaway trolley had to be redirected using a switch to achieve the same result. This effect holds even for professional philosophers among which some familiarity with the scenarios given can be presumed. Framing effects, in which people's moral judgments are affected by how and in what context an option is presented, are also frequently cited as an unwelcome influence on our moral thinking (Sinnott-Armstrong 2008).

These findings lead us to a possible skeptical argument. In making moral judgments, we rely on moral intuitions. But if, as the evidence suggests, these intuitions are sensitive to morally extraneous factors the presence of which we are frequently unaware of and sometimes cannot rule out, then our intuitions require confirmation. But the only thing we have to confirm our moral intuitions are *more moral intuitions*. The justification of our moral beliefs seems to have no hinges to turn on.

How unreliable do framing effects make moral judgments? According to one very reasonable measure of reliability, the mean probability that a subject will *not* change her moral judgment depending on framing or order is 80%—not so bad (Demaree-Cotton 2016; cf. Andow 2016). Moreover, as in the case of emotion manipulation studies more generally, effect sizes tend to be small, and framing effects rarely alter the polarity of people's judgments. That is to say, subjects' judgments are somewhat affected by being in one frame or another, but people do not, strictly speaking, change their minds.

Moreover, debunking arguments aiming to show that moral intuitions are unreliable face one crucial limitation: they rely on moral intuitions themselves, in particular regarding which factors count as morally irrelevant and which do not (Rini 2015). In order for such arguments to get off the ground, then, at least some moral judgments *must* be considered reliable.

Bridging the Gap

The guiding question of this chapter was: given the empirical presuppositions of normative theories of moral judgment and agency, what is the normative significance of empirical facts about our moral psychology? Most importantly, how should we think about the relationship between the two in light of *the gap*? Let's provide at least a tentative answer to this question.

I have surveyed a variety of topics that moral psychologists and empirically informed philosophers are currently working on, ranging from more specific issues such as which normative ethical theory fares best in light of empirical scrutiny, to whether human beings tend to have what it takes to satisfy the requirements of moral agency, to the influence of moral judgment on nonmoral thinking and, finally, the reliability of moral cognition in general.

It is rather clear that, though empirical research has no *direct* normative implications, there are ways to make empirical research normatively *relevant*. Empirical information always needs to be coupled with normative bridging principles to develop genuine moral impact. Note, however, that this is not an indictment of empirically informed moral philosophy, as the situation is exactly symmetrical with respect to purportedly "pure" normative inquiry, which equally fails to have any genuine normative implications unless coupled with empirical bridging principles that connect it to the real world.

In addition to the three positive ones mentioned below, I have one negative lesson to offer about trying to make empirical data normatively significant. It may seem trivial, but is easily—and frequently—ignored: avoid hasty, sweeping generalizations. Claims such as "moral intuitions are unreliable/reliable," "people are free/there is no such thing as free will," or "people are essentially good/bad" are unlikely to be true unless appropriately qualified to add nuance, in which case the bolder version of the claim turns out to be not just untrue and imprecise, but also unhelpful. Rather, empirically informed normative inquiry should be conducted in a piecemeal fashion. Exactly how, and to what extent, are intentionality attributions driven by normative judgment? How strong is the influence of framing effects on moral beliefs? In what sense may people have stable or fragmented personality traits, and how do they manifest? How does human decision-making work, when does it break down, and what causes it to do so? These complex questions cannot be answered with bold, attention-grabbing slogans—not correctly, at any rate.

Here is the first positive lesson I believe can be drawn: empirical data can develop normative relevance by *undermining the empirical presuppositions of various normative ethical theories regarding what kind of creature we are*. This means that when it comes to *the gap*, the *ought implies can* principle is at least as important as the *no ought from an is* principle. If we literally cannot act in the way postulated by a moral theory, then it cannot be the case that we ought to act in that way. To be sure, it is true that moral theories are not in the business of merely describing the world. Ultimately, normative inquiry is about what is good or right, and the normative power of the factual only goes so far. But it makes little sense to come up with fancy

ideals no one can bring herself to care about, while ignoring the things we do care about because they do not comport with the clever principles we came up with in our study. This point has been very clearly articulated by Owen Flanagan (1991), who calls it the "principle of minimal psychological realism" (32ff.). We see it at work, for instance, in sections (iii) and (iv) above.

My second lesson has it that empirical moral psychology can uncover that *the etiology of our moral intuitions sometimes undermines their justification*. Psychological debunking arguments of this sort all share the same basic structure: (1) There is a class C of moral judgments that is generated by cognitive process P. (2) P is unreliable with respect to C. (3) C is unjustified. (Or, alternatively, a subject S would be unjustified in holding a belief out of C if S arrived at that belief on the basis of P.)

Actually, debunking arguments are a motley bunch rather than a monolithic strategy. All debunking arguments try to show that a given belief has been generated by dubious processes. But there are various ways of spelling out this dubiousness. It is useful to distinguish six different types of debunking: (a) off-track debunking: a moral belief is based on a cognitive process that does not track the (moral) truth, e.g., evaluative tendencies that are evolutionarily adaptive, but not morally trustworthy (see section (ix) above). (b) Hypersensitivity debunking: many moral judgments are driven by feelings of disgust. But disgust is a hypersensitive "better safe than sorry" mechanism that generates an unhealthy amount of false positives and should thus be viewed with skepticism (Kelly 2011). (c) Hyposensitivity debunking: empathy is the (potential) source of at least as many moral judgments as disgust. But empathy is a hyposensitive mechanism that generates many false negatives due to its inherent partiality toward the near and dear (Prinz 2011). (d) Obsoleteness debunking: some judgmental processes used to be epistemically viable, but no longer are because the natural and social scaffolding they used to fit has disappeared. Our intuitive morality has been shaped to deal with the demands of stable, intimate, small-scale tribal groups in the Pleistocene. We are ill-equipped to deal with environments very unlike this one—namely, the one we currently happen to live in (Greene 2013). (e) Inconsistency debunking: in some cases, we can build inconsistent pairs of moral judgments, one or both of which we thereby know has to be given up because the difference between the two moral judgments may be based on nothing but a morally irrelevant factor (Campbell and Kumar 2012). (f) Ignoble origins debunking: this is the "original" type of debunking made famous by nineteenth (and early twentieth) century renegades such as Marx, Nietzsche, and Freud. It aims to uncover the ugly distal history of certain moral views by showing that they originated in processes, events, or dispositions that are either inherently undesirable or at least inconsistent with the targeted moral outlook. Christianity preaches love and compassion, but is founded on resentment and envy; capitalism is founded on the ideal of equal rights and fairness, but these ideals actually just serve the interests of the ruling class; and so on (Prinz 2007, 215ff.). The power of debunking arguments is discussed in sections (i), (ii), and (viii)–(x).

A final lesson is this: often, but certainly not often enough, empirical information can develop normative significance by enabling us to use this information for the *reflexive improvement of moral judgment and agency* (Rini 2013). We cannot discount

implicit biases unless we know how, why, when, and under what conditions they operate. Empirical research can tell us when and how the tools we wish to deploy in moral cognition and action are unsuitable or likely to be broken. Sections (i), (ii), and (v)–(x) nicely illustrate the usefulness of this lesson.

The problem is that we have no way of knowing introspectively when this is the case. In fact, we have no way of knowing, in general, what causes our thoughts and desires, and our folk theories of how our thinking works are often hopelessly inadequate. Empirical research is essential for this reflexive purpose, and ignoring or dismissing it reckless and foolish.

References

Aharoni, E., Sinnott-Armstrong, W., & Kiehl, K. A. (2012). Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology, 121*(2), 484.

Alfano, M. (2013). *Character as moral fiction*. Cambridge University Press.

Alfano, M. (2016). *Moral psychology: An introduction*. Cambridge: Polity.

Alfano, M., & Loeb, D. (2014). Experimental moral philosophy. *Stanford Encyclopedia of Philosophy*, 1–32.

Andow, J. (2016). Reliable but not home free? What framing effects mean for moral intuitions. *Philosophical Psychology, 29*(6), 904–911.

Appiah, A. (2008). *Experiments in ethics*. Cambridge, MA: Harvard University Press.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54*, 462–479.

Baumeister, R. F., Masicampo, E. J., & Nathan DeWall, C. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*(2), 260–268.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language, 25*(4), 474–498.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs, 37*(4), 293–329.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition, 57*(1), 1–29.

Brink, D. O. (1984). Moral realism and the sceptical arguments from disagreement and queerness. *Australasian Journal of Philosophy, 62*(2), 111–125.

Campbell, R., & Kumar, V. (2012). Moral reasoning on the ground. *Ethics, 122*(2), 273–312.

Clark, C. J., et al. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology, 106*(4), 501–513.

Clarke-Doane, J. (2012). Morality and mathematics: The evolutionary challenge. *Ethics, 122*(2), 313–340.

Cova, F., & Naar, H. (2012). Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology, 25*(6), 837–854.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.

Cushman, F. (2013). The role of learning in punishment, prosociality, and human uniqueness. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its evolution*. Cambridge, MA: MIT Press.

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass, 10*(2), 117–133.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*(2), 353–380.

Darley, J. M., & Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*(1), 100–108.

Darwall, S., Gibbard, A., & Railton, P. (1992). Toward Fin de Siècle Ethics: Some trends. *Philosophical Review, 101*(1), 115–189.

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology, 29*(1), 1–22.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Doris, J. M. (2009). Skepticism about persons. *Philosophical Issues, 19*(1), 57–91.

Doris, J. M., & Murphy, D. (2007). From My Lai to Abu Ghraib: The moral psychology of atrocity. *Midwest Studies in Philosophy, 31*(1), 25–55.

Doris, J., & Plakias, A. (2008). How to argue about disagreement: Evaluative diversity and moral realism. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, *The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 303–331). Cambridge, MA: MIT Press.

Doris, J. M., & Stich, S. P. (2005). As a matter of fact: Empirical perspectives on ethics. In F. Jackson & M. Smith (Eds.), *The Oxford handbook of contemporary philosophy*. Oxford: Oxford University Press.

Enoch, D. (2009). How is moral disagreement a problem for realism? *The Journal of Ethics, 13*(1), 15–50.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Fitzpatrick, S. (2014). Moral realism, moral disagreement, and moral psychology. *Philosophical Papers, 43*(2), 161–190.

Flanagan, O. J. (1991). *Varieties of moral personality: Ethics and psychological realism*. Cambridge, MA: Harvard University Press.

Foot, P. (2001). *Natural goodness*. Oxford: Oxford University Press.

Fraser, B., & Hauser, M. (2010). The argument from disagreement and the role of cross-cultural empirical data. *Mind & Language, 25*(5), 541–560.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 3. The neuroscience of morality: Emotion, brain disorders, and development*. Cambridge, MA: MIT Press.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin Press.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (Neuro)science matters for ethics. *Ethics, 124*(4), 695–726.

Greene, J. D., et al. (2014). Are "counter-intuitive" deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Social Cognitive and Affective Neuroscience, 9*(9), 1368–1371.

Greene, J. D., Nystrom, L. E., et al. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greene, J. D., Sommerville, B. D., et al. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*(1), 128–133.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS One, 7*(9), e45457.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society, 99*(1999), 315–331.

Harman, G. (2009). Skepticism about character traits. *The Journal of Ethics, 13*(2/3), 235–242.

Holton, R. (2010). Norms and the Knobe effect. *Analysis, 70*(3), 1–8.

Huemer, M. (2005). *Ethical intuitionism*. New York: Palgrave Macmillan.

Hume, D. (1739/2000). *A treatise of human nature*. Oxford: Oxford University Press.
Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.
Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology, 21*(3), 384–388.
Joyce, R. (2006). *The evolution of morality*. Cambridge: MIT Press.
Kahane, G. (2011). Evolutionary debunking arguments. *Noûs, 45*(1), 103–125.
Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind & Language, 27*(5), 519–545.
Kahane, G., et al. (2012). The neural basis of intuitive and counterintuitive moral judgement. *Social Cognitive and Affective Neuroscience, 7*(4), 393–402.
Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209. doi:10.1016/j.cognition.2014.10.005.
Kahneman, D. (2011). *Thinking, fast and slow*. London: Macmillan.
Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. New York: Oxford University Press.
Kelly, D. (2011). *Yuck!: The nature and moral significance of disgust. A Bradford book*. Cambridge, MA: MIT Press.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(3), 190–194.
Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33*(4), 315–329.
Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 2). Cambridge, MA: MIT Press.
Knobe, J., & Leiter, B. (2007). The case for Nietzschean moral psychology. In B. Leiter & N. Sinhababu (Eds.), *Nietzsche and morality*. Oxford: Oxford University Press.
Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.
Kristjánsson, K. (2012). Situationism and the concept of a situation. *European Journal of Philosophy, 20*(S1), E52–E72.
Kumar, V. (2017). Moral vindications. *Cognition*. Vol. 167, 124–134.
Kumar, V. (forthcoming). The ethical significance of cognitive science. In S.-J. Leslie & S. Cullen (Eds.), *Current controversies in philosophy of cognitive science*. Routledge.
de Lazari-Radek, K., & Singer, P. (2012). The objectivity of ethics and the unity of practical reason. *Ethics, 123*(1), 9–31.
Leiter, B. (2007). Against convergent moral realism: The respective roles of philosophical argument and empirical evidence. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, *The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 333–337). Cambridge, MA: MIT Press.
Levy, N. (2015). Less blame, less crime? The practical implications of moral responsibility skepticism. *Journal of Practical Ethics, 3*(2), 1–17.
Levy, N. (2016). Dissolving the puzzle of resultant moral luck. *Review of Philosophy and Psychology, 7*(1), 127–139.
Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology, 25*(5), 661–671.
Libet, B. W. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), 529–566.
Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. New York: Penguin.
Martin, R., & Barresi, J. (2003). Personal identity and what mattes in survival: An historical overview. In R. Martin & J. Barresi (Eds.), *Personal identity*. Oxford: Blackwell.
Mason, E. (2013). Objectivism and prospectivism about rightness. *Journal of Ethics and Social Philosophy, 7*(2).
May, J. (2014). Does disgust influence moral judgment? *Australasian Journal of Philosophy, 92*(1), 125–141.
Merritt, M. (2009). Aristotelean virtue and the interpersonal aspect of ethical character. *Journal of Moral Philosophy, 6*(1), 23–49.
Meyers, C. D. (2013). Defending moral realism from empirical evidence of disagreement. *Social Theory and Practice, 39*(3), 373–396.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.
Miller, C. (2003). Social psychology and virtue ethics. *The Journal of Ethics, 7*(4), 365–392.
Moore, G. E. (1903). *Principia Ethica*. New York: Dover Publications.
Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations, 9*(2), 203–219.
Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39*(1), 96–125.
Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.
Nichols, S. (2014). Process debunking and ethics. *Ethics, 124*(4), 727–749.
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663–685.
Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South*. Boulder, CO: Westview Press.
Paxton, J. M., Ungar, L., et al. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.
Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*(5), 586–604.
Phillips, J., Misenheimer, L., & Knobe, J. (2011). The ordinary concept of happiness (and others like it). *Emotion Review, 71*(3), 929–937.
Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations, 9*(1), 29–43.
Prinz, J. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.
Prinz, J. (2009). The normativity challenge: Cultural psychology provides the real threat to virtue ethics. *The Journal of Ethics, 13*(2–3), 117–144.
Prinz, J. (2011). Against empathy. *Southern Journal of Philosophy, 49*(s1), 214–233.
Rini, R. A. (2013). Making psychology normatively significant. *The Journal of Ethics, 17*(3), 257–274.
Rini, R. A. (2015). Morality and cognitive science. *Internet Encyclopedia of Philosophy*.
Robinson, B., Stey, P., & Alfano, M. (2015). Reversing the side-effect effect: The power of salient norms. *Philosophical Studies, 172*(1), 177–206.
Ross, L., & Nisbett, R. E. (1991). *The person and the situation*. Philadelphia: Temple University Press.
Sauer, H. (2012a). Psychopaths and filthy desks: Are emotions necessary and sufficient for moral judgment? *Ethical Theory and Moral Practice, 15*(1), 95–115.
Sauer, H. (2012b). Morally irrelevant factors: What's left of the dual process-model of moral cognition? *Philosophical Psychology, 25*(6), 783–811.
Sauer, H. (2014). It's the Knobe effect, stupid! *Review of Philosophy and Psychology, 5*(4), 485–503.
Sauer, H., & Bates, T. (2013). Chairmen, cocaine, and car crashes: The Knobe effect as an attribution error. *The Journal of Ethics, 17*(4), 305–330.
Schinkel, A. (2009). The problem of moral luck: An argument against its epistemic reduction. *Ethical Theory and Moral Practice, 12*(3), 267–277.
Schlosser, M. E. (2012a). Free will and the unconscious precursors of choice. *Philosophical Psychology, 25*(3), 365–384.
Schlosser, M. E. (2012b). Causally efficacious intentions and the sense of agency: In defense of real mental causation. *Journal of Theoretical and Philosophical Psychology, 32*(3), 135–160.
Schlosser, M. E. (2014). The neuroscientific study of free will: A diagnosis of the controversy. *Synthese, 191*(2), 245–262.
Schnall, S., et al. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*(8), 1096–1109.
Schultze-Kraft, M., et al. (2016). The point of no return in vetoing self-initiated movements. *Proceedings of the National Academy of Sciences, 113*(4), 1080–1085.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27*(2), 135–153.
Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics, 9*(3–4), 331–352.
Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, *The cognitive science of morality* (Vol. 2, pp. 47–76). Cambridge, MA: MIT Press.
Smith, M. (1987). The human theory of motivation. *Mind, 96*(381), 36–61.
Sneddon, A. (2009). Normative ethics and the prospects of an empirical contribution to the assessment of moral disagreement and moral realism. *Journal of Value Inquiry, 43*(4), 447–455.
Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience, 11*(5), 543–545.
Sripada, C., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language, 26*(3), 353–380.
Stanovich, K. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies, 127*(1), 109–166.
Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 113*(2014), 159–171.
Tiberius, V. (2014). *Moral psychology: A contemporary introduction*. Routledge.
Tobia, K. P. (2015). Personal identity and the phineas gage effect. *Analysis, 75*(3), 396–405.
Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision making, 4*(6), 479.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476–477.
Velleman, J. D. (2011). *How we get along*. Cambridge: Cambridge University Press.
Vranas, P. B. (2005). The indeterminacy paradox: Character evaluations and human psychology. *Noûs, 39*(1), 1–42.
Webber, J. (2013). Character, attitude and disposition. *European Journal of Philosophy, 21*(1), 1082–1096.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*(10), 780–784.
Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235–8240.
Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck. *Review of Philosophy and Psychology, 1*(3), 333–349.
Zimmerman, M. J. (2011). *The immorality of punishment*. Buffalo, NY: Broadview Press.