Why Moral Psychology is Disturbing

Regina A. Rini ▪ NYU Bioethics
gina.rini@nyu.edu

Abstract Learning the psychological origins of our moral judgments can lead us to lose confidence in them. In this paper I explain why. I consider two explanations drawn from existing literature – regarding epistemic unreliability and automaticity – and argue that neither is fully adequate. I then propose a new explanation, according to which psychological research reveals the extent to which we are disturbingly disunified as moral agents.

Please note: This is the penultimate draft of a paper that has been accepted for publication in *Philosophical Studies*. This version has not yet been through final copy-editing and may not reflect certain changes in the final published version.

1. **Moral judgment and moral psychology**

We rarely think about the psychological causes of our moral judgments, and perhaps this is wise. Sometime this information is disturbing. One minute I am thinking: 'obviously it is wrong to support that tax policy'. Then the next moment I am told by a developmental psychologist: 'well, you believe that because you went to Catholic elementary school'. Or this by a neuroscientist: 'you wouldn't believe that if you didn't have such strange brain chemistry'. And now I am taken aback, if only for a moment. What if they are right? What if those *are* the causes of my moral judgment? It seems to me as if I have good reasons for my judgment – but if the psychological processes had turned out differently I'd be just as confident in reasons for a different judgment.

This paper discusses the disturbing feeling that sometimes follows psychological information about the origin of our moral judgments. There is a long-running debate about whether or not we *should* be disturbed by learning such information, and whether being disturbed should lead us to revise or reject any of our judgments. But this paper does not take sides in that debate. Rather, it aims to clarify the experience at the center of the debate. I will explain why we are sometimes disturbed by learning the psychological causes of our moral judgments. I will propose that moral psychological research is disturbing because it often reveals an unsettling lack of unity between our conscious selves and the nonconscious evaluative processes that make up our minds.

The structure of the paper is as follows: In section 2 I characterize the sort of disturbing experience at issue. In sections 3 and 4 I discuss two seemingly plausible explanations for it: that psychological

research undermines the epistemic credentials of our moral intuitions, and that psychological research suggests our moral choices are automatic and uncontrolled. Both of these explanations are partly correct, but they leave something out. In section 5 I introduce a new account, centered on the discovery that we are *disunified moral agents*.

2. Doxastic embarrassment

2.1 Moral phenomenology

In this section I will characterize a certain phenomenological state.[1] But characterizing phenomenology is a tricky business, especially in the moral domain (Chappell 2013). Language does not always provide us the best tools for communicating the details of our inner lives. The best I can do is provide a description of an experience that I have had and hope that my description strikes close enough to a similar experience of yours. You will have to fill in the rest of the details through introspection and imagination.

It will be helpful to have a label for the experience I am about to describe. I will call it 'doxastic embarrassment'.

2.2 An instance of doxastic embarrassment

I begin in possession of some moral beliefs. For instance: I believe that it if I am standing beside some tracks and can prevent an out-of-control trolley from running over five innocent people by diverting the trolley to a sidetrack where it will kill only one, then it is morally permissible for me to do so. On the other hand, I believe, it is not morally permissible to stop a similarly out-of-control trolley by physically shoving a large man onto the tracks from an overpass.[2]

I am aware that I hold these moral beliefs and, on the whole, I am content with holding them. I have, let us say, a second-order attitude of approval toward them. On the whole, I think it is rational for me to maintain these beliefs. Or, at least, I do not think it is irrational for me to maintain them.

[1] By 'phenomenological' I do not mean *perceptual* phenomenology; I am not talking about colors or smells. What I will describe might be better called *intellectual phenomenology*: it pertains to the experience of a certain sequence of beliefs and emotions. To call this phenomenological is merely to stress that the immediate interest is about *what it is like* to go through this experience.

[2] These cases, of course, constitute the famous 'Trolley Problem' (Foot 1967; Thomson 1976).

Although I am aware that I have these beliefs and approve of having them, I do not quite know *why* I have them, causally speaking. I have no particular theory about how I came to believe these things, nor does my approval of them seem to depend on any causal theory.

Now something new happens. I learn of a psychological experiment that purports to reveal surprising facts about the causal origin of moral beliefs like mine. Let's say that I learn about some neuroscientific findings (Greene et al. 2004). It seems that when people judge it morally permissible to flip a switch to divert the trolley, their brains exhibit greater activity in areas correlated with cold cognition. By contrast, when people judge it impermissible to shove the large man from the footbridge, their brains show greater activity in emotional areas.[3] There is also a speculative evolutionary hypothesis about these beliefs, which runs as follows: people have an emotional revulsion to shoving the large man from the footbridge because it was evolutionarily adaptive for primates like us to find such acts of violence aversive. Shoving the large man involves 'up close and personal' violence, a sort of violence that our distant ancestors needed to avoid in order to make their social system works. By contrast, flipping a switch to divert a spatially distant object is not a form of action that would have been familiar to our ancestors. We could not have evolved an aversion to it. Hence, goes the speculative hypothesis, we react differently to the two cases because only one of them triggers the 'up close and personal' violence detector in our brains.[4]

So I hear about this psychological research, and I accept its empirical conclusions. That is, I am convinced that the research describes facts about the causal origins of my moral judgments. Upon accepting the empirical claims, I find that I am disturbed. I think something like this: 'All this time I thought it was reasonable to have different reactions to the trolley and the footbridge. But now I find out that I only thought this way because my brain is surprisingly primitive: some ancient brain bit can understand up-close-and-personal matters like shoving someone, but it can't understand new-fangled railroad switching technology. Is this really *all* it comes down to? Am I that simple – are my moral beliefs just artifacts of simian psychology?'

[3] Here I will leave Greene's *empirical* claims unchallenged, though certainly others have challenged them. See e.g. Kahane and Shackel (2010); Kahane et al. (2015).

[4] This speculative evolutionary account comes from Greene (2008) and Singer (2005). It is worth noting that Greene has amended his claims about the nature of the 'up close and personal' mechanism (Greene 2014), though the details do not matter for this example.

Once I have this thought, I begin to doubt these moral beliefs. I am no longer sure that it is rational for me to have them. I do not know which other beliefs I should have instead, but I am no longer confident that I should have *these.* Accepting the psychological account of these beliefs has *changed* my relationship to them.

2.3 Other examples

Importantly, the experience I have just described is not hypothetical. It is how I *actually reacted* to learning about Joshua Greene's studies (at least when I set aside doubts about his empirical claims). And this does not seem to be a personal eccentricity; other people have reacted similarly.[5] Of course, *you* may not have reacted to Greene's research the way I did. You may not have felt any wavering about your reactions to the trolley and the footbridge. Since it will be useful to have a clear example in mind, I offer a few other potential triggers of doxastic embarrassment:

- After hearing a clip of inoffensive stand-up comedy you are more likely to regard deontological moral violations as permissible (Strohminger, Lewis, and Meyer 2011).
- If you deliberate with your eyes closed, you are likely to condemn cheating more harshly (Caruso and Gino 2011).
- If you have been holding dirty banknotes recently, you are more likely to be uncooperative and ungenerous in economic exchange (Yang et al. 2013).
- After unscrambling sentences that prime 'science' concepts, you are likely to issue harsher moral verdicts (Ma-Kellams and Blascovich 2013).

There are many, many more that I could list. Not every example will bother every reader; not all of them bother me. If *none* of these bother you, and if you cannot think of any other example of the experience, then the rest of this paper may not be of direct use to you. But the recent explosion in philosophical discussion of the psychological debunking of morality suggests that this sort of experience is fairly common; a somewhat inchoate anxiety seems to attend these discussions.[6] Even if you have not

[5] Peter Singer thinks so as well, judging by his rhetoric: "[W]hat is the moral salience of the fact that I have killed someone in a way that was possible a million years ago, rather than in a way that became possible only two hundred years ago? I would answer: none." (Singer 2005, 348). This does not appear to be an argument, unless Singer expects his reader to share his reaction.

[6] See Doris and Stich (2007), Appiah (2008), Kauppinen (2007), Berker (2009), and others discussed below.

personally experienced that anxiety, perhaps the following explanation will help you understand those who do.

2.4 The significance of doxastic embarrassment

Importantly, all I have done so far is *describe* an experience. I have *not* claimed that this experience is a stopping point, that it settles anything. Most especially, I have not presented an *argument*. My worries about simian psychology are not premises logically entailing a conclusion. They are merely thoughts that seem to have *triggered* my new lack of confidence in my moral beliefs. But it may be that my reaction is irrational and I *should not* ultimately allow myself to be scared off my moral beliefs in this way.

Philosophers have considered this idea before. Here is Charles Stevenson (writing in the unfortunate idiom of his time):

> If certain of our attitudes are shown to have the same origin as the taboos of savages, we may become disconcerted at the company we are forced to keep. After due consideration, of course, we may decide that our attitudes, however they may have originated, are unlike many taboos in that they will retain a former function, or have since acquired new ones. Hence we may insistently preserve them. But in the midst of such considerations we shall have been led to see our attitudes in a natural setting, and shall be more likely to change them with changing conditions. Hence anyone who wants to change a man's attitudes can prepare the way by a genetic study. (Stevenson 1944, 123-124)

Note that while Stevenson says we "may become disconcerted" by psychological claims about our moral beliefs, he thinks that we can come to regain our confidence in them "after due consideration". This is a general theme. Even Nietzsche wrote:

> The inquiry into the *origin of our evaluations* and tables of the good is in absolutely no way identical with a critique of them, as is so often believed: even though the insight into some *pudendo origo* certainly brings with it *a feeling* of a dimunition in value of the thing that originated thus and prepares the way to a critical mood and attitude toward it. (Nietzsche 1901/1973 §254)

Both Nietzsche and Stevenson talk about doxastic embarrassment as a way to "prepare" for criticisms or change of moral attitudes. Both see it as a way to soften our attachments to our moral beliefs, not yet a sufficient reason actually to reject them.

Should doxastic embarrassment lead us to change our moral beliefs? Many philosophers think not. Thomas Nagel writes, "someone who abandons or qualifies his basic methods of moral reasoning on historical or anthropological grounds alone is nearly as irrational as someone who abandons a mathematical belief on other than mathematical grounds." (Nagel 1997, 105) In Nagel's view, moral thinking constitutes its own authoritative domain, such that merely learning new empirical facts *about* moral thinking should not (rationally speaking) lead to change. On this view, doxastic embarrassment is irrational.[7]

Similarly, Ronald Dworkin argues that *whatever* the psychological truth might be, it does not immediately give us reason to revise our moral beliefs. Imagine that science shows your beliefs about justice to actually be caused by processes that reflect only your own egoistic concerns. Dworkin does not think this is a reason to change your beliefs about justice:

> Your opinion is one about justice, not about your own psychological processes… You lack a normative connection between the bleak psychology and any conclusion about justice, or any other conclusion about how you should vote or act. (Dworkin 1996, 124-125)

So there is a live question about whether my experience of doxastic embarrassment reflects a dawning awareness of difficulties for my moral beliefs, or a sort of irrational quaver that will rightly fade upon reflection. I will not try to settle that question here.[8] The rest of this paper aims to *explain* doxastic embarrassment. When we have a clear explanation, it will be much easier to address the rationality question. I now turn to discussing two extant explanations, before proposing my own.

3. The Explanation from Epistemic Unreliability

3.1 Epistemically unreliable moral judgments

Much of the recent discussion of moral psychology focuses on the epistemic status of moral intuitions. It is claimed that empirical findings show our moral judgments to be unreliable guides to moral truth. Here

[7] There are also general epistemic reasons, not restricted to the moral domain, to resist causal debunking of our beliefs (White 2010). However, since it is not universally agreed that moral judgment does or should obey the same epistemic standards as other reasoning domains, I will restrict my discussion to distinctively moral judgment.
[8] Elsewhere I do, in effect, defend the rationality of some cases of doxastic embarrassment (Rini 2013). But the argument of this paper does not rely on that one.

is Walter Sinnott-Armstrong's 'Master Argument' against moral intuitionism (Sinnott-Armstrong 2008, 52):

> SA1) If our moral intuitions are formed in circumstances where they are unreliable, and if we ought to know this, then our moral intuitions are not justified without inferential confirmation.
> SA2) If moral intuitions are subject to framing effects, then they are not reliable in those circumstances.
> SA3) Moral intuitions are subject to framing effects in many circumstances.
> SA4) We ought to know (SA3).
> SA5) Therefore, our moral intuitions in those circumstances are not justified without inferential confirmation.[9]

According to Sinnott-Armstrong, there is psychological evidence that some of our moral intuitions are caused by framing processes; our intuitions are affected by the order in which we read about cases, or the particular words used to describe cases (SA3). Presumably these factors have nothing to do with moral truth (SA2). Therefore (SA5) we have reason to challenge the epistemic standing of our moral intuitions. As Sinnott-Armstrong puts it, "Moral beliefs that vary in response to factors that do not affect truth – such as wording and belief context – cannot reliably track the truth (54)."

A large number of recent papers have considered empirically-driven challenges to the epistemic credentials of moral intuition.[10] These authors are *not* directly addressing the topic of this paper; they are instead making a theoretical argument in moral epistemology. But we might think that they have also provided an explanation for doxastic embarrassment. If arguments of this sort are right, then psychological findings about moral judgment are disturbing because they trigger epistemic doubt. Perhaps this sort of argument is just a formalization of what runs through my mind when I experience doxastic embarrassment.

3.2 Inadequacy of the Epistemic Unreliability Explanation: The Truth-Tracking Assumption

[9] Sinnott-Armstrong's mention of inferential connection here is meant to narrow the target of his skeptical argument. He does not aim to debunk moral intuitions completely, but only to claim that they cannot be treated as free-standing sources of justification; they must be embedded in a broader coherentist framework. In effect, Sinnott-Armstrong is attacking Intuitionist views in moral epistemology (e.g. Audi 2008). He does mount a more generally skeptical challenge to moral intuition in Sinnott-Armstrong (2006).

[10] See Liao (2008), Berker (2009), Musschenga (2010), Kahane (2011), Mason (2011), Leben (2011), and Lazari-Radek and Singer (2012).

As natural as this suggestion might be, I do not think it is right. The reason for this is quite simple: the challenge to epistemic reliability assumes that the point of moral judgment is to track moral truth, yet one needn't accept this view about moral judgment in order to experience doxastic embarrassment. If doxastic embarrassment is experienced even by people who doubt that the point of moral judgment is truth-tracking, then epistemic unreliability cannot fully explain doxastic embarrassment.

Let me explain the point more carefully. First, what is the truth-tracking assumption? The basic idea is that (1) moral judgments like 'stealing is wrong' or 'you should give more to charity' express propositions that can have truth values, and (2) that a person's capacity for moral judgment is reliable when it responds correctly to whatever properties are the moral truth-makers. Put more intuitively, we can say this is the view that some moral judgments are true and some are false, and the point of moral judgment is to get it right as often as possible.

There are several ways that philosophers have doubted the truth-tracking assumption. One way is to be a non-cognitivist. Non-cognitivists deny that moral judgments express truth-evaluable propositions at all. Instead, moral judgments are taken to express conative attitudes such as approval or disapproval, or simply emotional states triggered by morally-salient situations. If you don't believe that moral judgments are truth-evaluable, then you are unlikely to believe their point is to track moral truth.[11]

Another way to doubt the truth-tracking assumption is to be a moral skeptic. You might accept that moral judgments are truth-evaluable, but believe that *none* are true (or perhaps that none are justified). For instance, you might think that moral propositions presuppose the existence of mind-independent moral facts, but since there are no such facts, all moral propositions are false (Mackie 1977). There are other ways to be a moral skeptic, but the unifying idea is: moral judgments are truth-evaluable, but none are true (or justifiably known to be true).

So non-cognitivists and some skeptics doubt that the point of moral judgment is to track moral truth. If epistemic unreliability, with its truth-tracking assumption, is what explains doxastic embarrassment, then we should expect that these people will not experience doxastic embarrassment. Since they don't believe that the point of moral judgment is to track moral truth, evidence that their own moral

[11] There are some very sophisticated forms of non-cognitivism that seek to preserve truth-related language, even the standard operations of moral epistemology, while denying that moral judgments are semantically truth-evaluable. See for instance Blackburn (1996).

judgment fails to track moral truth won't come as much of a surprise to them. If they are strongly motivated to confirm their metaethical views, perhaps they will even be happy.

The problem is that some of these people *do* experience doxastic embarrassment. We saw examples in the previous section. Stevenson is most well-known precisely for his contributions to non-cognitivism. Nietzsche's metaethics are harder to interpret, but it is not easy to read him as affirming a truth-tracking account of moral judgment. These are two famous instances, and I have met others. You probably have as well – ask around among your non-cognitivist and skeptical friends.

What does all this show? It shows that the explanation from epistemic unreliability is at best an *incomplete* explanation. Maybe the epistemic unreliability explanation does account for doxastic embarrassment in a particular group of people – those who accept a truth-tracking account of moral judgment. But this account leaves out others, and if we want to fully explain doxastic embarrassment, we will need to appeal to something else.

Some readers may now lose interest. They may be thinking: 'well, I *do* accept the truth-tracking assumption. What do I care about explaining the phenomenology of weirdos who don't agree with my metaethics?' If you are such a reader, then you should still keep reading. Maybe epistemic unreliability is *part* of what explains your doxastic embarrassment. But maybe it isn't the whole story.[12]

4. The Explanation from Automaticity

4.1 The automaticity of moral life

Another psychological challenge to moral judgment has emerged in recent years. It originates in experimental work by Jonathan Haidt (2001) showing that people are typically unaware of the reasons for their moral judgments. If pressed to justify a judgment, they are unable to do so or will confabulate plausible-sounding reasons that are not actually explanatory. The suggestion is that moral judgment, like many other behaviors studied in social psychology, is *automatic* - executed without conscious control and (at least sometimes) inaccessible to introspective examination (Bargh and Chartrand 1999).

[12] Another reason is that there may be logical problems with psychological debunking arguments. In this paper I have left their internal logic unchallenged, but elsewhere I point out a problem (Rini 2016). If, for whatever reason, you doubt that this sort of argument works, and yet you experience doxastic embarrassment, then that seems sufficient reason to keep reading.

A number of philosophers have noted that Haidt's results (taken at face value) cause trouble for standard conceptions of moral judgment. Jeanette Kennett and Cordelia Fine (2009) argue that, if Haidt is correct, then people may not be making moral judgments at all. Borrowing a distinction from Jones (2003) they contrast *reason-tracking*, in which an individual's actions are fitted to the reasons for action present in her environment, with *reason-responsiveness*, in which an individual is consciously aware of these reasons as reasons, and acts accordingly. To be a moral agent, they claim, is to make decisions in a reasons-responsive way – that is, to reflect upon our reasons and govern our decisions in accord with them. A decision made in a purely reason-tracking way, absent any sort of reflective assessment, is not a *moral* decision (or perhaps a decision at all). Hence the worry:

> An implication of Haidt's work is that our first personal experience of ourselves as reason responders is illusory… Haidt's data might mean that our moral concepts require drastic revision and that we need to take a much more modest deflationary view of our agency. (Kennett and Fine 2009, 85)

John Doris (2009) makes a related claim about the consequences of automaticity research. If very many of our decisions are automatic, then we rarely exercise reflective self-direction. Perhaps, Doris hints, even when it introspectively *seems* to us as if we are acting on reflection, this is actually confabulation. If that is true, and if agency requires reflective self-direction, then we may not be agents at all.

Interestingly, these authors turn away from the most skeptical possible conclusions of their arguments. Kennett and Fine argue that Haidt's research turns out not to foreclose some important roles for reasons-responsiveness in moral thought (Kennett and Fine 2009, 88-93). Doris breaks in another direction, arguing that we should revise our conception of agency to reduce the requirement for *reflective* self-direction (Doris 2009, 79).

I won't venture an opinion here on what we should ultimately make of automaticity research. I am interested in asking whether it can explain doxastic embarrassment. Perhaps it can. Perhaps what bothers me about learning the psychological origins of my moral beliefs is that this exposes their automaticity. All along I've imagined that I generate my moral beliefs in a reflective way - that I have *reasoned* from plausible starting points to tenable conclusions. But now I learn that some of my moral beliefs actually come from this automated psychological process, not from my conscious reasoning. Upon learning this, I am disturbed.

Notice that here, by contrast to the explanation from epistemic unreliability, the claim is not that my moral judgments are mistaken (e.g. by failing to correspond to moral truth). Rather, the claim is that I have generated my judgments in the wrong way, automatically rather than through conscious deliberation. Following Jones' terminology, I might get the moral reasons *right* by being a mere reasons-tracker, but I need something more to be a reasons-responder. The problem now is that I lack the something more – that my moral beliefs come from some automated bit of my brain, and not from conscious *me*. Is this discovery what animates my experience of doxastic embarrassment?

4.2 The inadequacy of the automaticity explanation: automaticity is familiar

The problem for this as an explanation for doxastic embarrassment is that it assumes we are *unfamiliar* with moral automaticity. If I really have thought up until now that my moral beliefs come from my reflective deliberation, then of course the findings of automaticity research will come as a shock. But is that really how I have been thinking about myself?

Perhaps not. Our perfectly ordinary conception of moral agency already contains elements of automaticity. Think here about Aristotelian habituation. The virtuous agent, per Aristotle, is one who has inculcated appropriate responses to situations and who does not need to detachedly reflect in order to see what is called for. Hanno Sauer (2012) has recently argued that our moral intuitions are automated because they are "educated". Even if we do not typically *consciously* engage with reasons, it remains true that "genuine moral reasons are effective in how people arrive at their verdicts: they figure effectively in the acquisition, formation and maintenance (that is, the education) of subjects' moral intuitions, and make a psychologically real difference to people's moral beliefs" (Sauer 2012, 263). Sauer's view is largely an updated form of Aristotle's, taking account of contemporary cognitive science.[13]

What this suggests is that we have a background conception of moral agency, traceable at least as far back as Aristotle, that does not insist upon a role for reflective assessment in the production of *particular* moral beliefs and actions. Think here of Bernard Williams' famous claim that, at least some times, reflecting upon a decision can demonstrate a flaw in moral character – reflecting on whether to rescue one's partner rather than a stranger involves "one thought too many" (Williams 1981, 18). It

[13] Sauer builds on McDowell (1994) and Pollard (2005). Kennett and Fine (2009, 93) make a similar point in their challenge to Haidt. See also Railton (2014) for related discussion of the rationality of uncontrolled moral cognition and see Velleman (2008) on the concept of 'flow'.

seems to be a perfectly ordinary part of our lives, and how we experience ourselves, that some of our moral choices are automatic.

To make the point vivid, imagine the following: You are crossing a busy street slightly ahead of me. A car suddenly comes rushing toward you. Instantly, without time to reflect, I lunge forward and push you out of the car's path. The car only very narrowly misses me.

Here it seems like I have acted automatically; I did not have time to think or reflect on my action. But when the moment is passed, I am not disturbed by my having acted automatically. Instead, I may be proud of what I have done. You and others might praise me for being brave, for putting others first, etc. All of this is compatible with everyone, including me, knowing that I was not deliberating about my moral reasons in the moment.

Cases like these suggest that automaticity *alone* cannot explain doxastic embarrassment. If my background conception of moral agency allows (at least some) automated production of moral beliefs and actions, then it is unlikely that my unease is explained by the discovery that some of my moral beliefs are generated automatically.[14]

5. The Explanation from Disunity

5.1 Lessons from previous explanatory attempts

My discussion of the explanations from epistemic unreliability and from automaticity served in part as ground-clearing; both are natural enough explanations, and it is useful to distinguish any other explanation from them. And I think that they can help us see the features a fully adequate explanation must account for.

The epistemic unreliability explanation was inadequate because it could not account for people who doubt the truth-tracking assumption. But it did call attention to this more basic idea: it is disturbing when psychological evidence suggests a moral judgment might be caused or influenced by the *wrong*

[14] What would be shocking is the discovery that reflection *never* plays a role in generating, revising, or sustaining moral beliefs. But not even Haidt claims this – his model allows a role for explicit moral reasoning, albeit "hypothesized to occur somewhat rarely outside of highly specialized subcultures such as that of philosophy, which provides years of training in unnatural modes of thought" (Haidt and Bjorklund 2008, 193).

sort of thing (as with framing effects). An adequate explanation will need to give some account of this sense of 'wrongness' that doesn't rely on the truth-tracking assumption.

The automaticity explanation was inadequate because it ignored our regular acceptance of automaticity in moral judgment. But it did call attention to this point: it is disturbing to learn that we sometimes *confabulate* reasons for our moral beliefs. Even if automaticity is accepted, we don't want to be inventing false stories about the reasoning behind our beliefs. So an adequate explanation for doxastic embarrassment will take account of confabulation.

I'll now combine these two observations to motivate what I believe is the best explanation for doxastic embarrassment. Doxastic embarrassment results from my awareness of a *gap* between the considerations that seem correct to me in my conscious thought, and the factors that actually drive my automated moral beliefs. What psychological research exposes is a form of *disunity* in my functioning as a moral agent. My conscious, reflective self is not appropriately unified with my automated, effective self. This, I will argue, is the core of doxastic embarrassment.

5.2 Agency as autonomy and efficacy

To get at this idea, we will need a clear sense of 'agency'. Christine Korsgaard offers one:

> The ideal of agency is the ideal of inserting yourself into the causal order, in such a way as to make a genuine difference in the world. Autonomy, in Kant's sense of not being determined by an alien cause, and efficacy, in the sense of making a difference in the world that is genuinely your own, are just the two faces of that ideal, one looking behind, and the other looking forward. (Korsgaard 2009: 89-90)

Korsgaard's account highlights two distinctive components of agential action: a motivating principle that belongs to the agent, and an effective execution of that principle in the world. Bodily movements that originate in alien causes (the 'knee jerk' response to the physician's mallet) are not autonomous and so not agential. Yet the mere willing of an autonomous principle, if persistently prevented from having anything like its intended effect in the world, is also not sufficient for agency.

It will be useful to think about the relationship between agency and automaticity. It might seem that research on automaticity is a problem for agency, since automated actions are uncontrolled by the conscious subject. But this is not quite correct.

Recall the case described at the end of section four, in which I act automatically to save you from a speeding car. Does the fact that I didn't have time to reflect make it true that I did not act agentially? This doesn't seem right. My bodily movement was not involuntary, nor the result of an alien cause. Though I did not reflectively endorse my action in the moment, it is nevertheless true (let us suppose) that I *would have done so* if I'd had the time. My action was caused by a dispositional commitment that I can and do reflectively endorse.

Now contrast this with the following case: As before, we are crossing the street when a car approaches quickly. This time, however, we have just come from the doctor, where I have been treated with an adrenaline shot. My reflexes are primed, and it is the sound of the car's engine that causes my muscles to jerk and my arms to thrust forward. More or less by accident, you are saved.

In this case, I think, it is clear that I have not acted agentially. Though my bodily movement accomplished an outcome I can endorse, it is not really *me* that acted. The adrenaline shot and the car's engine were the joint causes of my movement, and my commitments seem to have played no role.[15]

What these two cases suggest is that automaticity is not by itself incompatible with agency. In the first case I act automatically but still agentially, because my action results from my commitments. In the second case I do not act agentially, but it is not automaticity alone that explains this failure of agency. Automaticity is only problematic if the automatic process is one that lacks the right sort of connection with my commitments.

Notice also that our discussion parallels Aristotle's famous claim, in book 3 of the *Nicomachean Ethics*, that an action is involuntary when its "moving principle is outside, being a principle in which nothing is contributed by the person who is acting or is feeling the passion". In the first case, the moving principle of my action is within me, as a dispositional commitment to protecting you. But in the second case, the

[15] Here there is the complicated counterfactual matter of what I *would* have done if the adrenaline shot had not been present. Presumably (as per the first case) I would have automatically acted from my commitments and saved you – so my movement seems to be overdetermined, in a way that complicates analysis of moral responsibility (Frankfurt 1971). But I am trying to sidestep issues of responsibility here; the point of the case is just to clarify what is involved in agency. For discussion of consciousness and moral responsibility, see Sie (2009) and Levy (2014).

moving principle seems to be a purely physical combination of the engine sound and the adrenaline shot; I, qua agent, do not contribute anything to it. Both cases involve automaticity, but only the second is incompatible with agency.

This point is really only a reformulation of our earlier discussion of the automaticity explanation. But we are now ready to explore a *particular* way in which automaticity can be problematic – one that draws in lessons from the discussion of unreliability, and which I will argue is the crux of doxastic embarrassment. There are some cases in which the moving principle of an action does seem to be *within* an agent, in the sense that it shows a stable dispositional trait, but is not consistent with the agent's own conscious or reflective attitudes. These cases, I will say, are cases of agential disunity.

Consider the following case. Again, we are crossing the street and here comes the speeding car. Again I have no time to think. But this time I don't save you. I tense and lean back, and the oncoming car strikes you. Why didn't I save you? A psychotherapist could explain: though you and I have been friends for years, I secretly hate you. I am jealous of your professional success, and though I refuse to admit it to myself, I have always felt little jolts of joy when bad things happen to you.[16]

Have I acted agentially here? Now the question is quite difficult to answer, because I appear to harbor two rival moving principles. One moving principle, the one I consciously endorse, is inconsistent with my action. Consciously, I am horrified by my failure to act. It is counterfactually true (let's suppose) that if I'd had time to think, then I *would* have saved you.

But I also have this second moving principle, operating below conscious awareness, that aims at bringing harm to you. I am mostly unaware of this fact; I sincerely claim to be your friend and I would certainly not reflectively endorse my jealousy and hatred. When I am fully in control of myself, my resentment has very little effect on how I treat you. But sometimes, when I have little time to think, or when my mind is distracted, my secret hostility breaks through and finds expression in my actions.

So, again, have I acted agentially in this case? Unsatisfyingly, I do not think there is a determinate answer to the question. I've claimed that automatic action can be agential provided that it follows from a moving principle within the agent. What makes this case problematic is that my stepping back from saving you *does* follow from a moving principle– just not the principle that I do or would consciously

[16] This case parallels a regular source of interpersonal drama in fiction – the mistake that is maybe not entirely an accident. See, for example, John Knowles' *A Separate Peace*, Ian McEwan's *Atonement*, Margaret Atwood's *The Blind Assassin*, or Julian Barnes' *The Sense of an Ending*.

endorse. It is surely not true that my bodily movement resulted from alien causes, at least not in the same way as the adrenaline shot case.

It is very tempting here to insist that whichever moving principle I do or would reflectively endorse is my *real* moving principle, and to claim that secret hatreds lingering below conscious awareness are somehow outside myself, if not exactly alien causes. But I am not sure this is right; as Nomy Arpaly (2002) has argued, it is not always obvious that the things we explicitly endorse are the values rightly attributed to us.[17] I do not think the question admits of an easy answer, and I will not attempt one here.

What I will say is this: the case shows one way in which an agent can be *disunified*. When my consciously endorsed moving principle is not identical with the moving principle acting below consciousness, and when my nonconscious moving principles are sometimes able to express themselves in my automatic behavior, then my agency is not unified. Disunified agency, if it is agency at all, is not a good sort of agency. Note that, per Korsgaard's analysis, a disunified agent regularly fails to exhibit the requisite combination of autonomy and efficacy. When a person's consciously endorsed moving principles (which, let us suppose, are the locus of autonomy) are not the ones driving her automatic behavior, then autonomy and efficacy cannot both be achieved for the same moving principles.

It is a bad thing to be disunified. Possibly, when I am disunified, I am not an agent at all – I am a mere complex of psychological impulses, adorned with an inefficacious reflective capacity spinning happy confabulations. It is distressing to come into contact with this truth.[18]

5.3 Agential judgment and the standards of evaluation

[17] Further, there is empirical evidence that our attribution of values to another person is affected by our own moral commitments, suggesting that we do not have a non-normative way of making the distinction. See Newman, Bloom, and Knobe (2014) and Strohminger and Nichols (2014).

[18] Anyway, *I* think that it is bad to be agentially disunified. But I should admit that not everyone thinks this, especially not those spared a Kantian intellectual upbringing. Many Buddhists hold that conceiving of oneself as a single unified agent is not only mistaken but is *the root of suffering*. It may be that my analysis does not apply to people with radically different conceptions of the self and human agency. I would be interested to know whether people raised in this tradition experience doxastic embarrassment at all; if they do, then that is a problem for my theory. Thanks to Nic Bommarito for this point (and for the phrase 'Kantian intellectual upbringing').

This gets us very close to an explanation for doxastic embarrassment. But we will need one more piece of conceptual machinery, in order to extend the discussion of agential *action* to moral *judgment*, since most of the psychological research at issue concerns judgment rather than action.

Personally, I am inclined simply to say that evaluative judgment *is* a type of mental action.[19] But there will be disagreement on the relationship between judgment and action, and I don't wish to argue the point here. So we will need a slightly different explanation for doxastic embarrassment about moral judgment.

Even if judgments are not themselves mental actions, there is a tight connection between moral judgment and action. When I make a moral judgment, I am evaluating an actual or possible action and doing so by the same standards that I would wish to be efficacious in my own actions in similar situations.[20]

For example: suppose I am asked to judge whether it is better to save person A or person B in a moral dilemma, knowing only certain pieces of information about them. I regard some information as irrelevant to the choice: e.g. that A is on the right and B on the left. But other information does seem relevant to me: e.g. that A is a serial killer and B is an innocent child. The factors I treat as relevant in my judgment are the same factors I would wish to be efficacious if I ever were in these situations; I would wish that even in the heat of the moment, I'd remember to save the child rather than the killer, not getting distracted by which is on the right. Hence the connection between my judgment and a set of actions.

This is a fairly simple point, but stating it explicitly allows us to draw action and judgment together under Korsgaard's conception of agency. The idea is that I, as an agent, am constituted not only from my agential actions, but also from my moral judgments, since these express the moving principles I would wish to be efficacious in my actions. To put the point another way: it *matters to me* how and why I make evaluative judgments, just as it matters to me how and why I act.

Now, it might seem that judgments, unlike actions, cannot be afflicted by disunified agency. Since judgments are mental states, there is no risk of them failing to "make a difference in the world"; there

[19] Some philosophers of mind have argued that certain mental happenings (such as judgments) should be understood as *agential mental acts* (e.g. Geach 1957; Proust 2001).
[20] This is not too far off from Allan Gibbard's claim that when I *judge* what is to be done in a particular circumstance, I am *making a plan* for what I would do were I ever in that circumstance. See Gibbard (2003, 48-53).

seems to be no gap between autonomy and efficacy from which disunified agency could emerge. But in fact there is room for a gap. Suppose I am preparing to grade a large stack of student essays and I first reflect on the factors that I think are relevant to the grades I will assign. I worry that in the past I've given slightly higher grades to students who use philosophical buzzwords, even when their arguments are no better, and I reflectively decide that I should not grade in this way. But as I get into the grading, I become tired and I forget to keep my decision at the front of my mind. There is now a question about whether or not I will effectively implement my reflective decision: have I avoided letting my grading be influenced by the appearance of buzzwords, as I'd wished? That the question is possible shows there can be a gap between autonomy and efficacy *even in judgment*: I can autonomously intend that buzzwords not factor into my quality judgments, but this intention may be inefficacious when the judging takes place.

If there *is* such a gap, then I am agentially disunified with respect to these judgments. I suggest that this is what often happens in implicit bias.[21] Many people are egalitarians, explicitly disavowing the relevance of race or gender to the evaluation of an individual. When they make evaluative judgments about others, they sincerely believe that they do so by employing standards that have nothing to do with race or gender. Yet it can be empirically demonstrated that, even for many explicit egalitarians, actual evaluative judgments are systematically affected by race and gender bias.

Egalitarians who harbor implicit biases are disunified evaluative judges. The evaluative standards that they accept reflectively are not the same as the factors that drive the judgments they actually make. They do not reflectively endorse a form of evaluative judgment in which race and gender are treated as relevant. Yet when the time comes to actually make evaluative judgments, they unwittingly judge *as if* race and gender were relevant. Examples like these show how we can be agentially disunified in evaluative judgment as well as in action.

5.4 Doxastic embarrassment and agential disunity

[21] For an overview of the science of implicit bias see Jost et al. (2009). For philosophical discussion see the essays in Brownstein and Saul (forthcoming).

So, to draw this all together: when I am surprised by revelations of the causal origins of my moral judgments, and when this discovery causes my trust in these judgments to waver, why has this happened? Because I have been made aware of disunity in my agential judgments.

To return to the example with which I began: Joshua Greene's research appears to show that when I morally distinguish between the Trolley and Footbridge cases, I am not doing so because of my reflective commitment to any deontological principle. Rather, I am doing so because primitive bits of my mind have been evolved to evaluate up-close-and-personal violence in one way, and have not evolved to evaluate impersonal violence in the same way. My reflective self is disunified from how I actually make judgments. The reflective preferences I have about how I ought to judge are inefficacious.

Notice that the claim here is *not* necessarily that my reflectively-endorsed standards of evaluation lead to different *verdicts* than the psychologically efficacious standards of evaluation. Both the Doctrine of Double Effect (or something like it) and Greene's hypothesized up-close-and-personal mechanism distinguish between the Trolley and Footbridge cases.[22] The problem is about *how* I got to my verdict, not necessarily what the verdict says.

This helps us address a matter of contention in the literature on moral psychology. One might ask: if the psychologically efficacious causes of my judgment lead me to the same verdicts as the standards I reflectively endorse, why should I be bothered? After all, it is *not* that I am producing judgmental outcomes different from those I can reflectively accept (as is the case in implicit bias). Van Roojen (1999) and Kamm (2009) have made exactly this point about moral psychological research, citing it as a reason to pay little attention to the results of these experiments. So long as my judgments can be given sufficient justification within the terms of a normative theory, why does it matter how they came about causally?

The disunity explanation provides an answer. Recall the earlier speeding car case, in which I accidentally save you because an adrenaline shot makes my arm jerk forward uncontrollably. I can be glad of this outcome, thankful that it happened, and yet nevertheless find the involuntary sensation alienating or disturbing.

[22] Indeed, that is Greene's point: he says that the up-close-and-personal mechanism *explains* my deontological moral intuitions, and that deontological moral philosophy is a *rationalization* of my primate psychology (Greene 2008, 68).

This is what I claim happens in doxastic embarrassment. It is the *discovery* that my judgment results from a standard of evaluation other than the one I reflectively endorse that causes me to feel discomfort and sometimes even to doubt my moral beliefs. Doxastic embarrassment is a glimpse of my own disunity, of the fact that the conscious 'me' is only one among a collection of factors determining what I value, some of which do not agree with the conscious 'me'.

Now we can see how the other two explanations for doxastic embarrassment each get partly – but only partly – to the truth. Both of these explanations point to disunity, though they treat it as merely *diagnostic* of some other problem.

The explanation from epistemic unreliability claims that doxastic embarrassment results from the discovery that my moral beliefs do not track the moral truth. In order to show that the factors driving moral intuition (e.g. framing effects) are not truth-tracking, authors often say something like, "surely no one thinks that *this* is morally relevant". What they mean is that no one *reflectively* accepts the factor as an appropriate standard for evaluation. This is why even those who don't accept a truth-tracking account of moral judgment can still feel doxastic embarrassment: they can reflectively agree that they do not wish to make moral judgments in such-and-such a way, whether or not there is any moral truth to track. So the unreliability account treats agential disunity as a means of diagnosing the problem, when it is in fact a problem itself.

Similarly, the explanation from automaticity points to worries about agential disunity. To motivate the claim that many choices are "uncontrolled", automaticity theorists highlight cases in which choices seem to turn on factors that the conscious subject is unaware of or even explicitly disavows – e.g. (in Doris's (2009) most prominent example) walking more slowly after being exposed to elderly prime words (Bargh, Chen, and Burrows 1996). In these cases, disunity between the reflective self and efficacious behavior is used to diagnose automaticity. But these are only *some* forms of automaticity, and automaticity itself does not seem to bother us; we are okay when we automatically enact standards that we *can* reflectively endorse (as in the original, simple case of my saving you from the onrushing car). Automaticity only becomes disturbing when it is *also* disunity.

So agential disunity is common to both explanations from epistemic unreliability and from automaticity, and it does not have their limitations. Our conception of unified moral agency seems to be at the root of why learning psychological origins can lead to embarrassment at our moral beliefs.

6. An inconclusive conclusion

The central aim of this paper is complete. I have argued that the explanation for doxastic embarrassment is the recognition of agential disunity. Learning the psychological origins of our moral judgments can expose a gap between our conscious selves and the nonconscious parts of us that are efficacious in our judgments and actions. Encountering this gap is often disturbing.

All I have done here is explain doxastic embarrassment. I have not addressed its rationality. Perhaps, as Dworkin and Nagel would have it, doxastic embarrassment is an irrational sort of response, and we would be foolish to change our moral beliefs on its account. Or perhaps, as Nietszche and Stevenson would have it, doxastic embarrassment is a useful state, preparing the way for wise reappraisal of our values.

I can't settle the matter here. But I will conclude with some questions, inspired by the points above, which can help to guide the next step.

First: is there something special about agential disunity *in the moral domain*? It seems as if this may be the case. Many of our non-moral judgments and actions are driven by nonconscious processes whose workings would surprise us. Think about perception: read a description of the two-streams account of visual perception, or experiments involving blindsight, and you will quickly appreciate that your visual system does not operate at all as it might seem to you subjectively. But learning how ordinary vision works rarely leads us to be alienated from our perceptual states, or to hesitate from the judgments they support. There does seem to be something especially disturbing about agential disunity in the moral domain. What is it?

Second: if doxastic embarrassment *is* sometimes rational – that is, if being disturbed to discover our agential disunity ever rightly disposes us to change our moral judgments – then what sort of changes would be appropriate? Many philosophers will think it obvious that we should close the gap between autonomy and efficacy by changing our nonconscious psychology (e.g. through cognitive-behavioral therapy) so that our actions and judgements are sensitive to the same factors we reflectively endorse. That seems right for implicit bias – but is it always right? Cases like that of Huckleberry Finn, who was driven to do the right thing while reflectively endorsing otherwise, lead some philosophers to say there is nothing especially valuable about reflective endorsement. (e.g. Arpaly 2002). Could it ever be true that we should close the gap by altering our reflective endorsement to match the standards already

efficacious in our actions and judgments, rather than the other way around? How could we determine which cases are which?

These are intriguing questions, the answers to which would imply quite a bit about the rationality of doxastic embarrassment. The account given in this paper has allowed us to frame the questions, and I believe that it can also guide our answers. But I leave answers for another time.[23]

References

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

Arpaly, Nomy. 2002. *Unprincipled Virtue*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/0195152042.001.0001/acprof-9780195152043.

Audi, Robert. 2008. "Intuition, Inference, and Rational Disagreement in Ethics." *Ethical Theory and Moral Practice* 11 (5): 475–92.

Bargh, J.A., and T.L. Chartrand. 1999. "The Unbearable Automaticity of Being." *American Psychologist* 54: 462–79.

Bargh, John A., Mark Chen, and Lara Burrows. 1996. "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action." *Journal of Personality and Social Psychology* 71 (2): 230–44. doi:10.1037/0022-3514.71.2.230.

Berker, Selim. 2009. "The Normative Insignificance of Neuroscience." *Philosophy & Public Affairs* 37 (4): 293–329. doi:10.1111/j.1088-4963.2009.01164.x.

Blackburn, Simon. 1996. "Securing the Nots: Moral Epistemology for the Quasi-Realist." In *Moral Knowledge? New Readings in Moral Epistemology*, edited by Walter Sinnott-Armstrong Mark Timmons, 82–100. Oxford University Press.

Brownstein, Michael, and Jennifer Saul, eds. forthcoming. *Implicit Bias and Philosophy*. Oxford: Oxford University Press.

Caruso, Eugene M., and Francesca Gino. 2011. "Blind Ethics: Closing One's Eyes Polarizes Moral Judgments and Discourages Dishonest Behavior." *Cognition* 118 (2): 280–85.

Chappell, Timothy. 2013. "Why Ethics Is Hard." *Journal of Moral Philosophy* Advance Online.

Doris, John M. 2009. "Skepticism about Persons." *Philosophical Issues* 19 (1): 57–91.

Doris, John M., and Stephen Stich. 2007. "As a Matter of Fact: Empirical Perspectives on Ethics." In *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith, 1st ed., 1:114–53. Oxford: Oxford University Press. http://www.oxfordhandbooks.com/oso/public/content/oho_philosophy/9780199234769/oxfordhb-9780199234769-chapter-5.html.

[23] This paper has extensively benefited from discussion by conference audiences at Oxford and NYU, especially a set of superb comments by Nic Bommarito. It was also greatly improved by participants in the 2015 Mentoring Workshop for Women in Philosophy at the University of Massachusetts Amherst, including Dana Howard, Julia Nefsky, Tina Rulli, and most especially Amelia Hicks and Karen Stohr. I also owe thanks to Nomy Arpaly, Nora Heinzelmann, Guy Kahane, David Kaspar, Hanno Sauer, Amia Srinivasan, and an anonymous reviewer for *Philosophical Studies* for very helpful comments and discussion.

Dworkin, Ronald. 1996. “Objectivity and Truth: You’d Better Believe It.” *Philosophy & Public Affairs* 25 (2): 87–139.

Foot, Philippa. 1967. “The Problem of Abortion and the Doctrine of Double Effect.” *Oxford Review* 5: 5–15.

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

Geach, Peter. 1957. *Mental Acts: Their Content and Their Objects*. New York: The Humanities Press.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Greene, Joshua D. 2008. “The Secret Joke of Kant’s Soul.” In *Moral Psychology, Vol. 3. The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, edited by Walter Sinnott-Armstrong, 35–80. Cambridge, MA: MIT Press.

———. 2014. “Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics.” *Ethics* 124 (4): 695–726. doi:10.1086/675875.

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. “The Neural Bases of Cognitive Conflict and Control in Moral Judgment.” *Neuron* 44 (2): 389–400. doi:10.1016/j.neuron.2004.09.027.

Haidt, Jonathan. 2001. “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review* 108 (4): 814–34.

Haidt, Jonathan, and Fredrik Bjorklund. 2008. “Social Intuitions Answer Six Questions About Moral Psychology.” In *Moral Psychology, Vol 2. The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-Armstrong, 181–218. Cambridge, MA: MIT Press.

Jones, Karen. 2003. “Emotion, Weakness of Will, and the Normative Conception of Agency.” In *Philosophy and the Emotions*, edited by A. Hatzimoysis, 181–200. Cambridge University Press.

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin. 2009. “The Existence of Implicit Bias Is beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies That No Manager Should Ignore.” *Research in Organizational Behavior* 29: 39–69. doi:10.1016/j.riob.2009.10.001.

Kahane, Guy, Jim A. C. Everett, Brian D. Earp, Miguel Farias, and Julian Savulescu. 2015. “‘Utilitarian’ Judgments in Sacrificial Moral Dilemmas Do Not Reflect Impartial Concern for the Greater Good.” *Cognition* 134 (January): 193–209. doi:10.1016/j.cognition.2014.10.005.

Kahane, Guy, and Nicholas Shackel. 2010. “Methodological Issues in the Neuroscience of Moral Judgement.” *Mind & Language* 25 (5): 561–82. doi:10.1111/j.1468-0017.2010.01401.x.

Kamm, F. M. 2009. “Neuroscience and Moral Reasoning: A Note on Recent Research.” *Philosophy & Public Affairs* 37 (4): 330–45. doi:10.1111/j.1088-4963.2009.01165.x.

Kauppinen, Antti. 2007. “The Rise and Fall of Experimental Philosophy.” *Philosophical Explorations* 10 (2): 95–118.

Kennett, Jeanette, and Cordelia Fine. 2009. “Will the Real Moral Judgment Please Stand up? The Implications of Social Intuitionist Models of Cognition for Meta-Ethics and Moral Psychology.” *Ethical Theory and Moral Practice* 12 (1): 77–96.

Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press.

Lazari-Radek, Katarzyna de, and Peter Singer. 2012. “The Objectivity of Ethics and the Unity of Practical Reason.” *Ethics* 123 (1): 9–31. doi:10.1086/667837.

Leben, Derek. 2011. “Cognitive Neuroscience and Moral Decision-Making: Guide or Set Aside?” *Neuroethics* 4 (2): 163–74. doi:10.1007/s12152-010-9087-z.

Levy, Neil. 2014. *Consciousness and Moral Responsibility*. Also available as: eBook.

Liao, S. 2008. “A Defense of Intuitions.” *Philosophical Studies* 140 (2): 247–62. doi:10.1007/s11098-007-9140-x.

Ma-Kellams, Christine, and Jim Blascovich. 2013. “Does ‘Science’ Make You Moral? The Effects of Priming Science on Moral Judgments and Behavior.” *PLoS ONE* 8 (3): e57989. doi:10.1371/journal.pone.0057989.

McDowell, John. 1994. *Mind and World*. 3. Cambridge: Harvard University Press.

Musschenga, Albert W. 2010. “The Epistemic Value of Intuitive Moral Judgements.” *Philosophical Explorations* 13 (2): 113–28. doi:10.1080/13869791003764047.

Nagel, Thomas. 1997. *The Last Word*. Oxford: Oxford University Press.

Newman, George E., Paul Bloom, and Joshua Knobe. 2014. “Value Judgments and the True Self.” *Personality and Social Psychology Bulletin* 40 (2): 203–16. doi:10.1177/0146167213508791.

Nietzsche, Friedrich. 1973. *The Will to Power: In Science, Nature, Society and Art*. Edited by Walter Kaufmann. New Ed. Random House USA Inc.

Pollard, Bill. 2005. “Naturalizing the Space of Reasons.” *International Journal of Philosophical Studies* 13 (1): 69–82. doi:10.1080/0967255042000324344.

Proust, Joëlle. 2001. “A Plea for Mental Acts.” *Synthese* 129 (1): 105–28.

Railton, Peter. 2014. “The Affective Dog and Its Rational Tale: Intuition and Attunement.” *Ethics* 124 (4): 813–59. doi:10.1086/675876.

Rini, Regina A. 2013. “Making Psychology Normatively Significant.” *The Journal of Ethics* 17 (3): 257–74. doi:10.1007/s10892-013-9145-y.

———. 2016. “Debunking Debunking: A Regress Challenge for Psychological Threats to Moral Judgment.” *Philosophical Studies* 173 (3): 675–97.

Sauer, Hanno. 2012. “Educated Intuitions. Automaticity and Rationality in Moral Judgement.” *Philosophical Explorations* 15 (3): 255–75.

Sie, Maureen. 2009. “Moral Agency, Conscious Control, and Deliberative Awareness.” *Inquiry* 52 (5): 516–31.

Singer, Peter. 2005. “Ethics and Intuitions.” *Journal of Ethics* 9 (3-4): 331–52.

Sinnott-Armstrong, Walter. 2006. “Moral Intuitionism Meets Empirical Psychology.” In *Metaethics After Moore*, edited by Terry Horgan and Mark Timmons, 339–66. Oxford: Oxford University Press.

———. 2008. “Framing Moral Intuition.” In *Moral Psychology, Vol 2. The Cognitive Science of Morality: Intuition and Diversity*, 47–76. Cambridge, MA: MIT Press.

Stevenson, Charles L. 1944. *Ethics and Language*. New Haven, CT: Yale University Press.

Strohminger, Nina, Richard L Lewis, and David E Meyer. 2011. “Divergent Effects of Different Positive Emotions on Moral Judgment.” *Cognition* 119 (2): 295–300. doi:10.1016/j.cognition.2010.12.012.

Strohminger, Nina, and Shaun Nichols. 2014. “The Essential Moral Self.” *Cognition* 131 (1): 159–71. doi:10.1016/j.cognition.2013.12.005.

Thomson, Judith Jarvis. 1976. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59 (2): 204–17.

van Roojen, Mark. 1999. “Reflective Moral Equilibrium and Psychological Theory.” *Ethics* 109 (4): 846–57.

Velleman, J. David. 2008. “The Way of the Wanton.” In *Practical Identity and Narrative Agency*, edited by Kim Atkins and Catriona Mackenzie, 169–92. New York: Routledge.

White, Roger. 2010. “You Just Believe That Because...” *Philosophical Perspectives* 24 (1): 573–615. doi:10.1111/j.1520-8583.2010.00204.x.

Williams, Bernard. 1981. *Moral Luck*. Cambridge University Press.

Yang, Qing, Xiaochang Wu, Xinyue Zhou, Nicole L. Mead, Kathleen D. Vohs, and Roy F. Baumeister. 2013. “Diverging Effects of Clean versus Dirty Money on Attitudes, Values, and Interpersonal Behavior.” *Journal of Personality and Social Psychology* 104 (3): 473–89. doi:10.1037/a0030596.