The Normative Significance of Cognitive Science Reconsidered

Dustin Locke
Claremont McKenna College

Forthcoming in *The Philosophical Quarterly*
Draft of 8/22/19. Please cite only with permission.

Abstract

Josh Greene (2007) famously argued that his cognitive-scientific results undermine deontological moral theorizing. Greene is wrong about this: at best, his research has revealed that at least some characteristically deontological moral judgments are *sensitive to* factors that we deem morally irrelevant. This alone is not enough to undermine those judgments. However, cognitive science could someday tell us more: it could tell us that in forming those judgments, we treat certain factors *as reasons* to believe as we do. If we independently deem such factors to be morally irrelevant, such a result would undermine those judgments and any moral theorizing built upon them. This paper brings charity, clarity, and epistemological sophistication to debates surrounding empirical debunking arguments in ethics.

Word Count

9,935

1. Introduction

Josh Greene (2007) famously attempts to empirically debunk deontological moral theorizing.[1] Philosophers have been highly critical of Greene's arguments. The most widely-cited critique by far is Selim Berker (2009).[2] Berker proceeds by canvasing several arguments Greene might be offering, refuting each along the way. Berker thus presents us with a challenge: since *these* arguments are no good, is there a better argument from cognitive-scientific results to Greene's moral-theoretic conclusion? Other critics present us with the same challenge.[3]

The present paper takes up this challenge. I spell out a plausible way in which cognitive-scientific results could undermine deontological moral theorizing. At the heart of my argument is the thesis that cognitive science could reveal that in forming our characteristically deontological judgments (CDJs), we treat a morally irrelevant factor as though it were morally relevant. Such a result would, I argue, undermine both those CDJs and any deontological principles accepted on their basis.

Fortunately for deontologists—myself included—cognitive science has not *yet* shown that in forming CDJs we treat morally irrelevant factors as though they were morally relevant. As I argue, cognitive science has at most shown that some of our CDJs are *counterfactually sensitive to* such factors. The upshot is that we need more research, and this paper identifies the specific question we need to address. The question is not simply 'What factors are our moral judgments sensitive to?' but 'What factors do we *treat as reasons* to make the judgments we make?'

I begin by briefly reviewing Greene's empirical work, noting some important criticisms (section 2). I show that, even if these criticisms are successful, Greene's research provides us with

[1] At last check, Google Scholar reports that Greene (2007) has been cited 610 times.
[2] Google Scholar reports that Berker's paper has been cited 142 times. The second most cited critique is Frances Kamm (2009), which currently has 38 citations, and which explicitly endorses Berker's main line of criticism. [Guy Kahane (2011), which discusses one aspect of Greene's argument, along with arguments from Sharon Street (2006) and Richard Joyce (2006), has been cited 97 times.]
[3] See inter alia Kamm (2009), Kahane (2011), and Lott (2016).

a first step in a serious challenge to deontological moral theorizing. I then explain how that threat is best understood (section 3), arguing that, contrary to Berker (2009), Greene's neuroscientific findings do not simply drop out of the argument (section 4). Finally, I discuss what remains to be shown in order to successfully debunk deontological moral theorizing (sections 5 and 6).

A few quick notes before I begin. First, nothing in this paper should be construed as offering an argument *against* any deontological moral principle. What is under discussion here is an argument that our endorsement of deontological principles is in a certain sense ill-founded. In other words, we are in search of an argument that 'undermines' deontological moral theorizing, leaving aside any thought of 'rebutting' it.[4] Second, and relatedly, I have tried to keep the body of this paper as free of epistemological jargon as possible. Readers interested in more explicit connections to the epistemology literature can find what they're looking for in footnotes.

Third, Greene's neuroscientific results—in particular, the results of his fMRI studies—end up playing a more limited role in the argument that I propose than what Greene originally intended.[5] While it can be difficult to pin Greene's argument down, it's clear that he sees the (alleged) fact that characteristically deontological judgments are driven by *emotional* centers in the brain as crucially important. The argument I propose will involve no such claim. In my view, what matters is *what we're treating as reasons* to believe as we do, whether or not this happens through an emotional process. That said, Greene's neuroscientific results do not drop out of the picture entirely. As I explain in section 4, the fMRI studies provide evidence that certain characteristically deontological judgments are not only formed through an emotional process, but through an *unconscious* process. If the process of belief formation is unconscious, then we cannot simply *introspect* to determine what it is we are treating as reasons to believe as we do—we must

[4] Sudduth (2015).
[5] Thanks to an anonymous referee for suggesting that I make this point explicit.

instead conduct careful empirical studies, often in the form of comparing subjects' responses across a range of cases. In so far as we are using such studies to discover the inner workings of our minds—in particular, as a means to address the question 'What are we, as a matter of fact, treating as reasons to believe as we do?'—such studies are within the domain of cognitive science generally (social psychology in particular). Hence, while cognitive science in general, and neuroscience in particular, play a different role in the argument as I conceive it than they do in Greene's original version of the argument, both remain significant.

Finally, fourth, this paper's focus on Greene's argument should not mislead the reader into thinking that there is nothing at stake other than one particular argument aimed at one particular moral theory. What's at stake is whether we can safely go about constructing moral theories without concern for the results of cognitive-scientific inquiry into the underpinnings of our moral judgments. If my argument is sound, we cannot. Commenting on an analogous line of argument, Russ Shafer-Landau once said that 'this leaves [the proponents of a certain moral theory] in a precarious position, hostage to the fortunes of future empirical research' (2012: 9). While I wish there were some in-principle argument that moral theorizing has nothing to fear from cognitive-science, I'll be arguing as forcefully as I can to the contrary.

2. The Central Tension Principle and Sensitivity to Personal Force

Various cognitive scientists have argued in defense of *dual-process theories.* What all such theories have in common is the view that psychological processes within some domain—say, decision-making—divide into two types: those that are relatively fast and intuitive (type-1), and those that are relatively slow and reflective (type-2).[6] The so-called 'dual-process theory of morality' is

[6] Evans and Stanovich (2013).

just the very minimal thesis that some moral judgments are the products of type-1 processes, while others are the products of type-2 process.[7]

Greene and his colleagues have gone beyond the bare-bones dual process theory of morality to make a specific proposal about which moral judgments are generated by which types of processes. Some important details of Greene's theory have changed over the years, but his main thesis is what he (2014) calls

> The Central Tension Principle. Characteristically deontological judgments are driven by automatic emotional responses, while characteristically consequentialist judgments are driven by controlled conscious reasoning.

Greene stipulates 'a characteristically consequentialist judgment' (CCJ) to mean a judgment that is 'naturally justified in consequentialist terms (i.e., by impartial cost-benefit reasoning) and that [is] more difficult to justify in deontological terms' (2014: 699). Greene's stock example of a CCJ is the judgment that it is permissible to push the stranger in Footbridge.[8] Similarly, Greene stipulates 'a characteristically deontological judgment' (CDJ) to mean a judgment that is 'naturally justified in deontological terms (in terms of rights, duties, etc.) and that [is] more difficult to justify in consequentialist terms' (*ibid*). Greene's stock example of a CDJ is the judgment that it is wrong to push the stranger in Footbridge.

Greene and his colleagues have conducted various studies in support of the dual-process theory of morality in general[9] and the Central Tension Principle in particular.[10] These studies are

[7] Campbell and Kumar (2012) call this a 'minimalist dual process theory' of moral judgment.
[8] Here and henceforth I assume that readers are familiar with the most famous trolley cases.
[9] Greene et al. (2001)
[10] Greene et al. (2004) and Greene et al. (2008).

somewhat controversial.[11] One of these controversies is particularly relevant here. Kahane et al. (2012) have argued that the correlation Greene and his colleagues have found between CDJs and type-1 processes, and CCJs and type-2 processes, is a mere artifact of the particular selection of moral dilemmas Greene and his colleagues used in their studies. Kahane et al. provide fMRI and reaction-time data suggesting that, with respect to a different set of moral dilemmas, the pattern found by Greene and his colleagues is reversed: CDJs are generated by type-1 processes and CCJs are generated by type-2 processes.[12] Greene and his colleagues dispute these findings.[13]

Fortunately, we need not settle this debate. Contrary to what Greene sometimes suggests, his attempt to undermine deontological theorizing does not actually rest on the Central Tension Principle, but on the weaker claim that 'many' CDJs are generated by automatic emotional processes.[14] More specifically, Greene's arguments rests on the claim that *many CDJs that have played an important role in ethical theorizing are produced by automatic emotional processes.* This claim is plausible in light of Greene's research. Greene has plausibly found, for example, that the judgment that it is wrong to push the stranger in Footbridge tends to be generated by an automatic emotional response. Greene found the same thing for the judgment that most people make about the famous Transplant case. Regardless of whether we can plausibly generalize to anything like the Central Tension Principle, it is an important finding that these CDJs—these that have played an important role in actual ethical theorizing[15]—are generated by automatic

[11] McGuire et al. (2009), Berker (2009), Kamm (2009), Mikhail (2011) and Kahane et al. (2012). See also Greene's reply to McGuire et al. (2009) in Greene et al. (2009b) and Greene's reply to Kahane et al. in Joseph M. Paxton et al. (2014).

[12] Kahane et al. hypothesize that 'intuitive' moral judgments are generated by automatic emotional processes, while 'counterintuitive' moral judgments are generated by conscious controlled reasoning, and that in the cases Greene studied, the intuitive/counterintuitive distinction just happens to line up with the CDJ/CCJ distinction.

[13] Paxton et al. (2014).

[14] Greene explicitly grants that 'there might be exceptions' to the Central Tension Principle (2014: 1368).

[15] Most notably in Philippa Foot (1967), Judith Thomson (1976), and those who have continued their lines of research.

emotional responses. This is a key first step in a serious challenge to deontological moral theorizing.

The next step is to ask what it is about the actions in questions that triggers the automatic emotional responses that generate the relevant CDJs. In (2001), Greene and his colleagues guessed that it was the 'personal harm' involved in cases like Footbridge. Later, Greene et al. (2009a) concluded that the automatic emotional responses were triggered by the interaction of two factors: *intention to harm* and *personal force*. The latter is a term of art, defined as follows.

> Personal Force. 'An agent applies personal force to another when the force that *directly* impacts the other is generated by the agent's muscles, as when one pushes another with one's hands or with a rigid object [as opposed to pressing a button that opens a trapdoor beneath someone].' (Greene et al. 2009: 365, emphasis in original)

With respect to personal force, Greene (2014) sums up the results of Greene et al. (2009a) as follows.

> We have automatic emotional responses that support [CDJs]. But what triggers these emotional responses?... It seems that there are two key factors that explain why most people say "yes" to the *switch* case and "no" to the *footbridge* case[: intention to harm and personal force.] The effect of personal force is revealed by examining four footbridge variations. In the *footbridge pole* version, the agent pushes the victim with a pole. In the [*trapdoor*] version, the agent drops the victim onto the tracks through a switch operated trap door, while standing near the victim on the footbridge. In the *remote* [*trapdoor*] version, the switch is located elsewhere, away from the victim. We asked

separate groups of subjects to judge whether the action proposed is morally acceptable. The proportions of subjects responding "yes" to these cases are as follows: [16]

[Cases with personal force.]

A. Standard footbridge: 31 percent yes

B. Footbridge pole: 33 percent yes

[Cases without personal force.]

C. [Trapdoor]: 59 percent yes

D. Remote [trapdoor]: 63 percent yes

...Such experiments [tell] us that people's judgments are, at least sometimes, sensitive to things like mere personal force. Ought people's moral judgments be sensitive to such things? We all answer, "no," of course...

And thus we've earned an interesting normative conclusion: People, at least sometimes, do not make good moral judgments in response to moral dilemmas. (Greene 2014: 708 – 713)[17]

Taken literally, Greene's 'interesting normative conclusion' is completely banal: 'people, at least sometimes, do not make good moral judgments in response to moral dilemmas' (2014: 713). Surely we don't need psychological research to tell us that.

[16] Subjects were asked to indicate yes/no in response to the question 'Is it morally acceptable to...' and to rate the acceptability on a scale of 1 ('completely unacceptable') to 9 ('completely acceptable'). Greene et al. reported mean acceptability ratings of 3.89 (n = 154, SD = .22) for Footbridge and 5.02 (n = 82, SD = .29) for Trapdoor.

[17] Tellingly, Greene begins by asking what it is that 'triggers' the automatic emotional processes that generate (certain) CDJs, but concludes with an assertion about what it is these CDJs are 'sensitive' to.

Given Greene's project, it is natural to assume that he means that people make a 'bad' moral judgment *when they make certain CDJs*—e.g., the CDJ that it is wrong to push in Footbridge. On this interpretation, Greene's argument goes something like this.

(1) The judgment that it is wrong to push in Footbridge is sensitive to personal force.

(2) We ought not to make judgments that are sensitive to personal force.

(3) Thus, we ought not to[18] have judged that it is wrong to push in Footbridge.[19]

Has Greene established (1)? At best, what Greene et al. (2009a) showed is something along the lines of

(1*) The judgment that it is wrong to push in Footbridge is counterfactually sensitive to personal force.

where a subject S's judgment that p is counterfactually sensitive to a factor F if and only if, had F not been present, S would not have judged that p. Strictly speaking, Greene et al. (2009a) did not establish (1*), but the weaker claim that

> *Most subjects who judge it is wrong to push in Footbridge are such that had they instead considered an action that was all else equal but did not involve personal force, they would*

[18] Here 'ought not to' means that our judgment was formed on the basis of bad reasons—not that there aren't good reasons, so to speak, 'out there', to judge as we judge. In other words, the conclusion is that our belief is not 'doxastically justified', although it may still be 'propositionally justified'. See Turri (2010).

[19] A more sophisticated version of the argument—a version that respects internalist views in epistemology—might put the idea in terms of the epistemic impact of *learning* that one's judgments are sensitive to personal force. In the interest of readability, I will forgo such niceties here.

have (at least) less strongly agreed with the claim that the action was wrong,[20] *and a significant number of these subjects would have thought that the action was not wrong.*

Nonetheless, for the sake of simplicity, let us grant Greene the stronger (1*).

Since we're interpreting (1) as (1*), (2) must then be interpreted as

(2*) We ought not make judgments that are counterfactually sensitive to personal force.

Recall that (2) was supposed to be Greene's 'non-controversial normative assumption'. (2*), however, is anything but. What is relatively non-controversial is a principle I will call

Irrelevance. One ought not treat personal force as a morally relevant factor—that is, one ought not treat the fact that an act involves personal force as opposed to impersonal force as in and of itself a reason to believe that the act is wrong.[21]

The idea behind Irrelevance is that if two acts are in all other respects morally equivalent, but one involves personal force and the other impersonal force (plausibly, Footbridge vs Trapdoor), then if the one involving personal force is wrong, so is the one involving impersonal force.

Crucially, Irrelevance does not imply (2*). Consider a non-moral analogy. Upon walking into a new colleague's office, I formed the judgment that there was a table in the corner of the room. This judgment was counterfactually sensitive to the fact that the lights in my colleague's office were on: had the lights not been on, I would not have come to believe that there was a table in the corner of the room. It's also true that *the lights being on* is not a 'table-ly' relevant

[20] See footnote 16 above.

[21] Thanks to an anonymous referee for encouraging a more careful formulation of Irrelevance.

difference: if two rooms A and B differ only in that the lights are on in A and not in B, then there is a table in the corner of A if and only if there is a table in the corner of B. It follows from this, I'll grant, that I ought not treat the light's being on as in and of itself a reason to believe there is a table in the room.[22] But it does not follow that I ought not to have made the judgment that there was a table in my colleague's office.

The lesson here is not simply the familiar lesson that some facts that are not constitutive of the truth of a proposition can nonetheless be evidence for that proposition.[23] The lesson here is that some judgments are counterfactually sensitive to facts that are neither constitutive of their truth, nor even evidence of their truth, and yet there is nothing wrong with those judgments. This happens, roughly speaking, when such facts play a role in *enabling us to detect the evidence in favor of those judgments*. The lights being on is not, I'll assume, evidence that there is a table in the room. But the lights being on enables me to detect the presence of the table. More specifically, the lights being on enables me to appropriately respond to a fact that I ought to treat as a reason to believe there is a table in the room—namely, the fact that there is a table in the room.[24] In general, from the assumptions that F should not be treated as relevant to the truth of your judgment J, and that J is counterfactually sensitive to F (and you know this), it does not follow that you ought not to make J. Hence, Irrelevance does not imply (2*).

Abandon (2*) and sticking with Irrelevance, we might try arguing as follows.

The Footbridge-Trapdoor Consistency Argument

(4) Personal force ought to be treated as morally irrelevant. (Irrelevance)

[22] I don't mean to suggest that this assumption is non-controversial. See Dancy (2004).
[23] Nor is it the less well-known but equally important lesson that sometimes a certain cognitive attitude is inappropriate even if it in a certain sense 'fits' its object. See D'Arms and Jacobson (2000).
[24] Contrary to certain Cartesian epistemologists, when I see a table in the room and form a belief that there is a table in the room, I thereby treat the fact that there is a table in the room as a reason to believe there is a table in the room. See my [omitted].

(5) There is no morally relevant difference between Footbridge and Trapdoor other than (at most) the presence/absence of personal force.

(6) Thus, we ought not make opposing moral judgments about Footbridge and Trapdoor.

If this is the argument, then it appears that empirical results have dropped out—the premises here are squarely normative.[25] More importantly, the conclusion of *this* argument is one that actual deontologists are happy to accept! Actual deontologists do not typically insist on a moral distinction between Footbridge and Trapdoor—they (typically)[26] insist on a moral distinction between Footbridge and Switch.[27] To challenge the position of these deontologists—that is, to undermine deontological moral theorizing as it is typically done—one needs to do more.

For roughly the reasons just given, Greene appears to have given up on the argument from Irrelevance,[28] and now offers what appears to be a new argument.[29] I will not pause to consider Greene's new argument here.[30] In my view, Greene has given up on the argument from Irrelevance too quickly. In the next few sections, I'll show that empirical inquiry could in principle tell us, not just that that our moral judgments are *sensitive* to personal force, but something more specific about *how* our judgments are sensitive to personal force. And should we learn that our judgments are sensitive to personal force in a particular way, this would undermine plenty of actual deontological moral theorizing.

[25] But cf. Campbell and Kumar (2012), who argued that Greene's empirical studies provide indirect support for (5).
[26] Cf. Thomson (2008).
[27] Aka, 'Bystander at the Switch'.
[28] See the last paragraph of section IV of Greene (2014).
[29] See the 'indirect route' section of Greene (2014).
[30] See Lott (2016) for useful discussion and criticism.

3. Treating Personal Force as a Reason

In the passage from Greene (2014) quoted above, Greene begins by asking what 'triggers' the psychological processes that produce our CDJs. As [omitted] and I argue elsewhere [omitted], wriggle-words like 'trigger'—and related words like 'track'—have wrought a bit of havoc in moral epistemology. Consider again my judgment that there is a table in the corner of the room. This judgment is counterfactually sensitive to the lights being on: had the lights not been on, I would not have made that judgment. But it is a bit misleading to say that the lights being on *triggered* my judgment. For epistemological purposes, one needs to draw distinctions between various elements of the causal chain leading up to a judgment, all of which might in some (loose) sense be considered part of what 'triggers' the judgment.

Chief among the distinctions these epistemically important distinctions is that between facts that one *treats as reasons* to believe as one does versus facts that *enable* one to treat those facts as reasons.[31] When I form my belief that there is a table in the room, I treat the fact that there is table in the room as a reason to believe there is table in the room. This also happens in cases where the fact treated as a reason is not the content of the belief formed: upon seeing a table in the room, I might treat the fact that there is a table in the room as a reason to believe that if I reach out, there will be a place to set down my coffee mug. But in neither case do I treat the lights being on as a reason to believe as I do—rather, the lights being on is (part of) what enables me to treat the relevant fact as a reason for believing as I do.

Importantly, the treating of facts as reasons is something that very often happens through unconscious automatic psychological processes. I need not consciously think: *there is a table in the room; therefore, if I reach out, there will be a place to set my coffee mug.* Rather, in virtue of automatically and unconsciously forming the latter belief, via the psychological process through

[31] Dancy (2000) objects to the idea that we treat *facts* as reasons. For criticism, see my [omitted].

which one normally forms such a belief in response to the presence of a table—whatever that processes might be—I thereby unconsciously and automatically treat the fact that there is a table in the room as a reason to believe that if I reach out, there will be a place to set down my coffee mug.

Now suppose Greene could show, not just that our CDJs are counterfactually sensitive to personal force, but that in forming those CDJs we (unconsciously and automatically) treat the fact that the action involves personal force as a reason to believe the action is wrong. Given that personal force ought to be treated as morally irrelevant, we would then have evidence that, in forming our CDJs, we treat something as a reason to form those CDJs that we ought not to treat as a reason to form those CDJs. Such a result would undermine those CDJs—that is, would make it so that, assuming we don't have any independent reason to endorse those CDJs, we ought to abandon them.[32] By contrast, if we learn that the presence of personal force is merely facilitating our detection of features of the action that *are* morally relevant—in the same way that the lights being on facilitates my detection of the table—then our CDJs would not thereby be undermined, and indeed might even be vindicated.

To fully appreciate this point, it is helpful to compare it with Campbell and Kumar's (2012) position. Campbell and Kumar argue that by learning that our judgments are sensitive to some factor that is irrelevant to the truth of those judgments, we learn that *either* we're getting things wrong with respect to cases where the factor is present *or* we're getting things wrong with respect to cases where the factor is absent—crucially, we do not thereby learn which of the cases we are mistaken about. I agree with Campbell and Kumar on this point. However, when we learn, not just that our judgments are sensitive to some irrelevant factor, but that in forming some of those judgments we *treat that factor as a reason to believe as we do*, that information gives us

[32] In other words, empirical results could in principle provide us with a certain kind of *defeater* for our CDJs. See Sudduth (2015).

reason to abandon those judgments, provided that we don't have any independent reason for continuing to hold them. [33]

But don't we have independent reason to continue to endorse the standard CDJs? A deontologist might grant that in initially forming our CDJs, we unconsciously treat something as a reason that we ought not to treat as a reason. Still, she might insist, there is good reason to believe as we do. She might note that pushing in Footbridge involves not only personal force (a bad reason to believe the action is wrong) but also, e.g., that pushing in Footbridge does not 'respect humanity as an end in itself'.

In response, the empirical debunker might insist that the deontologist has not truly offered an *independent* reason to believe the action is wrong. [34] According to what Allen Woods calls 'the standard' and 'dominant' model of ethical theorizing (2007: p. 43), any general moral beliefs the deontologist might appeal to in order to explain why it is wrong to push in Footbridge—e.g., her belief that humanity is an end in itself and her belief about what this means for cases like Footbridge—are ultimately justified at least in part by her intuitions about particular cases.

> [On the standard model, the aim of ethical theorizing] is to give the most coherent and intuitively compelling account of all our moral intuitions, at all levels of generality, an account that both reconciles our intuitive judgments and also gives us the most satisfying explanation of why we consider them true.
>
> This description fits a lot of what is done in moral philosophy at the present time... It fits the aims and procedures, for instance, of most philosophers who make use of carefully crafted if artificial examples in order to test and refine moral principles –

[33] cf. White (2010).
[34] Campbell and Kumar (2012) make a similar point.

> examples such as those in which you happen to be positioned so as to throw the switch and alter the course of a runaway trolley, which will kill one group of people if you don't throw the switch and another group of people if you do. (Wood 2007: p. 44)

Wood goes on to note that even many Kantian's have embraced this model of ethical theorizing.

> It might seem that the use of such examples would have a consequentialist bias and therefore be alien to Kantian theories. But the point of many trolley problems is to enlist our intuitions *against* the thesis that it is always right to produce the best overall consequences, by calling our attention to cases in which these consequences have been produced by means of actions we intuitively regard as wrong. Kantian ethics itself has been influenced by this model, especially in the interpretation of Kant's famous formulas of universal law (FUL) and the law of nature (FLN)... If one interpretation of Kant's formula yields counterintuitive results, then another interpretation is proposed. The fate of Kantian ethics itself, as a moral theory, is then seen as depending on this enterprise of interpretation, and how well our best interpretation of Kant's principle fares against our intuitions about the most challenging examples against which we can test it. (Wood 2007: pp. 44 – 45, emphasis in original)

If she embraces the standard model, the deontologists could hardly appeal to her more general moral beliefs—such as her belief that humanity is an end in itself, and her belief about what this means for cases like Footbridge—to justify her intuitive judgments about cases. On the standard model, things go the other way around: her intuitive judgments about cases are at least part of what justify her in having the more general moral beliefs.

That said, the standard model of ethical theorizing is by no means uncontroversial.[35] Wood himself goes on to insist that the standard model could hardly get 'Kant's conception of ethical theory more wrong if it tried' (p. 45). It would take us too far afield to explore the alternative model that Wood finds in Kant. Suffice it to say that one potential way for deontologists to avoid empirical debunking is to embrace some such alternative to the standard model of ethical theorizing. Indeed, should it turn out that the standard model, together with certain results from cognitive science, leads to an undermining of deontological theorizing, that would provide deontologists with all the more reason to embrace alternative models.

There is, however, another way one might try to escape the conclusion of the debunking argument. The results of Greene et al. (2009a) make it plausible that there is a minority of subjects who believe *both* that it is wrong to push in Footbridge and wrong to press the button in Trapdoor. These people have consistent views in line with standard deontological principles. Are *these people* threatened by the argument under consideration?

On the view defended here, the relevant issue is what, ultimately, such people are treating as reasons to believe as they do. There are lots of possibilities to consider, but there are two that are particularly salient. On what I'll call 'possibility (1)', such people don't ever treat personal force as a reason to believe an action is wrong. On what I'll call 'possibility (2)', such people (unconsciously) treat personal force as a reason to believe pushing in Footbridge is wrong and then, on the basis of consistency reasoning, also believe it is wrong to press the button in Trapdoor.

On possibility (1), subjects are treating *other* features of Footbridge—features not having to do with personal force—as reasons to believe it is wrong to push. Since these same features are present in Trapdoor, they also believe it is wrong to press the button in Trapdoor. If this is what's

[35] Thanks to an anonymous referee for encouraging me to discuss this and the following issue.

going on with you, and you *know* this is what's going on with you, then you are, as far as the particular debunking argument of this paper is concerned, in the clear.

On possibility (2), however, there's still trouble. On possibility (2), subjects are unconsciously treating personal force as a reason to believe it's wrong to push in Footbridge. Then, on the basis of this belief—but, crucially, without recognizing what they unconsciously treated as a reason to form this belief—they use *consistency reasoning* to arrive at the belief that it must also be wrong to press the button in Trapdoor. Schwitzgebel and Cushman (2012) have gathered empirical evidence suggesting that even professional philosophers are prone to forming beliefs on the basis of irrelevant factors and then, by consistency reasoning, drawing like conclusions about like cases. Of course, this consistency reasoning might be happening more or less unconsciously. Moreover, one needn't have previously encounter Footbridge specifically—one might have encountered other cases, perhaps some time ago, where one unconsciously treated personal force or some other irrelevant factor as a reason to believe as one does, and now, by consistency reasoning, arrived at the view that it must also be wrong to press the button in Trapdoor. If this is what's going on with you, and you *know* this is what's going on with you, then you're in epistemic trouble: when you look far enough back in the history of your beliefs, you find that you ultimately formed them on the basis of bad reasons.

Finally, what should you do if you, like me, don't *know* if you are of the possibility (1) sort or the possibility (2) sort? In other words: what should you do if you don't *know* whether you have formed your belief on the basis of a bad reason? This is a difficult question over which epistemologists disagree. But here's one plausible answer: you ought to be suspicious of your belief to the degree that you suspect that you formed that belief on the basis of a bad reason. I will return to this difficult issue of *higher-order* uncertainty in the final section. For now, let us move on.

So far, I've suggested that were Greene able to show that in forming the CDJ that it is wrong to push in Footbridge, we treat the fact that this action involves personal force as a reason to believe it is wrong, this would plausibly undermine that judgment. Now even if that were so, it could also be that in forming the characteristically consequentialist judgment (CCJ) that it is permissible to switch in Switch, we treat the fact that the action *does not* involve personal force as a reason to believe the action is *not* wrong.[36] Learning this would mean that the CCJ is undermined as well. However, it is possible that there's a cognitive asymmetry in how we respond to personal force. It could be that although we treat the presence of force as a reason to believe the action is wrong, we do not treat the *absence* of personal force as a reason to believe the action is *not* wrong.

This sort of cognitive asymmetry is familiar in other domains. I treat the fact that a label says 'Gluten Free' as a reason to believe that the product is gluten free. But suppose I don't see 'Gluten Free' on the label. Since many products that are, in fact, Gluten Free, don't contain such labeling, I usually don't treat the fact that the product's label does *not* say 'Gluten Free' as a reason to believe the product is not gluten free. Rather, I investigate further, looking most likely at the list of ingredients. Here there's a cognitive asymmetry when it comes to 'Gluten Free' labeling: when the label is present, I treat its presence as a reason to believe the product is gluten free, but when the label is absent, I don't treat its absence as a reason to believe the product is not gluten free. If I do end up believing that the product is not gluten free, I typically do so by treating other facts as reasons to believe as I do (e.g., the fact that the ingredients list says 'wheat').

It is possible that there is a similar cognitive asymmetry when it comes to personal force: when personal force is present, we (unconsciously) treat its presence as a reason to believe the

[36] As Thomson (2008) has in effect suggested.

action is wrong, but when personal force is absent, we don't treat its absence as a reason to believe the action is not wrong. It's possible that if we do end up believing that such an action is not wrong, we do so by treating *other* facts as reasons to believe as we do—e.g., the fact that more lives will be saved if the action is taken. Of course, there are other psychological possibilities to consider, each with their own epistemic implications.

The crucial question is which if any of our moral judgments are formed in a way that involves treating the presence/absence of personal force as a reason to believe as we do? Unfortunately, current research in cognitive science does not answer this question. But it could in principle. I explain how in section 5. First, I need to take care of a loose end.

4. The Neuroscientific Results Reconsidered

Now that we see that the crucial question is not 'What features are our moral judgments sensitive to?' but rather 'What features do we treat as morally relevant?', we can also see why, contra Berker (2009), Greene's neuroscientific results do not 'drop out of' the argument. By 'neuroscientific results', Berker means Greene's empirical conclusions about what type of psychological processes—automatic and emotional vs controlled and conscious—generate our moral judgments. These 'neuroscientific results' stand in contrast to Greene et al.'s (2009a) stimulus/reaction results about what features of a moral dilemma—e.g., the presence of personal force—our moral judgments are sensitive to.

Why does Berker think that Greene's neuroscientific results drop out of Greene's argument? After surveying and dismissing alternative reconstructions of Greene's argument, Berker reaches a reconstruction that is in some respects similar to the argument discussed in the previous section. Specifically, Berker considers the following (2009: 321).

The Argument from Morally Irrelevant Factors

(10) The emotional processing that gives rise to deontological intuitions responds[37] to factors that make a dilemma personal rather than impersonal.

(11) The factors that make a dilemma personal rather than impersonal are morally irrelevant.

(12) So, the emotional processing that gives rise to deontological intuitions responds to factors that are morally irrelevant.

(13) So, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force.

Berker says that his 'most pressing worry' about the argument from irrelevant factors is that

> the neuroscientific results seem to be doing no work in this argument. The epistemic [status] of consequentialist versus deontological intuitions now appears to be purely a function of *what sorts of features out there in the world [i.e., the presence/absence of personal force] they are each responding to...* The 'emotion-based' nature of deontological intuitions has no ultimate bearing on the argument's cogency. (2009: 325 – 6, emphasis in original)

There is, I think, something quite right about Berker's complaint. The argument, as Berker reconstructs it, says that what's wrong with our deontological intuitions is that they are formed via a psychological process that 'responds to' factors that are morally irrelevant. For the purposes of this argument, it would seem to not matter whether the psychological process in question was

[37] Here's another term to add alongside 'triggered by', which I discussed in the previous section.

automatic and emotional or controlled and conscious. Similarly, the argument that I've offered says that what's wrong with certain CDJs is that they are formed in a way that involves treating something that is not morally relevant as though it were morally relevant. As far as this argument goes, it ultimately does not matter whether those judgments are formed via automatic emotional processes or conscious controlled reasoning: it's possible, through either sort of process, to treat something that is not morally relevant as though it were morally relevant.

However, this does not mean that the neuroscientific results are completely irrelevant to the argument. More precisely, it does not mean that the neuroscientific results don't play some *evidentiary role* in the argument. As I've formulated it, the argument rests on the premise that in forming our CDJ about, e.g., Footbridge, we treat the presence of personal force as a reason to believe the action is wrong. Now suppose that we were forming our moral judgment about Footbridge entirely through a process of controlled conscious reasoning. In that case, it is at least somewhat plausible that, were we treating personal force as morally relevant, and were we to sufficiently reflect on what it was we were doing when we formed our judgment about Footbridge, we would have at least some, however minimal, introspective awareness that that is what we were doing. Since many of us have engaged in such reflection, and yet we are not, I'll assume, introspectively aware that we are treating personal force as morally relevant, it follows that if we do form our CDJ about Footbridge through controlled conscious reasoning, we are probably not treating personal force as morally relevant. Hence Greene's empirical result that we form our CDJ about Footbridge through an automatic emotional process is not idle: it makes the claim that we are treating personal force as morally relevant at least *defensible* in light of what would otherwise be a piece of evidence to the contrary—namely, that, try as we might, we have no introspective awareness that we are treating personal force as morally relevant.[38]

[38] This point is a bit trickier than I am letting on, for it is also possible to not be (higher-order) conscious of what one is consciously doing.

This is not to say that Greene's neuroscientific result provides *evidence for* the claim that in forming our CDJ about Footbridge, we treat personal force as though it were morally relevant. The point here that Greene's neuroscientific result pushes the question of whether we are treating personal force as morally relevant out of armchair psychology and into the hands of cognitive scientists: if the process via which we form our moral judgment about Footbridge is an automatic emotional process, the armchair philosopher's insistence that *he just knows* he isn't treating personal force as morally relevant when he thinks about Footbridge, becomes rather unconvincing. He might be right, but he'll need more than his powers of introspection to tell.

5. Where Things Stand Empirically

I have argued that in order to undermine our CDJs, the empirical evidence would have to support the hypothesis that in forming those CDJs, we treat the presence of personal force as a reason to believe the action is wrong. And to *selectively* undermine those CDJs, the empirical evidence would have to *fail* to support the hypothesis that, when forming the relevant CCJs, we treat the *absence* of personal force as a reason to believe the action is *not* wrong. Do we have such evidence? Not exactly.

We do have evidence that our CDJs are counterfactually sensitive to personal force, and this is at least *some* evidence that in forming those CDJs we treat the presence of personal force as a reason to believe the action is wrong. But this is not much evidence. After all, there are innumerable factors upon which people's CDJs counterfactually depend that no one would suggest people are treating as reasons to form those CDJs: level of intoxication[39] and general mood[40] are just two examples. Had Heather been drunk, she might not have judged that pushing was wrong in Footbridge. It doesn't follow that she treats *the fact that she is sober* as a reason to believe

[39] Duke and Begue (2015).
[40] Valdesolo and DeSteno (2006) and Strominger et al. (2011).

that pushing is wrong. Similarly, had Heather been shown a funny video clip prior to being presented with Footbridge, she might not have judged that pushing is wrong. Again, it does not follow that she treats *the fact that she didn't watch a funny video clip* as a reason to believe that pushing is wrong.[41]

Moreover, Greene et al. (2009a) found evidence that we do not *generally* treat the presence of personal force as a reason to believe that an action is wrong. That is, the presence of personal force does not in all cases lead subjects to judge an action to be less permissible than what they would have judged it to be had the action not involved personal force. Greene et al. found that *when an action involves intentional harm,* subjects rate the action as worse if it involves the use of personal force than if it does not—e.g., subjects on average rate the action in Footbridge worse than they rate the action in Trapdoor. However, when the act in question did not involve an intention to harm, the addition of personal force made no difference to how subjects on average rated the action. For example, subjects rated the action in Obstacle Collide no worse than they rated the action in Loop Weight.

> Obstacle Collide. An agent unintentionally collides with a stranger on a footbridge as he runs to a switch that he uses to divert a runaway trolley off of a main track and onto a sidetrack, thus saving five strangers on the main track, but the stranger with whom he collided on the footbridge is unintentionally knocked onto the sidetrack and killed by the trolley.

> Loop Weight. An agent flips a switch diverting a trolley onto a sidetrack where it runs over and kills one stranger but is then stopped by a heavy weight before the sidetrack

[41] This is related to the familiar distinction between so-called 'explanatory reasons' and 'motivating reasons'. See my [omitted] for discussion of the epistemic significance of this distinction.

connects back to the main track, further down which are five other strangers, who are thus saved.

Neither case involves intentional harm. However, Obstacle Collide involves the use of personal force to harm an agent, whereas Loop Weight does not. And yet Greene et al. (2009a) found no statistically significant difference in subjects' responses to these cases. More generally, Greene et al. found that 'personal force exhibited no effect in the absence of intention [to harm]' (369). They conclude that 'the effect of personal force depends entirely on intention [to harm]' (364). This conclusion has been supported by a recent meta-analysis of empirical research on the role of the means/side-effect distinction in moral cognition.[42]

Let us assume, then, that subjects do not *generally* treat the presence of personal force as a reason to believe an action is wrong. Still, it could be that when an action involves intentional harm, subjects treat the fact that the action involves personal force as a reason to believe the action is wrong. If so, subjects are treating something as morally relevant that they ought not to be treating as morally relevant. But there is another possibility: the presence of personal force *facilitates* treating intention to harm as a reason to believe an action is wrong, just as the lights being on *facilitates* treating the fact that there is a table in the room as a reason to believe there is a table in the room. Call this the 'Facilitation Hypothesis'.

How might the presence of personal force facilitate treating intentional harm as a reason to believe an action is wrong? Here is just one possibility—call it the 'Highlighter Hypothesis': the presence of personal force draws our attention to the action itself—as opposed to its consequences—allowing us to more fully appreciate the nature of that action—specifically, the fact that the action involves intentional harm. This can happen in at least two ways: either the

[42] [omitted]

presence of personal force gets us to *pay more attention* to the intentional harm than we otherwise would, or the presence of personal force gets us to *weigh more heavily* the intentional harm than we otherwise would.

Learning the truth of the Facilitation Hypothesis in general—or the Highlighter Hypothesis in particular—would not undermine the relevant CDJs. According to the Facilitation Hypothesis, what we're treating as morally relevant is *the intentional harm*—it is just that our treating intentional harm as morally relevant is at least to some extent facilitated by the presence of personal force. Thus, to assume that the Facilitation Hypothesis undermines our CDJs is to assume that we ought not treat intention to harm as morally relevant, or at least that we ought not give it the moral weight that we in fact give it in the presence of personal force. But such an assumption goes far beyond the uncontroversial assumption that personal force is morally irrelevant—indeed, such an assumption simply begs the question against the deontologists who thinks that intention to harm is morally relevant and is morally weighty enough to make, e.g., pushing in Footbridge wrong.

It's important to stress that the Facilitation Hypothesis is consistent with, and not an alternative to, the conclusion reached by Greene et al. (2009a). Greene (2014) says that the conclusion of Greene et al. (2009a) is that intention to harm and personal force 'interact' in the 'technical statistical sense, meaning that the influence of one factor is influenced by the presence/absence of another factor, as in an interaction between medications' (709). The *facilitation* hypothesis suggested here is one way in which this interaction might work.[43]

[43] Indeed, facilitation is a common way that medications interact. Neither penicillin nor streptomycin, when used alone, is effective against E. Coli. However, when used in combination, penicillin does just enough damage to the cell wall of E. Coli to allow streptomycin to penetrate the cell wall and kill it. Here it is the streptomycin, and not the penicillin, that kills the E. Coli. Nevertheless, the death of E. Coli is counterfactually sensitive to the presence of penicillin, because penicillin *facilitates* the effect of streptomycin.

It is worth emphasizing, however, that there is nothing in Greene et al. (2009a) that tells in favor of the Facilitation Hypothesis over some other version of Greene et al.'s conclusion. There are several ways in which personal force and intention to harm might interact to produce the results Greene and his colleagues found. The point here is that some of these ways—the ones that involve treating personal force as a reason—spell trouble for the relevant CDJs, while others—the ones that do not involve treating personal force as a reason—do not.

It should be possible to run experiments that shed light on whether subjects treat personal force as a reason to believe certain actions are wrong. Here's one sort of experiment we might run, suggested to me by the psychologist [omitted] (personal communication). This experiment puts the 'treating-personal-force-as-a-reason' hypothesis up against the Highlighter Hypothesis. We saw above that Greene et al. (2009a) found that when comparing pairs of actions, both of which involved an intention to harm someone as a means to an end, but only one of which involved the use of personal force, subjects rated the action that involved personal force as worse than the action that did not. According to the Highlighter Hypothesis, the presence of personal force serves to highlight the presence of *intentional harm*, which is, we're assuming, a bad-making feature of the action. If the Highlighter Hypothesis is correct, we should see the reverse effect when comparing subjects' reactions to pairs of *good* acts, one of which involves personal force and the other does not. If the Highlighter Hypothesis is correct, the presence of personal force in the one case ought to call subjects' attention to the feature in virtue of which the action in question is a good act, and thus, subjects ought to rate that action as *better* than they rate the good action that does not involve personal force. If, by contrast, subjects are treating personal force as a reason to believe that an action is bad, we would not expect this result.[44] As I mentioned, the

[44] As with any empirical study, such results could be explained by other factors (e.g., the 'Knobe effect'). Such factors will have to be carefully investigated and, to the extent that empirical science allows, 'ruled out', if the favored hypothesis is to be confirmed. Thanks to an anonymous reviewer for making this point.

Highlighter Hypothesis is just one version of the Facilitation Hypothesis. But I see no reason why similar experiments couldn't be run for other versions of the Facilitation Hypothesis, not to mention other versions of the treating-personal-force-as-a-reason hypothesis.

Everything I said in the previous paragraph is extremely rough. To carry-out experiments like these we must proceed cautiously, clearly delineating the hypotheses in question, what we would expect to find if those hypotheses were true, and what other resources might be available to explain away unpredicted results. But this is just par for the course for psychological research.

6. Where Things Stand Epistemically

As things currently stand, we do not yet have reason to conclude or deny that we are treating personal force as a reason to believe that certain actions are wrong. Where does this leave us, epistemically speaking, with respect to the relevant CDJs? Is agnosticism on the issue of whether we are treating personal force as a reason sufficient to undermine those CDJs? Our answer will depend on which if either of the following two principles, which I have adapted from Allan Hazlett (2012), is correct.

> Undercutting Principle. If it is rational for you to believe that your reasons for believing p are not good reasons, then you ought to stop believing p (unless you have independent reason to keep believing p).

> Feldman's Principle. If it is rational for you to suspend judgment about whether your reasons for believing p are good reasons, then you ought to stop believing p (unless you have independent reason to keep believing p).

Our current state of ignorance about whether we are treating personal force as a reason to endorse the relevant CDJs means that we satisfy the antecedent of Feldman's Principle but not the antecedent of the Undercutting Principle. Moreover, while the Undercutting Principle is widely (but not universally) endorsed, Feldman's Principle is considerably more controversial. Hazlett (2012) argues that the common arguments put forth for Feldman's Principle at best support the Undercutting Principle. Moreover, Hazlett argues that there is good, prima facie reason for rejecting Feldman's Principle—namely, that it apparently conflicts with the intuitively plausible view that sometimes epistemic peers might reasonably 'agree to disagree'. Katia Vavova (2014) argues for similar conclusions.

There's no space to rehearse Hazlett and Vavova's arguments here. Suffice it to say that if the empirical debunking of our CDJs is forced to rest on Feldman's Principle, the argument will be considerably more controversial than if it rests merely on the Undercutting Principle. Debunkers are thus left with two options: (1) produce the missing empirical evidence in support of the thesis that not only are our CDJs counterfactually sensitive to personal force, but that in forming those CDJs we treat the presence of personal force as a reason to believe as we do, or else (2) provide a convincing argument in support of Feldman's Principle. As I see it, this is where the debate over the normative significance of cognitive science currently stands.

Acknowledgements

[omitted]

Works Cited

Berker, Selim. 2009. 'The Normative Insignificance of Neuroscience'. *Philosophy and Public Affairs* 37: pp. 293 – 329

Campbell, Richmond and Victor Kumar. 2012. 'On the Normative Significance of Experimental Moral Psychology'. *Philosophical Psychology* 25: 311 – 330

D'Arms, Justin and Daniel Jacobson. 2000. 'The Moralistic Fallacy: On the 'Appropriateness' of Emotions'. *Philosophy and Phenomenological Research* 61: pp. 65 – 90

Dancy, Jonathan. 2000. *Practical Reality*. Oxford: Oxford University Press

Dancy, Jonathan. 2004. 'Ethics Without Principles'. Oxford: Oxford University Press

Duke, Aaron A. and Laurent Begue. 2015. 'The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas'. *Cognition*: 134: 121 – 127

Evans, Jonathan St. B. T. and Keith E. Stanovich. 2013. 'Dual-Process Theories of Higher-Cognition: Advancing the Debate'. *Perspectives on Psychological Science* 8: 223 – 241

Foot, Philippa. 1967. 'The Problem of Abortion and the Doctrine of Double Effect'. *Oxford Review* 5

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. 'An fMRI Investigation of Emotional Engagement in Moral Judgment'. *Science* 293: 2105 – 8

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: 389 – 400

Greene, Joshua. 2007. 'The Secret Joke of Kant's Soul' in Sinnott-Armstrong, Walter, ed., *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development Cambridge*. MA: MIT Press.

Greene, Joshua D., Sylvia A. Morelli, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2008. 'Cognitive Load Selectively Interferes with Utilitarian Moral Judgment'. *Cognition* 107: 1144 – 54

Greene, Joshua D., Fiery A. Cushman, Lisa E. Stewart, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2009a. 'Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment'. *Cognition* 111: pp. 364 – 371

Greene, Joshua D. 2009b. 'Dual- Process Morality and the Personal/Impersonal Distinction: A Reply to Mcguire, Langdon, Coltheart, and Mackenzie'. *Journal of Experimental Social Psychology* 45: 581 – 84

Greene, Joshua. 2014. 'Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics'. *Ethics* 124: pp. 695 – 726

Hazlett, Allan. 2012. 'Higher-Order Epistemic Attitudes and Intellectual Humility'. *Episteme* 9: 205 – 223

Joyce, Richard. 2006. *The Evolution of Morality*. MIT Press. Cambridge, MA.

Kahane, Guy. 2011. 'Evolutionary Debunking Arguments'. *Nous* 45: 103-125

Kahane, Guy, Katja Wiech, Nicholas Shackel, Miguel Farias, Julian Savulescu, and Irene Tracey. 2012. 'The Neural Basis of Intuitive and Counterintuitive Moral Judgment'. *SCAN*: 7: 393 – 402

Kamm, Frances. 2001. *Morality, Mortality Volume II: Rights, Duties, and Status*. Oxford University Press.

Kamm, Frances. 2009. 'Neuroscience and Moral Reasoning: a Note on Resent Research'. *Philosophy and Public Affairs* 37: 330 – 345

Lott, Micah. 2016. 'Moral Implications from Cognitive (Neuro)Science? No Clear Route'. *Ethics* 127: 241 – 256

McGuire, Jonathan, Robyn Langdon, Max Coltheart, and Catriona Mackenzie. 2009. 'A reanalysis of the personal/impersonal distinction in moral psychology research'. *Journal of Experimental Social Psychology* 45: 577 – 580

Mikhail, John. 2011. 'Emotion, Neuroscience, and Law: A Comment on Darwin and Greene'. *Emotion Review* 3: 293–95

Paxton, Joseph M., Tommaso Bruni, and Joshua D. Greene. 2014. *Social, Cognitive, and Affective Neuroscience* 9: 1368 – 1371

Peacocke, Christopher. 2004. *The Realm of Reason*. Oxford: Oxford University Press

Shafer-Landau, Russ. 2012. 'Evolutionary Debunking, Moral Realism, and Moral Knowledge'. *Journal of Ethics and Social Philosophy* 7: 1 – 37

Street, Sharon. 2006. 'A Darwinian Dilemma for Realist Theories of Value'. *Philosophical Studies* 127: pp. 109 – 166

Nina Strohminger, Richard L. Lewis, and David E. Meyer. 2011. 'Divergent Effects of Different Positive Emotions on Moral Judgment'. *Cognition* 119: 295 – 300

Rini, Regina. 2016. 'Debunking Debunking: a Regress Challenge to Psychological Threats to Moral Judgment'. *Philosophical Studies* 173: 675 – 697

Rini, Regina. *Forthcoming*. 'Why Moral Psychology is Disturbing'. *Philosophical Studies*

Schwitzgebel, Eric and Fiery Cushman. 2012. 'Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers'. *Mind and Language* 27: 135 – 153

Sudduth, Michael. 2015. 'Defeaters in Epistemology'. *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002, http://www.iep.utm.edu/

Thomson, Judith Jarvis. 1976. 'Killing, Letting Die, and the Trolley Problem'. *The Monist* 59: 204 – 21

Thomson, Judith Jarvis. 2008. 'Turning the Trolley'. *Philosophy and Public Affairs* 36: 359 – 374

Turri, John. 2010. 'On the Relationship between Propositional and Doxastic Justification'. *Philosophy and Phenomenological Research* 80: 312 – 326

Valdesolo, Piercarlo and David DeSteno. 2006. 'Manipulations of Emotional Context Shape Moral Judgment'. *Psychological Science* 17: 476 – 77

Vavova, Katia. 2014. 'Moral Disagreement and Moral Skepticism'. *Philosophical Perspectives* 28: 302 – 333

White, Roger. 2010. 'You Just Believe That Because...' *Philosophical Perspectives* 24: 573 – 615

Wood, Allen. 2007. *Kantian Ethics.* Cambridge: Cambridge University Press