

# Rules

Ron Mallon & Shaun Nichols

Is it wrong to torture prisoners of war for fun? Is it wrong to yank on someone's hair with no provocation? Is it wrong to push an innocent person in front of a train in order to save five innocent people tied to the tracks? If you are like most people, you answered "yes" to each of these questions. A venerable account of human moral judgment, influential in both philosophy and psychology, holds that these judgments are underpinned by internally represented principles or rules and reasoning about whether particular cases fall under those rules. Recently, this view has come under sustained attack from multiple quarters, and now looks to be in danger of being discarded. In this chapter we consider this evidence, and find that it does not support the elimination of rules from moral psychology.

## I. Moral rules and moral reasoning

Long traditions in religion, in law, in philosophy, and psychology connect moral judgment to moral rules. According to traditional rule-based accounts of morality, an action is wrong if it violates a moral rule. According to "rule utilitarians," (e.g., Brandt 1985) it is morally wrong to violate a rule that is justified by a balance of good consequence, while deontologists hold that there are rules – such as the prohibition against treating another person as a means to one's own end – that are wrong to violate whatever the consequences (Kant 1785/1964, Ross 1930). The central thread of these traditional approaches is *prescriptive*. For instance, deontologists maintain that murdering one innocent person to save two others really is wrong. It shouldn't be done. But for the purposes of this paper, our interests are entirely on the proper *descriptive* characterization of moral judgments. And there is a closely related *descriptive* claim that is also suggested by traditional rule-based accounts: the way a person actually comes to form a moral judgment depends on the person's application of a rule. An action is judged to be morally impermissible if the action violates a moral rule that is embraced by the judge.

In addition to the rich philosophical tradition of rule-based accounts of morality, there is a rich empirical tradition that adverts to rules as essential to certain normative judgments. In the literature on the moral/conventional distinction, it's widely agreed that at least for judgments of "conventional" violations (e.g., standing up during story time), these judgments depend on knowledge of local rules (see, e.g., Turiel et al. 1987). Thus there is reason to think that people make at least some normative judgments by drawing on their knowledge of rules. In addition, by appealing to agents' knowledge of local rules, we get an obvious explanation for cross-cultural differences in normative judgments. For example, people in the US but not people in China would think it wrong not to tip servers in local restaurants. The obvious explanation for this difference is that people in the US embrace a rule about tipping (in the US) and people in China do not embrace that rule about tipping (in China). Thus, there is independent reason to think that rules do play a role in at least some normative judgments.

Our aim here is not to defend the view that moral rules are the only factor in generating moral judgment, but rather to insist that moral rules are *one* crucial factor in the psychological processes that lead to moral judgments. In particular, we are defending an *internal* account of rules on which such rules are mentally represented and play a causal role in the production of

judgment and behavior.<sup>1</sup> Such rules come into play when they are thought to apply to a situation - i.e. when features of the situation instantiate properties that are represented in the rule, and one kind of rule we are concerned with represents properties traditionally considered to be of moral relevance (e.g. intention, injury). Consider, for example, the considerable evidence that judgments of moral responsibility for an act's consequences are sensitive to the intention with which the act was performed (e.g. Schultz et al 1981; Shultz & Wright 1985; Shultz et al 1986). A plausible explanation for this sensitivity is that judgments of responsibility typically require the satisfaction of rules that specify the relevance of intention. And this is exactly the kind of explanation that is offered in detail in "information-processing" theories of moral cognition (see Darley & Shultz 1990 for a review).

In arguing for the importance of moral rules, we follow influential traditions in both philosophy and psychology. In philosophy, much work on moral judgment can be seen as including an attempt to decide which properties and principles we ordinarily think of as morally relevant. Similarly, the tradition of information processing approaches to moral psychology can be seen also as attempting to discern which properties and rules give rise to moral judgments and behaviors (again, see Darley and Shultz 1990). Despite these influential traditions, moral rules have recently come to seem retrograde, a relic of best discarded views of moral judgment.

## II. Rules and Social Intuitions

A recent and provocative challenge to a rule-based approach to moral judgment is Jonathan Haidt's (2001) "social intuitionist" model. Haidt argues against a prominent role for moral reasoning in the production of moral judgment. Rather, he writes "[Understanding of moral truths occurs] not by a process of ratiocination and reflection, but rather by a process more akin to perception, in which one 'just sees without argument that they are and must be true' (Harrison, 1967, p. 72)" (2001, 814). In place of processes of reasoning, Haidt argues that moral judgments are typically caused by moral intuitions (including moral emotions) that are "a kind of cognition" but "not a kind of reasoning" (2001, p. 814).

Since all of the theories that are being considered are, effectively, causal models of moral judgment, it is useful to depict them with flow charts in which the boxes represent psychological mechanisms and the arrows represent causal relations. We might then depict a "rationalist" model as in figure 1. This is the model that Haidt has in his sights.

\*\*\*Figure 1 about here\*\*\*

In place of the rationalist model, Haidt offers a complex *social intuitionist* model, a crucial part of which we have illustrated in figure 2.

\*\*\*Figure 2 about here\*\*\*

Haidt marshals an impressive array of considerations to support his intuitionist model. Here, it might be helpful to focus on one sort of evidence that nicely illustrates Haidt's position: cases of moral dumbfounding. Haidt and colleagues presented the following case to subjects:

Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least, it

---

<sup>1</sup> Such a view thus rules out certain connectionist views of the mind on which (1) the mind is a connectionist network or networks and (2) such networks are nonrepresentational. Fortunately, these views are not true.

would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it okay for them to make love? (2001, 814)

Haidt reports that subjects presented with such cases and asked for a moral judgment typically answer "immediately" that the action was "wrong," and begin "searching for reasons" (814). Only, these cases have been designed by Haidt and colleagues to undermine the most obvious reasons for a moral judgment (for example, the risks and harms are eliminated, *ex hypothesi*). Thus, many subjects find themselves unable to provide reasons that justify their judgment. When faced with this dumbfounding, do subjects give up? No. They simply insist that their judgment is correct in the absence of reasons.

The conclusion Haidt draws from this and other evidence is that reasoning typically plays no role in the production of moral judgment. While Haidt allows that reasoning sometimes play a role in moral judgment and behavior, he holds (contra the rationalist model) that most moral reasoning is post facto, not causing moral judgment but rather being deployed in the process of justifying one's moral responses to others - just as the subjects in the dumbfounding case search for reasons for their judgments. If this is so, then why do we think we know the reasons for our judgments? Our ignorance is readily explained by a host of experimental evidence (some of it reviewed in Nisbett and Wilson's seminal [1977]; see Wilson 2002 and Stanovich 2004 for more recent coverage). that (i) people are not consciously aware of the processes that connect their mental states to the causal effects of those states and (ii) people routinely confabulate explanations for their behavior.

Haidt is relying here on what are sometimes called "dual process" models of cognition (Chaiken and Trope 1999), the fundamental idea of which is that cognition is subserved by two very different kinds of mechanisms. On the one hand, there are mechanisms characterized by conscious control, in which the "reasoner searches for relevant evidence, weighs evidence, coordinate evidence with theories, and reaches a decision" (Haidt 2001, 818). Call these *controlled* processes. On the other, there are processes that operate "quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness" (Ibid, 818). Call these *intuitive* processes. Haidt's move is, then, to suggest (on the basis of a range of evidence) that most moral judgment results from this second kind of system, and so he concludes that it is not a kind of reasoning.

There are a variety of ways to question Haidt's account of moral judgment without relying on what Haidt (2001, 815) calls philosophers' "worship of reason". Indeed, one of the best developed, dual process literatures within social psychology - that of implicit social attitudes - looks to provide an empirical basis for doubting Haidt's view.

Consider the substantial literature documenting unconscious (implicit) racial attitudes. It is a hallmark of this literature that a person can, on various indirect measures (for example, the "Implicit Attitudes Test" or IAT), exhibit racial biases, while on explicit measures (like self report) that same individual is not racially prejudiced.<sup>2</sup> In the most common version of the task, pictures of faces (black or white) are paired with positively or negatively valenced words in a sorting task. In such a test, subjects typically find it easier to perform the task when white faces

---

<sup>2</sup> Greenwald et al. 1998; see Nosek et al. (2007) for a review, and Kelly et al. (this volume) for more discussion.

are paired with good words and black faces with bad words than vice versa. And this effect occurs even when subjects appear to have explicit nonracist attitudes on paper and pencil questionnaires (e.g. Greenwald et al. 1998). The literature fits nicely within the dual-process framework, with the automatic, implicit processes producing one set of outcomes while consciously controlled processes produce another. But which processes are "in charge"? On Haidt's model, what drives moral responses are, in the first place, implicit, intuitive processes, and the reasoning processes come along after the fact. Is that what goes in the literature on racial attitudes? We suggest not.

Instead, the literature seems to show that conscious controlled processes exert substantial control over explicit verbal responses and behavior, and that, as these processes become overtaxed or exhausted (e.g. by increasing cognitive load or by so-called "ego depletion"), verbal responses and behavior come to align more and more closely with implicit attitudes. For example, Richeson and Shelton (2003) found that white subjects demonstrating an anti-black/pro-white IAT effect and then interacting with an African American confederate subsequently performed worse on tasks requiring effort and attention (i.e. a Stroop task).<sup>3</sup> The suggestion is that interacting with the African American confederate required additional control resources (to suppress automatic anti-black responses), thereby depleting a resource that would be needed on subsequent tasks. And Bartholemew et al. (2006) show that alcohol consumption interferes with the capacity to intentionally regulate implicit biases, an effect they link to a compromise of control mechanisms.

This is an important research program, and one that has revealed a startling array of implicit attitudes that may exert surprising effects on behavior. But the program does not show that intuitions run the show, only to be served by ex post facto reasoning. Rather, the correct model here seems quite a bit more like Plato's image of the soul as the charioteer (Phaedrus 246-254e) who must hold the reigns of the spirited horses (the automatic processes) closely so as to control them. When the charioteer becomes exhausted or drunk, the horses go wild, but that hardly shows that this is the typical case. Rather, it seems that where controlled processes maintain their integrity, connections between intuitions and behavior may be checked. At least the dual-process research on racial cognition provides a substantial body of evidence that in an important domain of real-world moral behavior, implicit racial biases processes are checked by controlled processes.<sup>4</sup> We take these to provide substantial reason to doubt that Haidt's general model of moral judgment is correct.

---

<sup>3</sup> Richeson et al. (2003) present brain imaging data suggesting that executive control inhibits the expression of implicit processes.

<sup>4</sup> Why not think that the implicit biases literature shows a competition between two automatic or intuitive processes – one generating anti-racist moral intuitions and one generating evaluative racial biases? The question for this alternate model is why compromising control (e.g. via exhaustion or drunkenness) alters the balance of power among intuitions (e.g. leaving anti-racist intuition weaker than evaluative racial bias, when normally it is stronger). An account that pits control against the automatic processes explains this, but an account that pits automatic processes against one another does not.

Still, this is far from definitive. Perhaps racial cognition is the exception with respect to the role of controlled processes. Perhaps elsewhere in moral cognition, conscious control plays little role.<sup>5</sup>

In truth, no one has made any serious attempt to count moral judgments in ordinary life. Nor is it obvious how one might approach such a daunting task. It is therefore difficult to find evidence that would definitively answer the question of whether moral judgments are typically caused in a particular way.

But even if intuitions do predominate to the extent that Haidt thinks they do, it doesn't settle the question of whether moral rules or moral reasoning figure in the production of moral judgment. This is because the distinction Haidt (following the dual-process orientation) offers between conscious 'reasoning' processes as opposed to automatically produced "moral intuitions" is simply the distinction between processes that are under direct conscious control, and those that are not.<sup>6</sup> And this distinction cross-cuts the category of inferential, rule-based processes that are at the core of the view we defend. To see this, consider that Haidt's case against the relevance of moral reasoning to moral judgment hangs largely on his characterization of reasoning processes as conscious or introspectively accessible. It is because, for example, the dumbfounded subjects seemingly employed no reflective reasoning to reach their moral judgments, and were unable to employ it to defend these judgments, that Haidt claims reasoning plays no role.

One might think that this is an odd way to characterize "reasoning" processes - indeed, Nisbett and Wilson's (1977) seminal work reports a variety of results involving complex cognitive processes - processes that are tempting to characterize as examples of reasoning - that seem to fail to be introspectible. For example, Storms and Nisbett (1970) experimented with insomniac subjects by placing them into three groups: arousal, relaxation, and control. Subjects in the first and second conditions were given placebo pills to take 15 minutes before bed on two consecutive nights, but the arousal condition subjects were told that the placebo would produce "rapid heart rate, breathing irregularities, bodily warmth, and alertness - symptoms, in other words, of insomnia" (Nisbett and Wilson 1977, 237). In contrast, relaxation subjects were told just the opposite, that the pill would produce "lowered heart rate, breathing rate, body

---

<sup>5</sup> Haidt produces a range of evidence to argue that the failure of introspective access on display in his dumbfounding cases is a typical feature of moral reasoning. One sort of response to Haidt, provided by Pizarro and Bloom (2003), is to question whether it matters if reflective reasoning typically plays little role in moral judgment. Pizarro and Bloom maintain that Haidt undersells the role of reasoning in moral judgment, for even one-off reasoning can have an enduring significance for an individual when it results in a lasting change to the individual's motivational make up. For instance, many individuals reason to the conclusion that it is immoral to eat mammals, and this often has long term ramifications on the individual's behavior. Even if Haidt has succeeded in showing that reasoning rarely drives judgment, this wouldn't be a great problem for defenders of the relevance of reasoning and moral rules. For defenders could, like Pizarro and Bloom, insist on the importance of these (numerically rare) cases.

<sup>6</sup> This is a little quick, for Haidt thinks unconscious processes typically share a host of other properties, including features of computational architecture. So, Haidt (2001) suggests that unconscious processes are computed nonsequentially and nonsymbolically (perhaps in a connectionist network). Because we see no reason to think that rule-based processes must be conscious, and because we have doubts that this aspect of dual-process theorizing is well motivated, we ignore this here.

temperature, and a reduction in alertness" (Ibid, 238). According to the subjects' reports, arousal subjects got to sleep significantly faster and relaxation subjects took significantly longer to get to sleep (with no change in control subjects).

The explanation Storms and Nisbett offered for this effect was simply that arousal condition subjects reattributed their insomnia symptoms to action of the pill, while relaxation condition subjects assumed their arousal must be particularly intense since they still felt their symptoms despite having taken a pill that would relax them. Nisbett and Wilson (1977, 238) are particularly struck by the fact that in post-experimental interviews, subjects "almost uniformly insisted that after taking the pills they had completely forgotten about them." That is, though the resulting behavior appears to be the result of a process of reasoning involving the comparison of an introspective assessment of one's arousal states with the an expected state, the subjects cannot recover this process of reasoning when asked later.

One explanation for this failure – the one that parallel's Haidt's explanation of his dumbfounding subjects – is that the reasoning processes involved are subserved by automatic, unconscious processes.<sup>7</sup> Suppose that is true. Then these processes would not be reasoning in Haidt's sense, even though they seem to involve complex inferences, determinations of relevant information, and other features characteristic of intelligent cognition. The right thing to conclude from this is that Haidt's use of 'reasoning' excludes mental processes that may nonetheless be inferential, rule based, and highly 'intelligent'. For example, Haidt's moral dumbfounding cases show nothing about whether moral rules or inferential processes were involved in the production of the moral judgments. They show only that whatever these processes are, they either (a) fail to be introspectively accessible shortly after completion or (b) fail to 'deactivate' quickly in the face of countervailing evidence. Option (a) seems the right explanation of at least some of the classic data regarding failures of self-knowledge – for example, the Nisbett and Storms data reviewed above. There is good reason to think an inferential process is occurring, but for whatever reason (because it is implicit, or because it is not well encoded in memory) this process cannot be recovered in response to questioning soon after. Option (b) might also explain Haidt's dumbfounding results: after all, the subjects do produce reasons for their judgments, but these reasons are refuted by the experimenter. It's only after this process of refutation that the subjects are dumbfounded, but perhaps once the judgment is made it is not quickly abandoned.

Having said all this, we are now in a position to simply stipulate that moral rules may play an important causal role in inferences without that process being consciously accessible, and therefore without being "reasoning" in Haidt's sense. Because Haidt's attack on conscious reasoning leaves the door wide open to rational, rule-governed inference at the unconscious level, his critique doesn't address whether moral rules play a role in moral judgment.

### **III. Rules and the Moral/Conventional Distinction**

A more direct challenge to the importance of moral rules comes from James Blair's (1995) explanation of performance on the moral-conventional task (e.g. Nucci 2001, Smetana 1993, Turiel 1983). Blair's work, like Haidt's, stresses the importance of a relatively automatic mechanisms underlying moral responses. But while Haidt discusses the relatively broad class of

---

<sup>7</sup> Another possibility is that the subjects simply cannot remember the contents of (what was) a conscious process.

what he calls intuitions (which includes moral emotions), Blair emphasizes the special importance of emotional response in moral judgment and behavior.

Blair offers a sophisticated account of the mechanisms underlying moral judgments in the moral/conventional task. In order to appreciate this model, it will be useful to review features of the moral/conventional task. Previous researchers found that a wide range of populations – adult and child, developmentally delayed and developmentally normal, autistic and nonautistic -- seem to treat “moral” violations (e.g. unprovoked hitting) differently than “conventional” violations (e.g. standing up during story time) (see, e.g., Blair 1996, Nichols 2002, Nucci 2001, Turiel et al. 1987) . Subjects tend to appeal to harm to the victim in explaining why a moral violation is wrong; for example, children say that pulling hair is wrong because it hurts the person. By contrast, children’s explanations of why conventional transgressions are wrong often proceed in terms of social acceptability; standing up during storytime is wrong because it’s rude or impolite, or because “you’re not supposed to.” Further, conventional rules, unlike moral rules, are viewed as dependent on authority; if the teacher at another school has no rule against chewing gum, children will judge that it’s not wrong to chew gum at that school, but even if the teacher at another school has no rule against hitting, children claim that it’s still wrong to hit.<sup>8</sup>

Blair maintains that the capacity to draw the moral/conventional distinction derives from the activation of a Violence Inhibition Mechanism (VIM). The idea for VIM comes from Konrad Lorenz’s (1966) suggestion that social animals like canines have evolved mechanisms to inhibit intra-species aggression: when a conspecific displays submission cues, the attacker stops. Blair suggests that there’s something analogous in our cognitive systems, the VIM, and that this mechanism is the basis for our capacity to distinguish moral from conventional violations.

According to Blair, VIM is activated by displays of distress and results in a withdrawal response (1995, 4). Moreover, VIM activation generates an aversive experience.<sup>9</sup> On Blair’s view, it is this feeling of aversiveness that generates the responses to the moral items on the moral/conventional task.

Thus stated, Blair’s account of VIM is an importantly incomplete account of moral judgment for people are wont to make moral judgments in cases where they witness no distress displays. For example, we suspect many people are likely to make a moral judgment upon hearing that in Salt Lake City during the summer of 2003, Mark Hacking shot his wife Lori in her sleep, disposed of her body in a trash dumpster, reported her missing, and personally led thousands of volunteers in a week long search for signs of her, all to conceal that he had lied to family and friends that he had been admitted to medical school. But in hearing this story, we witness the distress of neither Lori Hacking nor her family and community. Similarly, in canonical presentations of the moral-conventional task, there are no displays of distress.

---

<sup>8</sup> Kelly et al. (2007) launch an important critique of the moral/conventional distinction. We won’t draw on their critique here however. Rather, we will argue that Blair’s theory doesn’t work even if the moral/conventional distinction is viable.

<sup>9</sup> Blair’s theory has a further wrinkle. VIM-activation initially simply produces a withdrawal response. This VIM-activation becomes aversive through “meaning analysis”: “the withdrawal response following the activation of VIM is experienced, through meaning analysis, as aversive” (1995, p. 7). But Blair (1995) does not elaborate on what the “meaning analysis” comes to. Moreover, our critique will apply regardless of whether the notion of “meaning analysis” is invoked. Hence, in our discussion of Blair we will simplify and say that VIM causes the experience of aversion.

Blair extends the VIM model to these cases via classical conditioning:

During normal development, individuals will witness other individuals displaying distress cues resulting in the activation of VIM. On many occasions the observers may role take with the distressed victims; they will calculate representations of the victim's internal state (e.g. "she's suffering"; "what a poor little boy"; "he must be cold and hungry").

There will thus be pairings of distress cues activating VIM with representations formed through role taking. It is suggested here that the representations formed through role taking will become, through classical conditioning, trigger stimuli for VIM. .... Thus, an individual may generate empathetic arousal to just the thought of someone's distress (e.g. "what a poor little boy") without distress cues being actually processed. (1995, 4-5).

On the Blair model, then, there are two processes important to the development of typical moral response, with the former ontogenetically prior to the latter:

Distress Cues --> VIM

Role taking --> Representations of Distress --**conditioned response**--> VIM

His explanation of the moral-conventional distinction is, then, simply that the violation of moral rules gives rise (via role taking) to representations of a victim's distress. And these representations activate VIM (in normally developed individuals).

Blair's account competes with the moral rules account we defend, because it proposes to account for all moral judgment via the association of moral violations with victims' distress, mediated by an aversive emotion-like response. For present purposes, it's useful to simplify the model as an "emotionist" model, as depicted in figure 3.

\*\*\* figure 3 about here\*\*\*

This model offers a clear and radical alternative to a rule-based or rationalist account of moral judgment.<sup>10</sup> For Blair's emotionist model maintains that moral judgment is caused by a particular kind emotional response that is caused by cues of suffering. In contrast, the moral rules account we want to defend suggests that assessing situations for the presence more theoretical properties (like "intention," "right," "injury" and so forth) is crucial to the activation of moral judgment.

So which account is right? We think it is clear that Blair's account will not explain moral judgment because distress, and the representation of distress, are not sufficient for judgments that an act (or act type) is "wrong." Perhaps the easiest way to illustrate this is by exploiting the venerable distinction between judging something *bad* and judging something *wrong*. Many occurrences that are regarded as bad are not regarded as wrong. Toothaches, for instance, are bad, but they aren't wrong. The moral/conventional task gets its interest primarily because it gives us a glimpse into judgments of *wrong*. This is reflected by the fact that the items in the moral/conventional task are explicitly *transgressions*, and the very first question in standard moral/conventional tasks checks for the *permissibility* of the action. As we'll see, the problem with Blair's account is that, while the proposal might provide an account of judging something

---

<sup>10</sup> Indeed, Blair's more recent work offers a less radical model of moral judgment (see, e.g., Blair et al. 2008). We focus on his original theory because it provides such a delightfully clear and testable emotionist model of moral judgment.

bad (in a certain sense), it does not provide an account of judging something wrong (Nichols 2002; 2004).

If Blair's theory is right, VIM leads to a distinctive aversive response. As with toothaches, we might regard the stimuli that prompt this aversive response as "bad". Furthermore, it might be important to treat stimuli that produce VIM-based aversion as "bad" in a distinctive way. Now, what is the class of stimuli that are bad in this sense? Well, anything that reliably produces VIM activation. Distress cues will be at the core of this stimulus class (Blair, 1995, 1999a). The class of stimuli that will be accordingly aversive will include distress cues from victims of natural disasters and accidents and even distress cues in paintings and drawings. Thus, the class of stimuli that VIM will lead us to regard as "bad" includes natural disaster victims, accident victims, and superficial distress cues. But it is quite implausible that these things are regarded as *wrong*. Natural disasters are, of course, bad. But, barring theological digressions, natural disasters aren't regarded as *wrong*. Similarly, if a child falls down, skins her knee, and begins to cry, this will produce aversive response in witnesses through VIM. Yet the child's falling down doesn't count as a moral transgression. This was put to the test in a recent experiment by Leslie, Mallon & Dicorcia (2006). Children were presented with a scenario in which a child exhibited distress cues, but only because the child was a "crybaby". Children were told a story in which two characters, James and Tammy were eating their lunch. Each of them had a cookie, but James wanted Tammy's cookie as well as his own. Their teacher says "In this school, anybody can eat their own cookie if they want to. Anybody can eat their own cookie." This makes Tammy very happy and she eats her cookie, but that makes James unhappy, and he cries. When presented with this scenario, both four year old children and children with autism thought that it was perfectly okay for Tammy to eat her cookie, even though it made James cry. So the mere fact that Tammy's action led to James' distress wasn't enough to generate a judgment of a moral transgression.

Thus, while Blair's theory might provide an account of how people come to judge things as *bad* in a certain sense, it does not provide an adequate account of moral judgments of *wrong* on the moral/conventional task. The natural explanation for why the problem arises for Blair's account is that the account fails to include *rules* in the processes that generate moral judgment. If we invoke rules, then we can easily explain why it isn't judged wrong to have a toothache, fall off your bike, or eat your own cookie (knowing full well that it will make Jimmy cry). None of these are judged wrong because they don't violate the internally represented rules. Now there are a variety of detailed ways that Blair's account might be amended. For example, rather than pairing representations of distress with VIM activation, Blair might pair transgression types (e.g. intentional harm) with VIM activation. Alternatively, Blair might hold that in some of these cases other cognitive mechanisms or what he calls "executive functions" may override the connection between VIM and moral judgment. But we suggest that these emendations to Blair's account, to the extent they are successful, are simply ways of supplementing the account with internally represented rules.

#### **IV. Rules and Moral Dilemmas**

The final important threat to rules comes from an emerging body of work on the psychological factors involved in assessing moral dilemmas. As with Blair's work, one of the important results from research on dilemmas has been to show how emotions can impact moral judgment (Greene et al. 2001). And this gives rise to an emotion-based account of moral

judgment that threatens to displace a rule-based account. However, again, we will argue that moral rules play a vital role in the psychology underlying the assessment of moral dilemmas.

This recent work on moral dilemmas grows out of a large body of research in philosophy which draws out our intuitions about a wide range of dilemmas and attempts to determine a set of principles that captures our intuitions about the cases.<sup>11</sup> The most intensively studied dilemmas are the “trolley cases”, which serve to isolate different factors that might affect our intuitions. In the *bystander* case, we are asked to imagine that a person sees a train approaching that will kill five innocents on the track, and the only way to prevent the deaths of these five is to flip a switch that will divert the train to a side track. Diverting the train to the side track will lead to the death of the person on the side track, but it is the only way to save the five people on the main track. Philosophers have maintained that the intuitive position is that it is acceptable to flip the switch to divert the train, leading to the death of one instead of five (e.g., Thomson 1976). In the *footbridge* case, the situation is quite similar except that there is no side track, and the only way for the protagonist to save the five is to push a large stranger off of a footbridge in front of the oncoming train, which will kill the stranger. In this case, philosophers maintain that the intuitive position is that it is wrong to push the stranger (Foot 1967, Quinn 1989, Thomson 1976). This can seem puzzling since, on a simple utilitarian calculus, the cases seem quite parallel: five lives can be saved for the price of one. One goal of the philosophical investigations has been to develop a unified normative theory that will accommodate intuitions about such cases. This goal has been exceedingly difficult to meet, and few would maintain that philosophers have succeeded in finding a unified normative theory that fits with the full range of our intuitions about moral dilemmas.

This work in philosophy was unapologetically *a priori*, but recently researchers have conducted interview and survey experiments with these sorts of cases (Petrinovich & O’Neill 1996, Mikhail 2000, Greene et al. 2001, Hauser et al. 2007). The results have largely confirmed what philosophers maintained about the bystander and footbridge cases: most people have the intuition that it is acceptable to flip the switch in *bystander* but that it is not acceptable to push the stranger in *footbridge*. The interesting subsequent question concerns the psychological underpinnings of these judgments. Why do people judge pushing the stranger as inappropriate but turning the train as appropriate?

In an important recent discussion, Joshua Greene proposes that the response in footbridge-style cases is generated by the fact that these actions are “personal” and such actions generate greater emotional engagement than “impersonal” actions. The personal/impersonal distinction is drawn as follows:

A moral violation is personal if it is: (i) likely to cause serious bodily harm, (ii) to a particular person, (iii) in such a way that the harm does not result from the deflection of an existing threat onto a different party... A moral violation is impersonal if it fails to meet these criteria. .... Pushing someone in front of a trolley meets all three criteria and is therefore “personal,” while diverting a trolley involves merely deflecting an existing

---

<sup>11</sup> In addition to this descriptive philosophical project, there is a related *prescriptive* project which attempts to characterize the normative theory that *should* guide our judgments in these cases. Some of this literature seems to take the view that identifying what our intuitions are to dilemma cases (which is in some sense a descriptive question) will provide important input for characterizing the normative theory that *should* guide our judgments (e.g. Thomson 1976)

threat, removing a crucial sense of “agency” and therefore making this violation “impersonal” (Greene & Haidt 2002, 519).<sup>12</sup>

Thus, Greene maintains that footbridge cases elicit inappropriateness judgments because they trigger emotional responses, and they trigger emotional responses because they are *personal*. In support of this account, Greene and colleagues provide evidence that emotional processing plays a key role when people consider footbridge cases (Greene et al. 2001; Greene et al. 2004). The proposal is that the key difference between footbridge and bystander – the difference that generates the different response – is that footbridge is *personal* and bystander is *impersonal*.

Greene incorporates this distinction into a dual process approach to moral judgment, as depicted in figure 4.

\*\*\*Figure 4 about here\*\*\*

According to Greene, impersonal dilemmas tend to activate the ‘reason’ path, whereas personal dilemmas run through the ‘emotion’ path. Like Haidt, Greene’s model suggests that we can arrive at moral judgments either through reasoning or through emotions. But unlike Haidt, Greene never suggests that our reasoning-based judgments are rare. He does, however, suggest that reason-based judgments are characteristically utilitarian.

Greene’s focal cases are dilemmas like Bystander and Footbridge, but of course, if his theory aims merely to provide a theory of judgments about trolley cases, it is of limited interest. Greene’s approach is particularly interesting and provocative if it is taken to suggest a general account of moral judgment. That is, one might maintain that quite generally, personal acts generate moral condemnation through the emotional pathway in the dual process model. However, if we take this as a general account of moral judgment, there are numerous *prima facie* counterexamples, cases in which manifestly personal (and emotionally salient) acts are not judged impermissible. Just as Blair’s account runs afoul of cases in which distress is insufficient for moral judgment, the appeal to “personal” acts runs afoul of cases in which acts are personal but permissible. For example, some acts of self-defense, war, and punishment are plausibly personal and emotional, but regarded as permissible nonetheless. For instance, many people think that spanking their own child is permissible, even though it is obviously personal and emotional. Similarly, there is cultural variation in the harms that are judged impermissible. Among Yanomamö men, wife beating is judged permissible, despite being personal and emotional (Chagnon 1992). Closer to home, in much of Western culture, male circumcision is permissible though obviously personal. So the idea that actions are judged to be wrong when they’re personal faces *prima facie* concerns.

Greene does discuss situations in which being personal is insufficient to result in a moral judgment of wrongness. Thus, he allows for cases in which an action may be judged appropriate despite being personal. However, the cases he discusses in which the personal is insufficient to

---

<sup>12</sup> One aspect of Greene’s description requires some clarification. What is meant by ‘moral violation’ in this passage? It doesn’t seem to mean *transgression* because diverting the trolley in the bystander case is cast as an ‘impersonal moral violation’, but it’s doubtful that that action is a transgression at all. Indeed, if transgressions are measured by judgments of permissibility, the available evidence indicates that diverting the train isn’t a transgression – in all of the extant studies, a majority of participants judge it to be permissible (Greene et al. 2001, Hauser et al. forthcoming, Mikhail 2000). Since ‘violation’ is easily confused with ‘transgression’, we will set aside this terminology in characterizing Greene’s account.

generate moral judgment are cases in which the utilitarian calculus is obviously in favor of acting. For instance, in one dilemma, the “crying baby” case, subjects have to decide whether it’s permissible to smother one’s baby in order to prevent the Nazis from discovering and killing your entire family (including the baby). Greene maintains that although such cases involve personal acts, many people draw on utilitarian considerations to reach the judgment that the actions are permissible. This explanation comports well with the general account he offers in which each of the two systems (emotions, utilitarian reasoning) in his dual-process model are in competition with one another. Taking these examples and Greene’s model together, Greene’s model (generalized as an account of moral judgment) suggests the following two regularities obtain:

(a) If an action is personal, it must maximize utility to be permitted. (e.g. crying baby)

(b) If an action maximizes utility, it must be personal to be prohibited. (e.g. footbridge)

We agree with Greene that something like utilitarian considerations can play an important role in reasoning about these kinds of dilemmas, but we doubt that regularity (a) obtains and that this model is complete.<sup>13</sup> For we doubt that the appeal to utilitarian considerations could explain the permissibility of the cases above. For instance, it’s implausible to maintain that the reason people think it’s okay to punish people is because it maximizes utility (cf. Haidt & Sabini 2000; Fehr & Gächter 2002; see Prinz & Nichols this volume for a brief presentation of these results). If this is correct, then Greene’s appeal to personal acts together with utilitarian considerations fails to provide an adequate account of moral judgment.

Even if one concedes that the appeal to personal acts doesn’t explain moral judgment generally, one might maintain that it explains the asymmetry in judgments in footbridge and bystander trolley cases. However, we think that even in this restricted domain, the appeal to personal acts and utilitarian considerations is inadequate to explain the phenomena. The reader will not be surprised to learn that we think that rules provide a critical supplement to the factors that Greene invokes. Unlike the appeal to *personal acts*, the rule-based approach with which we began has an obvious explanation for why personal acts like self-defense, punishment, and circumcision are not judged impermissible. The judge doesn’t embrace a rule against them. Thus, once again, we find that the traditional rule-based account gives a very natural explanation for some obvious facts about people’s moral judgments.<sup>14</sup> We suggest that rules play an important

---

<sup>13</sup> In fact, we also doubt that (b) is correct for we suspect subjects may judge a utility maximizing lie to be morally wrong, at least in some cases. Of course, it may be that such cases show the limits of generalizing Greene’s explanation beyond the tightly circumscribed moral dilemmas (e.g. trolley cases) characteristic of this literature.

<sup>14</sup> One important response to our view is that the culturally specified rules act as *overrides* to the natural moral judgment, which itself is a basic and universal emotional reaction. We regard it as an open question whether the rules act as ‘overrides’ to a basic emotion. But we do maintain that if we identify ‘moral judgment’ with this basic emotional response, then we have no longer provided an account of moral judgment that answers to the primary question at hand, which concerns why people think what they do about morally fraught situations. The standard measures here have been, of course, their spontaneous judgments about scenarios. Some of the scenarios are difficult moral dilemmas; others, like the scenarios in the moral/conventional task, involve less conflict but still generate spontaneous judgment. Our point is that even if rules provide an ‘override’ to the emotional response, this override is typically already factored in to the spontaneous judgments that we are trying to explain.

role, not just in judgments about the problematic cases we've pressed, but also in judgments about the moral dilemmas that dominate this literature.

The rule-based approach does owe an answer to the original challenge, however. Why do people judge that choosing five over one is acceptable in the bystander case but not in the footbridge case? The traditional answer is that the different status of these actions is explained by what the rules do and do not forbid. One important proposal is that a rule like "Do not kill persons" forbids one from intending to bring about a person's death, but it does not necessarily forbid acting in a way that brings about a person's death as an unintended but foreseen side effect. Hence, in the bystander case, it is permissible to divert the trolley even though it has an effect (the killing of an innocent) that it would be impermissible to intend. There has been protracted debate on how exactly to characterize what moral rules do and do not forbid (Foot 1967, Thomson 1976, Quinn 1989). Sorting all this out has been enormously complicated and has not produced any consensus, but we need not delve into this debate. For the important point is simply that the traditional advocate of rule-based moral judgment maintains that we can explain the intuitions about the trolley cases in terms of what the rules do and do not forbid.

The rule-based approach offers an alternative to the emotion-based explanation of the asymmetry between the footbridge and bystander cases. However, in light of the apparent failure of philosophers to achieve consensus on how to characterize what the rules do and do not forbid, the rule-based explanation of the asymmetry between the footbridge cases and bystander cases may seem ad hoc. Is there an independent way to support the claim that the asymmetry between the footbridge cases and the bystander cases is explained by what the rules do and do not forbid? Our hypothesis is that it is a common feature of many rules, not specific to personal contexts, that they exhibit the asymmetry reflected in the footbridge and bystander cases. We recently conducted a set of experiments on adults to test this hypothesis (Nichols & Mallon 2006).

The idea behind the experiments was to create cases that parallel the footbridge and bystander cases, but which do not count as personal. We did this by altering the scenarios such that the possible human victims were replaced with teacups. Here are the two central cases:

**Impersonal bystander case:**

When Billy's mother leaves the house one day, she says "you are forbidden from breaking any of the teacups that are on the counter." Later that morning, Billy starts up his train set and goes to make a snack. When he returns, he finds that his 18 month old sister Ann has taken several of the teacups and placed them on the train tracks. Billy sees that if the train continues on its present course, it will run through and break five cups. Billy can't get to the cups or to the off-switch in time, but he can reach a lever which will divert the train to a side track. There is only one cup on the side track. He knows that the only way to save the five cups is to divert the train to the side track, which will break the cup on the side track. Billy proceeds to pull the lever and the train is diverted down the side track, breaking one of the cups.

Did Billy break his mother's rule? YES NO

**Impersonal footbridge case:**

When Susie's mother leaves the house one day, she says "you are forbidden from breaking any of the teacups that are on the counter." While Susie is playing in her bedroom, her 18 month old brother Fred has taken down several of the teacups and he has

also turned on a mechanical toy truck, which is about to crush 5 of the cups. As Fred leaves the room, Susie walks in and sees that the truck is about to wreck the cups. She is standing next to the counter with the remaining teacups and she realizes that the only way to stop the truck in time is by throwing one of the teacups at the truck, which will break the cup she throws. Susie is in fact an excellent thrower and knows that if she throws the teacup at the truck she will save the five cups. Susie proceeds to throw the teacup, which breaks that cup, but it stops the truck and saves the five other teacups.

Did Susie break her mother’s rule? YES NO

The results were clear. In two different experiments, most subjects said that the rule was broken in the impersonal footbridge case, but fewer than half said that the rule was broken in the bystander case. The difference between the responses was highly significant (see figure 1).

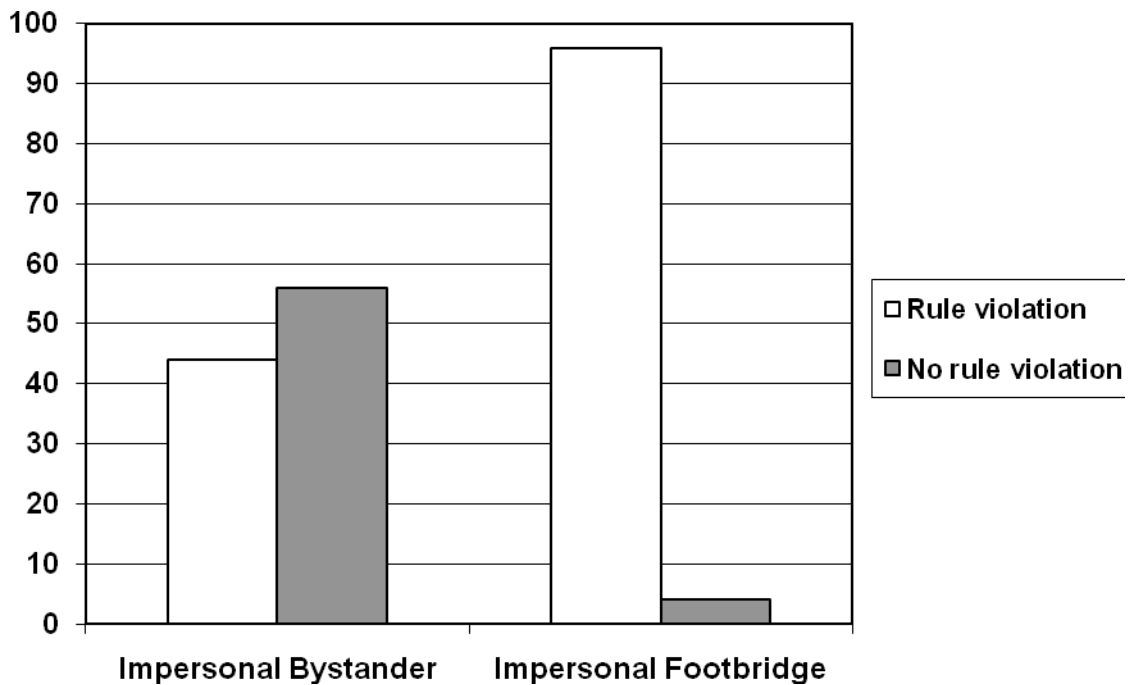


Figure 1: Responses to impersonal moral dilemmas (Nichols & Mallon 2006, experiment 2)

These results support the rule-based interpretation of the results on the original trolley cases. For it seems that even in impersonal cases, subjects distinguish footbridge-style cases from bystander-style cases. This provides independent reason to think that traditional rule-based accounts illuminate the asymmetric responses on the original bystander and footbridge cases. Namely, the judgment of impermissibility in the footbridge cases is guided by a moral rule that is not violated in the bystander cases.

## V. Rule Violations and All-In Permissibility

In the previous sections, we have tried to beat back the emotion-based attacks on moral rules in recent moral psychology. However, it is not clear that we can wring a complete account of moral judgment from the appeal to rules alone. The results discussed above leave open the possibility that people might judge that an action violates a rule and also judge that the action is acceptable, *all things considered*. For instance, stealing bread violates a rule, but if a person is starving, we might think that it is okay for him to steal the bread, *all things considered*.

In the impersonal footbridge case, we know that participants tend to regard the action as a rule violation (“it breaks her mother’s rule”). But we don’t yet know whether participants would regard the action as impermissible in the all-in sense. So in addition to the rule violation question, we also asked a question to get at judgments of all-in permissibility:

All things considered, was it okay for Susie to throw the teacup? YES NO  
Most participants said Yes to the all-in question on the impersonal footbridge case. Indeed, most subjects said both that Susie broke her mother’s rule and that what she did was, all things considered, okay.

So, participants recognize a distinction between breaking a rule and all-in impermissibility, at least in the case of some rules. But perhaps this complication doesn’t arise in the case of moral rules. In philosophical ethics, one important view, absolutist deontology, maintains that if an action violates a moral rule, it is thereby the wrong thing to do, all things considered (e.g. Fried 1976). In addition, some rule-utilitarians (e.g. Brandt 1985) also endorse the primacy of moral rules, even in the face of consequences that favor breaking the rule. On such views, breaking a moral rule is (at least typically) sufficient for all-in impermissibility. If either account is a correct view of the role of moral rules in actual moral judgments, then we have a very plausible explanation of the original footbridge results. In those case, the action violates a moral rule, and violating a moral rule suffices for generating a judgment of all-in impermissibility.

However, there is now converging evidence from several different methodologies that suggest that emotions do make a critical contribution to moral judgment (e.g. Blair 1995, Greene et al. 2004, Valdesolo & DeSteno 2006, Koenigs et al. 2007). In particular, judgments of “all-in permissibility” are plausibly influenced by emotional activity. This suggests that no adequate account of moral judgment can be given in terms of rules alone. But of course this doesn’t mean that we should give up rules altogether. On the contrary, we’ve spent this entire chapter arguing that an adequate account of moral judgment cannot ignore the role of rules. A natural proposal at this stage is that emotions contribute to the salience and importance of the rules. That is what elevates certain rules, and certain applications of rules, above other morally relevant considerations (Nichols & Mallon 2006).

It’s important to note that, although this model of non-utilitarian moral judgments invokes separate processes, it does so in a way that is quite different from typical dual process models offered in recent moral psychology (e.g. Greene & Haidt 2002). Typical dual process accounts depict two systems vying for control: the rational versus the emotional; the smart versus the stupid; the slow versus the quick. As we noted above, this is especially clear in Greene’s account of what happens in moral dilemmas. The rational system votes for pushing the man in front of the footbridge, the emotional system votes against pushing the man, and the stronger signal carries the day. Like Greene, we think that utilitarian considerations contribute to judgments of all-in permissibility. Furthermore, we would allow that utilitarian considerations can be in competition with non-utilitarian patterns of thought. Indeed, that’s what makes the

dilemmas vexing in the first place. But we have a different view of the psychological character of *non-utilitarian* judgment. Our model of non-utilitarian judgment invokes separate factors – rules and emotions – but they are not in competition with each other. Rather, the emotions and the rules combine to produce the judgment. Of course there might still be competition when these judgments are pitted against other considerations. But the explanation for why people judge, for example, that it’s wrong to push the man in footbridge, appeals to multiple systems working together to produce the judgment. In particular, it depends on the integration of emotional reactions and internally represented rules. Thus, instead of thinking of this as a dual *process* model, on which moral judgment comes from either reason or emotion, we promote a dual *vector* model of non-utilitarian moral judgment. This is depicted in figure 5.

\*\*\*Figure 5 about here\*\*\*

When people make non-utilitarian moral judgments, as in the footbridge case and the moral transgressions in Turiel-style cases, both emotional and rule-based processes are implicated. As reflected in the figure, however, it remains quite unclear how the emotions and rules interact to generate the judgment.

To reinforce the point that rules are critical to such cases of moral judgment, we want to consider one last experiment that has been recently celebrated (e.g., Kelly forthcoming; Nado et al. forthcoming). In an experiment involving emotion induction, Wheatley and Haidt gave participants a case that involved “no violation of any kind” (2005, 782):

Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He [tries to take/often picks] topics that appeal to both professors and students in order to stimulate discussion.

The trick was that the subjects had received hypnotic instructions that would lead them to feel disgust at the mention of a certain word (either ‘often’ or ‘take’), and the experiment was set up so that half of the subjects would get their disgust trigger and the other half wouldn’t. What they found was that subjects in the disgust condition did give significantly higher ratings on a scale of moral wrongness (p. 783). However, it is critical to pay attention to the actual numbers here. On a scale from 1 (“not at all morally wrong”) to 100 (“extremely morally wrong”), the mean rating for disgust-subjects was still only 14. And that is obviously well on the side of ‘not morally wrong’. Thus, the one data point we have in which emotion is induced in the absence of a rule, we find that moral condemnation does *not* emerge.

## VI. Conclusion

The recent wave of research showing a role for emotions in moral judgment has generated a great amount of interest, and we think that this is entirely justified. The discovery and elucidation of the important role that emotions play is perhaps the most exciting pattern of results in moral psychology. However, we should not let our enthusiasm for the emotion-based results lead us into thinking that moral judgment (or even just “deontological” moral judgment) can be identified with an emotional response. Rather, the capacity for moral judgment is rich and complex. The evidence suggests that emotions are one important part of the story, but rule-based accounts also capture an important element of the psychological underpinnings of moral judgment. While we think this much is clear, we think it remains unclear how rules interact with emotions to produce the patterns of moral judgments that we see.

References:

- Bartholow, B.D., Dickter, C.L. and Sestir, M.A. 2006. Stereotype Activation and Control of Race Bias: Cognitive Control of Inhibition and Its Impairment by Alcohol. *Journal of Personality and Social Psychology* 90: 272–287.
- Blair, R. (1995). A cognitive developmental approach to morality: investigating the psychopath, *Cognition*, 57, 1-29.
- Blair, J., Marsh, A., Finger, E., Blair, K. & Luo, J. (2006). Neuro-cognitive systems involved in morality. *Philosophical Explorations*, 9 (1), 13-27.
- Brandt, R. (1985). *A Theory of the Good and the Right*, Oxford: Clarendon Press.
- Chagnon, N. (1992). *Yanomamö*, 4<sup>th</sup> edition. New York: Harcourt Brace Jovanovich.
- Chaiken, S. & Trope, Y. (Eds.). 1999. *Dual process theories in social psychology*. New York: Guilford.
- Darley, J.M. & Shultz, T.R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, 41, 525- 556.
- Davidson, P.E. Turiel, and A. Black (1983). “The Effect of Stimulus Familiarity on the Use of Criteria and Justifications in Children’s Social Reasoning,” *British Journal of Developmental Psychology*, 1, 49-65.
- Davis, N. (1991). Contemporary deontology. In P. Singer (ed.) *A Companion to Ethics*, 205-218.
- Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5. Reprinted in *Virtues and Vices*, Oxford University Press.
- Fried, C. (1978). *Right and wrong*. Cambridge, MA: Harvard University Press.
- Greene, J. (forthcoming). Cognitive neuroscience and the structure of the moral mind. In P. Carruthers, S. Laurence, and S. Stich (eds.) *The Innate Mind: Structure and Content*. New York: Oxford University Press.
- Greene, J. & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science*, 612
- Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment, *Science*, 293, 2105-08.
- Greenwald, A., McGhee, D. & Schwartz, J. 1998. “Measuring Individual Differences in Implicit Cognition: The Implicit Association Test,” *Journal of Personality and Social Psychology*, 74(6): 1464-1480.

- Haidt, J. (2001). The emotional dog and its rational tail. *Psychological Review*.
- Haidt, J., & Sabini, J. (2000). What exactly makes revenge sweet? (Unpublished manuscript, University of Virginia).
- Hauser, M., Cushman, F., Young, L., Jin, R., and Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, 22.
- Kant, I. (1785/1964). *Groundwork of the Metaphysics of Morals*, translated by H.J. Paton. New York: Harper & Row.
- Kelly, D. forthcoming. Moral Disgust and Tribal Instincts: A Byproduct Hypothesis. *Connected minds: Cognition and interaction in the social world. Proceedings of Cognition Conference 2007*.
- Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, Affect and the Moral/Conventional Distinction. *Mind & Language*, 22.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A. 2007. Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, 446, 908-911.
- Leslie, A., Mallon, R. & Dicorcia, J. (2006). Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Social Neuroscience*, 1, 270-283.
- Mikhail, J. (2000). *Rawls' Linguistic Analogy*. Unpublished Ph.D. thesis, Cornell University.
- Nado, J., Kelly, D., and Stich, S. forthcoming. Moral judgment. In *Routledge Companion to the Philosophy of Psychology*, ed. by John Symons & Paco Calvo.
- Nichols, S. (2002). Norms with feeling: towards a psychological account of moral judgment, *Cognition*, 84, 221-236.
- Nichols, S. (2004). *Sentimental rules: on the natural foundations of moral judgment*. New York: Oxford University Press.
- Nichols, S. & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100.
- Nisbett, R. and Wilson, T. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes", *Psychological Review* 84, 231-259.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at **age 7**: A methodological and conceptual review (Pp. 265–292). In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior*. Psychology Press.

- Nucci, L. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.
- Petrinovich, Lewis, and O'Neill, Patricia. 1996. "Influence of Wording and Framing Effects on Moral Intuitions." *Ethology and Sociobiology* 17, 145-171.
- Pizarro, D. A., & Paul Bloom (2003). The intelligence of the moral intuitions: A reply to Haidt (2001). *Psychological Review*, 110, 193–196
- Prinz, J. (forthcoming). *The emotional construction of morals*. Oxford: Oxford University Press.
- Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect, *Philosophy and Public Affairs*, 18.
- Richeson, J., Baird, A., Gordon, H., Heatherton, T., Wyland, C., Trawalter, S. & Shelton, N. 2003. "An fMRI investigation of the impact of interracial contact of executive function," *Nature Neuroscience*, 6(12): 1323-1328.
- Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, 14, 287-290.
- Ross, W. (1930). *The Right and the Good*. Oxford: Clarendon Press.
- Smetana, J. (1993). Understanding of social rules. In M. Bennett (ed.) *The development of social cognition : the child as psychologist*. New York: Guilford Press, 111-141.
- Stanovich, K. (2004). *The Robot's Rebellion*. University of Chicago Press.
- Storms, M. & Nisbett, R. (1970). Insomnia and the attribution process. *Journal of Personality and Social Psychology*, 2, 319-328.
- Thomson, J. (1976). Killing, letting die, and the trolley problem, *The Monist*, 59, 204-217.
- Turiel, E. (1983). *The development of social knowledge: morality and convention*, Cambridge: Cambridge University Press.
- Turiel, E., Killen, M., & Helwig, C. (1987). Morality: Its structure, functions, and vagaries. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children*. University of Chicago Press.
- Valdesolo, P., and DeSteno, D. 2006. "Manipulations of Emotional Context Shape Moral Judgment." *Psychological Science* 17: 476–77.
- Wheatley, T., and Haidt, J. 2005. "Hypnotically Induced Disgust Makes Moral Judgments More Severe." *Psychological Science* 16: 780–84.
- Wilson, T. (2002). *Strangers to Ourselves*. Harvard University Press.

Zelazo, P., C. Helwig, A. Lau (1996). "Intention, Act, and Outcome in Behavioral Prediction and Moral Judgment," *Child Development*, 67, 2478-2492.

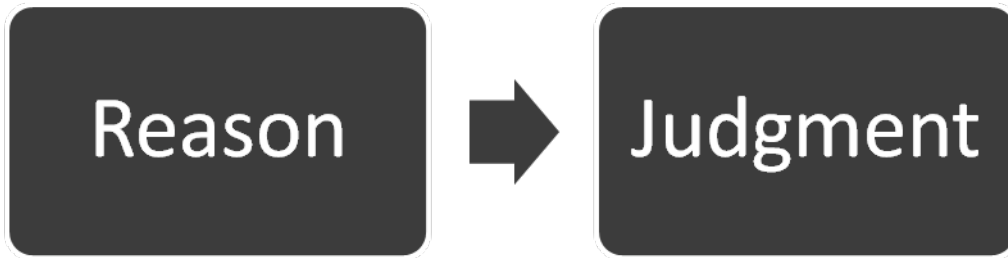


Figure 1: Rationalist model of moral judgment

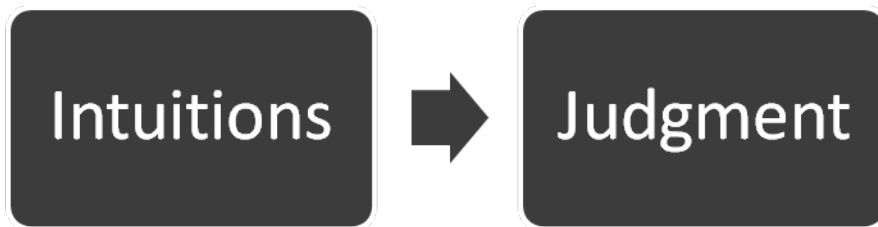


Figure 2: Intuitionist model of moral judgment

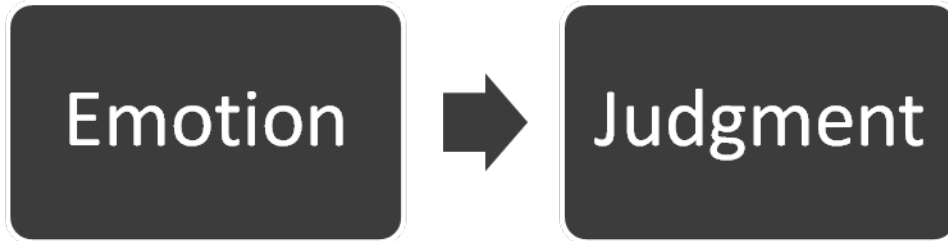


Figure 3: Emotionist model of moral judgment

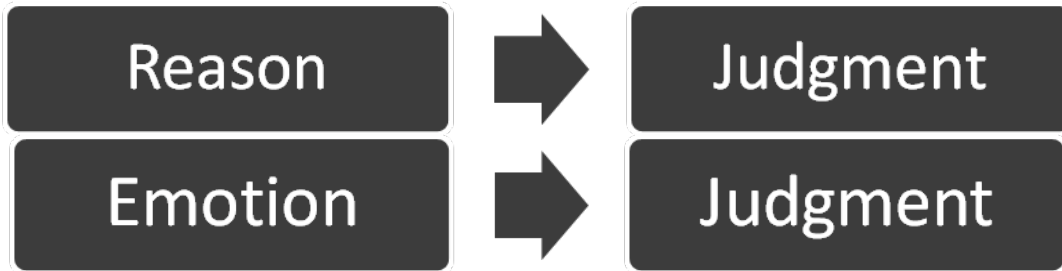


Figure 4: Dual process model of moral judgment

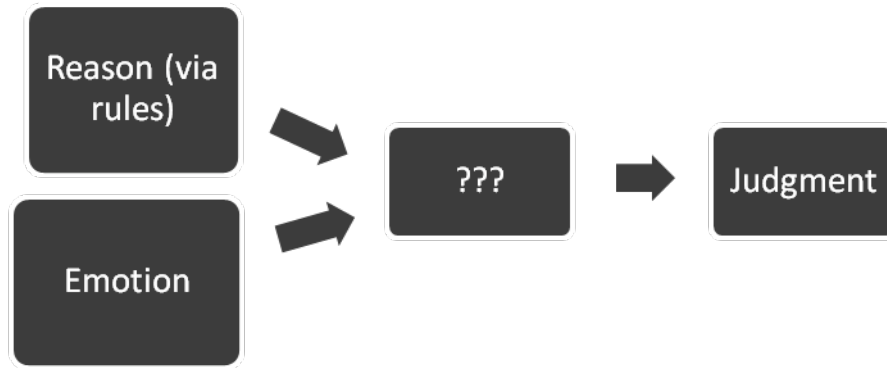


Figure 5: Dual vector model of (non-utilitarian) moral judgment