Critiquing empirical moral psychology

Bryce Huebner, Department of Philosophy, Georgetown University

Commonsense psychology seems to have it that most morally significant decisions arise through practical reasoning and moral delibaeration. No doubt, there are social psychological data that put pressure on reflective views of human agency.[1] However, even on the assumption that non-rational and irrational processes are implicated in many morally significant decisions (cf., Doris, 2002, forthcoming), there still seems to be room for rational deliberation in structuring a morally significant life (cf., Annas, 2005; Kamtekar, 2004). Yet, recent rumblings from the emerging field of empirical moral psychology (hereafter EMP) suggest reason for a wide-ranging and thoroughgoing reevaluation of the role of deliberation and reflection in moral cognition. Research in EMP has attempted to establish 1) that moral intuitions are typically produced by *reflexive* computations that are implicit, fast, and largely automatic (cf., Cushman, Young, & Hauser, 2006; Greene & Haidt, 2002; Haidt, 2001; Hauser, Young, & Cushman, 2007) and, 2) that practical deliberation can only intervene to offer post-hoc justifications of these reflexive moral intuitions (Haidt, 1993). Such justifications may have long-term implications (Haidt & Kesebir, 2010; Paxton & Greene, in press); but even so, moral deliberation and practical reflection are supposed to play a less important role in moral cognition than are the reflexive intuitions that are produced by a distinctively moral system. My goal in this paper is to motivate a skeptical conclusion about the deeply reversionary claims about moral intuitions that are common in EMP by demonstrating that current empirical methods provide too blunt of a tool to offer genuine insight into the computational processes that are responsible for the production of our moral intuitions.

I begin by briefly explaining the thought experimental methodology that has come to prominence in EMP. I argue that because such methods operate on a near-glacial scale, relative to the speed at which reflexive information processing occurs in a human brain, they cannot support the conclusion that moral intuitions are fast, automatic, and largely unconscious. I do not wish to claim that moral intuitions are not produced by computational systems that are fast and automatic. They may well be. Rather, I intend to call attention to the fact that these methods leave room for a significant degree of reflective and deliberative processing in the production of moral intuitions. However, even in light of this skepticism about the methods that are currently employed in EMP, all is not lost. I close by urging the consideration of an alternative theoretical approach to EMP that is more consonant with the use of thought experimental methods. I suggest that the results of moral psychology experiments ought to be reconceived as providing insight into the experience of navigating unfamiliar moral dilemmas, even if they cannot provide insight into the properties of reflexive computational mechanisms.

[1] It is now widely agreed that heuristics, biases, and other reflexive strategies play a critical role in navigating various kinds of social interaction (Bargh & Chartrand, 1999; Wegner, 2002). Many interpersonal evaluations depend on spontaneous, stereotype-based judgments and implicit biases (Nosek, Banaji, & Greenwald, 2002; Valian, 1999); many decisions about who, or what to treat as a morally significant agent depend on non-conscious strategies of dehumanization and anthropomorphization (Haslam, 2006; Haslam, Kashima, Loughnan, Shi, & Suitner, 2008); and, the desire to conform to social norms often plays a crucial role in governing morally relevant behavior (Haney, Banks, & Zimbardo, 1973; Milgram & Sabini, 1978).

1. THE METHODOLOGY OF EMPIRICAL MORAL PSYCHOLOGY

Psychologists have often attempted to study moral cognition empirically; but, in the mid-1990s, psychologists and cognitive scientists began to deploy the explanatory strategies of the cognitive and neurobiological sciences to explain how moral intuitions are implemented in the mind/brain. Unlike investigations that are carried out from the perspective of social and developmental psychology, experiments in EMP do not target behavioral regularities or the phenomenology of social interaction (Wegner & Gilbert, 2000); rather, EMP targets the computational mechanisms and cognitive architecture that is responsible for the production of the explicit intuitions about justice, rights, and welfare that arise as people attempt to resolve moral quandaries (Haidt & Kesebir, 2010, p. 799). Before proceeding with my analysis of this research, two points of clarification are in order: the first concerns the model of explanation that is adopted by proponents of EMP; the second concerns the way in which proponents of EMP construe the moral domain.

First, the decision to approach moral psychology with the tools of the cognitive sciences belies a commitment to explaining moral capacities in terms of the informational mechanisms by which these capacities are implemented. There is a broad consensus within cognitive science that the mind is best understood in computational terms, as an information processing system that is composed of a number of functionally specifiable component systems. Within this theoretical framework, person-level capacities (e.g., the capacity to parse phonemes, the capacity to visually track multiple objects, or the capacity to make deontic judgments about right and wrong) are explained by appeal to the overall organization of subpersonal systems that have the function of processing particular kinds of information. This explanatory project "yields an explanatory payoff when we come to see that something having the kinds of components specified, organized in the way specified, is bound to have the target property" (Cummins, 1983, p. 99). So, as a species of cognitive science, EMP begins by cataloging the moral intuitions that are expressed in morally salient situations; these intuitions are then explained by appeal to the sub-personal states and processes by which they are produced (e.g., by appealing to the representational states and transformational rules that facilitate action segmentation, causal reasoning, mental state ascription, and the assignment of affective valence). In short, reliable patterns in person-level intuitions about what to do and how to live serve as the explananda for EMP, and claims about the computational mechanisms that are responsible for the production of these representations serve as the explanans for these phenomena. As I argue below, there are ongoing empirical debates over which processes are employed in the production of moral intuitions; for my purposes, the most significant of these questions concerns the extent to which the processes responsible for the production of moral intuitions operate reflexively in the face of a morally salient situation.

Second, the assumption that moral computations can be subjected to targeted empirical investigation depends on the claim that there is a psychologically plausible—even if philosophically untenable—distinction between moral and non-moral situations. To establish the existence of this distinction, EMP looks to the results from an older tradition in moral psychology. Over the past 30 years, developmental psychologists have repeatedly demonstrated that young children draw a distinction between moral rules, which prohibit harming or cheating another person, and conventional rules, which govern how one ought to behave in some particular, local context. Children judge that violations of these moral rules are always worse and more punishable than violations of conventional rules; moreover, they judge that moral rules apply universally and cannot be revised by any authority (Turiel, 1983). This distinction between moral and conventional rules arises reliably across cultures (Hollos, Leis, & Turiel, 1986; Nisan, 1987; Nucci, Turiel, & Encarnacion-Gawrych, 1983); and, Larry Nucci and Ernst Turiel (1993) have shown that even Amish children believe that even God cannot modify moral rules against physically harming another person. Furthermore, in a recent study using a battery of adult-appropriate cases (ranging from vehicular and sexual assault, to reckless behavior, and violations of etiquette and social contracts), my colleagues and I have shown that the four features that children employ in distinguishing between moral and conventional transgression (wrongness; punishibility; universality; and authority dependance) also underwrite a moral-conventional distinction in mature cognition (Huebner, Lee, & Hauser, 2010). Building on this ubiquitous distinction between moral and conventional rules, proponents of EMP operationalize moral cognition as the distinct cognitive domain that is concerned with harming or cheating another person (Dwyer, 1999; Nichols, 2004; Turiel, 1983).

With these preliminaries in mind, I turn to the structure of a typical experiment. In most cases, participants are presented with moral scenarios that include ominous threats (e.g., out of control trolleys headed toward innocent people) and agents who can respond to these threats (e.g., by flipping switches or pushing people in front of trolleys to stop them). Intuitions about the permissibility of the actions that are described in these scenarios are collected, statistical analyses are carried out, and patterns are extracted from the data. This much is consistent with standard survey methods used in social psychology: experimenters ask participants to read about some weird things before asking them to convert their thoughts into a number along a specified continuum (Scholl, 2008) Such methods immediately raise many hackles. In these sorts of studies, relatively minor differences in the presentation of a scenario can modulate, attenuate, and even reverse the intuitions that will be elicited. Moreover, the simplifying and idealizing assumptions that are employed in constructing such scenarios can lead to the recruitment of representational resources that would otherwise remain inactive, “thereby evoking responses that may run counter to those evoked by alternative presentations of relevantly similar content” (Gendler, 2007, p. 69). Finally, such scenarios tend to succeed in generating a stable pattern of response, where they do, because they lead people “to represent relevant non-thought experimental content in light of the thought experimental conclusion” (Gendler, 2007, p. 69). But, as a descriptive project, EMP has some resources for defusing the force of these worries.

Experiments in EMP start from empirical hypotheses about where and when contextual factors will modulate, attenuate, or reverse moral intuitions. Moral psychologists thus treat these effects as variables to be explained in terms of moral computations triggered by different scenarios. These experiments attempt to catalog the aspects of our moral psychology that are contextually invariant, as well as those that are contextually dependent. “Where investigations employing different experimental scenarios and subject populations reveal a clear trend in responses, we can begin to have some confidence that we are identifying a deeply and widely shared” set of moral principles or convictions (Doris & Stich, 2006). However, treating patterns of intuitions as evidence for claims about the architecture of the moral mind requires a further theoretical justification.

If EMP is to have the critical bite that it purports to have, moral psychologists cannot rest satisfied with a mere catalog of intuitions evoked by moral dilemmas. Instead they must attempt to uncover the computational mechanisms responsible for the production of our moral intuitions. Only if moral intuitions are always (or almost always) produced by reflexive computational mechanisms can EMP establish the radical claim that deliberation and reflection play only a minimal role in our moral lives. Moral intuitions must be interpreted as evidence for the existence of a *dedicated moral faculty*, or at least a set of interfaced computational systems operating over structured representations of 'actions', 'events', 'agency', and 'outcomes' (cf., Dwyer, Huebner, & Hauser, 2010; Hauser et al., 2007; Jackendoff, 2007a, for friendly criticism; Mikhail, 2007, 2008a; Turiel, 1983) Assuming that this is the case, experimental scenarios can be treated as queries that are presented to the moral faculty in the way that vision scientists present visual illusions and attentional tasks as queries to the visual system. By presenting scenarios that vary systematically in the source of the threat, structure of the causal chain, intentions of an agent, and outcome of an action, moral psychologists hope to *wiggle* each of the components of the moral faculty as a way of seeing how the components are arranged and interfaced.

In relying on scenarios that skirt the borders of fantasy and science fiction, moral psychologists take on the additional explanatory burden of demonstrating that the intuitions that are offered in such experiments can function as genuine queries to a dedicated moral faculty, and that they can facilitate the recovery of 'deep and invariant' computational principles. The plausibility of this methodology, thus, crucially turns on the assumption that the moral intuitions elicited by thought experimental prompts are the output of a reflexive computational system that operates "without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion" (Haidt & Bjorklund, 2008, p. 188). But, while research in EMP critically assumes that the intuitions reported by naïve respondents are produced reflexively, this hypothesis has received relatively little direct confirmation. To bolster the radical claim, various simplified architectural models have been proposed to explain the capacity for making moral judgments. Yet, there have been few points of agreement about the role that is played by various cognitive and affective mechanisms in our moral psychology (cf., Cushman, Greene, & Young, in press; Huebner, Dwyer, & Hauser, 2009). Perhaps the lack of convergence is indicative of the relative immaturity of EMP as a scientific discipline. If so, these disputes will eventually be resolved as more data is collected. However, this lack of convergence, and the seemingly recalcitrant disputes over the role of various computational mechanisms in our moral psychology, could also be indicative of irreconcilable difficulties inherent in the methodology itself. Over the next three sections, I turn to the empirical and theoretical justifications that have been offered for this crucial assumption, suggesting that none of them is sufficient to warrant the conclusion that moral intuitions are produced by a dedicated moral faculty.

2. Moral dumbfounding and implicit computations

A set of experiments carried out by Jon Haidt and his colleagues (Haidt, Bjorklund, & Murphy, n.d.; Haidt & Hersh, 2001) the source of the first argument that is typically advanced to establish that moral intuitions are produced by computational mechanisms operating reflexively, automatically, and outside of conscious awareness.[2] In the most well known experiment, participants were presented with the following scenario:

[2] It is important to note that Haidt does not draw a computational conclusion from this data. His project is situated more firmly within the explanatory paradigm of social psychology than is the research carried out by many other proponents of EMP. I return to this point briefly in the penultimate section of this paper.

> 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 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? (Haidt & Hersh, 2001, p. 814).[3]

After offering their intuition about whether it was okay for the brother and sister to make love, participants were asked to justify their answer. But, although participants reliably offered the predicted moral intuition that incest between siblings is wrong—really, really WRONG—they were rarely able to offer any justification for their intuition that took into account all of the relevant contextual information. Instead, they offered justifications that had been explicitly ruled out in the scenario: they appealed to the possibility of pregnancy, the negative impact on their relationship, and the likely social sanctions when other people found out. Having exhausted these attempts at justifying their intuition, many of the participants eventually asserted that incest is *just wrong,* acknowledging that they had no idea *why* it is wrong. This sort of 'moral dumbfounding' has played a critical role in the development of EMP; it has been taken as evidence that moral intuitions are reflexively produced by a system that is fast and automatic. It has also been taken as evidence for the claim that deliberation and reasoning are typically deployed only to provide post-hoc rationalizations of these reflexive intuitions. But, should we accept this interpretation of the data?

Before adopting this revisionary hypothesis about moral intuitions, we must consider the plausibility that reasoning and deliberation are still playing a role, even if their precise role is not immediately obvious in this experiment. As I noted above, simplifying and idealizing assumptions often lead participants to rely on representational resources that would otherwise remain inactive, thus representing "non-thought experimental content in light of the thought experimental conclusion" (Gendler, 2007, p. 69). This occurs a narrative leads a person to dynamically construct an abstract representation of the possibilities that are entailed by the description of a situation. As these 'mental spaces' are constructed in working memory, the details of the narrative are integrated with previously experienced or imagined situations, drawing on a wide variety of resources to construct a representation of the counterfactual situation; however, the structure of a mental space tends to remain minimal, allowing for rapid revision as the narrative unfolds (Fauconnier & Sweetser, 1996, p. 12). Crucially, the concepts expressed in a narrative

[3] Bjorklund, Haidt, & Murphy (2000) also examined a second narrative, which read as follows:

> *"Jennifer works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research. One night Jennifer is leaving the lab when she sees a body that is going to be discarded the next day. Jennifer was a vegetarian, for moral reasons. She thought it was wrong to kill animals for food. But then, when she saw a body about to be cremated, she thought it was irrational to waste perfectly edible meat. So she cut off a piece of flesh, and took it home and cooked it. The person had died recently of a heart attack, and she cooked the meat thoroughly, so there was no risk of disease. Is there anything wrong with what she did?*

I believe that there is a parallel to the argument that I develop below for this scenario. However, I leave that argument for another time.

often act as filters, constraining the elements that are included in the mental space. So, the structure of the narrative can lead to the recruitment of representational resources that were otherwise inactive, and to the representation of some aspects of the narrative in light of other previously held beliefs. This account of the construction of mental spaces raises two worries about the inference from moral dumbfounding to the claim that deliberative reasoning is only likely to occur in providing post-hoc rationalizations of automatic moral intuitions.

First, the participants in this study may have constructed the relevant mental spaces in ways that would lead them to be *inattentive* to important details of the scenario. Although Haidt and his colleagues presented the narrative in terms that were as clear and concise as possible, participants who are presented with a case of consensual incest may begin to immediately evaluate the case in a way that would prevent them from attending closely to mitigating information. This scenario is relatively long, and although participants listened to the entire scenario before being asked to offer an intuition about the action, they still had time to carry out reflective and deliberative evaluations. Moreover, if participants began to deliberate about why incest is wrong prior to hearing the details of the case, this would increase the likelihood that these consciously considered reasons would be more accessible in the second phase of the task. This would yield a pattern of justifications that looked like moral dumbfounding although conscious and deliberative reasoning strategies had been deployed.

Second, there is a concern about the *incredulity* that participants are likely to feel toward the details of the scenario. While the scenario includes information intended to deflate standard arguments against incest, it is not unreasonable to suppose that participants would generate a mental space that represents the case differently than described. Since the dynamic construction of mental spaces draws elements from both the narratives that evoke them and previously experienced or imagined situations, a representation of the case exactly as it is described would require participants to suppress a wide range of their previously held beliefs in light of the details of the experimental prompt. However, it is strange to suppose that participants would tend to construct a morally relevant space that would update previously held moral beliefs in the relevant respects. Even if they use multiple forms of contraception, it is still possible that Julie will become pregnant. Moreover, even in the best imaginable cases, such acts of incest are unlikely to be consensual in any robust sense; rather, they are likely to trade on power asymmetries that suggest something deeply wrong with the relationship. So, it is even less likely that incest would make a brother and sister feel closer to one another. Finally, even on the assumption that they intend to do so, there is little reason to suppose that such an action would be kept secret. While we cannot be sure of the extent to which participants were likely to be incredulous, the anecdotal evidence that I have collected after presenting this case to numerous students suggests to me that incredulity is likely to be quite prevalent.

Of course, this is only an alternative interpretation of the data; and, it would be unwise to make any general claim about the nature of the moral mind on the basis of a single experiment. In fact, even on the assumption that rational and deliberative processes are likely to play some role in the evaluation of this case, it is also quite likely that participants in this experiment would have an immediate aversive reaction to any description of sibling incest. This opens up a series of questions about the extent to which such intuitions are likely to be revisable on the basis of reflective and deliberative reasoning. Perhaps, in this sort of case, the intuitions are unlikely to be revised. However, the more important question for the purpose of EMP is the extent to which moral intuitions more broadly, as opposed to intuitions grounded on an innate aversion to incest, are likely to depend on immediate responses of this sort (cf., Prinz, 2009). Even if we have reason to posit an innate aversion to incest, something further would be required to demonstrate that moral intuitions more broadly are the output of a dedicated moral faculty.

2.1 The selective accessibility of implicit principles

In an attempt to extend Haidt's data to a more central class of moral intuitions, Fiery Cushman and his colleagues (2006) examined the extent to which an analog of moral dumbfounding arises for intuitions expressed in response to moral dilemmas. Participants were presented with a number of dilemmas modeled on familiar philosophical thought experiments, including a pair of trolley-type cases that differed in the extent to which an innocent person's death was an intended, or merely foreseen consequence of a morally significant action. In the first case, the agent could flip a switch that diverted a boxcar onto a side track, but also caused a person to fall onto that track and be killed. In the second case, flipping the switch caused the person to fall onto the main track where he would be hit by the boxcar, but this would slow the boxcar down before it hit the five people.[4] At the end of the experiment, participants who expressed different intuitions regarding the permissibility of these two actions were asked to explain *why* they had done so. Although participants consistently judged that it was worse to kill one person as a means to saving five others than to kill one as a foreseen side-effect of achieving that same end, they rarely recovered a plausible explanation for this difference in intuitions. Cushman and his colleagues take these data to support the claim that moral intuitions are produced by a dedicated moral faculty, and to show that deliberative capacities are only utilized in offering post-hoc justifications of these intuitions.[5] While participants tend to offer consistent and reliable patterns of response to moral dilemmas, the principles that

[4] The precise wording of these scenarios was as follows (Cushman et al, supplementary materials):

> *"Standing by the railroad tracks, Dennis sees an empty, out-of-control boxcar about to hit five people. Next to Dennis is a lever that can be pulled, sending the boxcar down a side track and away from the five people. But pulling the lever will also lower the railing on a footbridge spanning the side track, causing one person to fall off the footbridge and onto the side track, where he will be hit by the boxcar. If Dennis pulls the lever the boxcar will switch tracks and not hit the five people, and the one person to fall and be hit by the boxcar. If Dennis does not pull the lever the boxcar will continue down the tracks and hit five people, and the one person will remain safe above the side track."*

> *"Standing by the railroad tracks, Evan sees an empty, out-of-control boxcar about to hit five people. Next to Evan is a lever that can be pulled, lowering the railing on a footbridge that spans the main track, and causing one person to fall off the footbridge and onto the main track, where he will be hit by the boxcar. The boxcar will slow down because of the one person, therefore preventing the five from being hit. If Evan pulls the lever the one person will fall and be hit by the boxcar, and therefore the boxcar will slow down and not hit the five people. If Evan does not pull the lever the boxcar will continue down the tracks and hit the five people, and the one person will remain safe above the main track."*

[5] It is worth noting that Cushman and his colleagues (2006) also found that participants were able to recover a plausible justification for their judgments when, for example, the difference between two scenarios was that one required direct and forceful contact with an identifiable individual, while the other did not. As my argument below should make clear, this is important because this is precisely the sort of consideration that would salient in the construction of the mental spaces that would be used to evaluate these cases both individually and comparatively.

justify these intuitions cannot always be recovered for use in conscious reflection and deliberation. So, the argument runs, if we are to explain the consistent and reliable pattern of intuitions, it is necessary to posit a moral faculty that runs reflexive computations over implicit moral principles to yield the relevant moral intuitions.

But, could the alternative hypothesis that depends on the construction of 'mental spaces' explain this range of phenomena? As I noted above, mental spaces are only partial representations of a situation; they are constructed dynamically in working memory to establish enough of a local understanding to allow for the evaluation of a particular situation or action (Fauconnier & Turner, 1998, 2003). Although mental spaces that are utilized frequently can become entrenched in long term memory, the scenarios in this experiment are likely to be unfamiliar and strange enough that they would lead participants to construct a new mental space for evaluating each new scenario. So, we must consider the mental spaces that would be constructed in light of each scenario. In the first case, the outcome of flipping the switch is described *first* as "sending the boxcar down side track, away from five people" and only *second* as lowering a railing and "causing one person to fall off the footbridge and onto the side track, where he will be hit by the boxcar". This narrative is likely to lead participants to construct a mental space that focuses attention first on the number of lives that are saved. In the second case, however, the outcome of flipping the switch is described *first* as lowering a railing and "causing one person to fall off the footbridge and onto the main track, where he will be hit by the boxcar" and only *second* as causing the boxcar to slow down, "preventing the five from being hit". In this case, participants are likely to construct a mental space that focuses attention on the single person who is about to be killed, and only secondarily on the number of lives that will be saved.

Each of these mental spaces would elicit the predicted intuition, and it would do so without recourse to a general principle like the doctrine of double effect. However, in justifying these intuitions, participants would need to carry out a new comparison where the action of flipping the switch, the number of lives at stake, and the outcome of the action are identical across scenarios. These direct mappings would be highlighted in considering the cases side-by-side, and understanding the morally important difference between these scenarios would require examining the intentions of the actors: the intention to kill one person *as a means to* saving five others must be compared against the intention to save five people even though doing so will also lead to the death of one other person. But, these intentions need not be attributed in evaluating the cases independently. So, the fact that participants express moral intuitions conforming to the "doctrine of double effect" cannot, by itself, establish that these intuitions are produced by a computational system that utilizes the doctrine of double effect as an implicit moral principle.

I can only offer this as an alternative hypothesis that cannot be ruled out on the basis of these data. But, the fact that such an interpretation cannot be ruled out is important. It suggests that behavioral data alone are unlikely to discriminate between cognitive architectures that include a dedicated moral faculty that is fast and automatic, and a cognitive architecture that relies on a domain general capacity for dynamically constructing mental spaces on the basis of the information that is presented within a thought experimental scenario. But things are even worse than this might suggest. Indeed, we have good reason to assume that some participants will rely on conscious and reflective processes in arriving at the intuitions expressed in EMP experiments.

2.2. A general concern about behavioral data and claims about mechanisms

Having spent time collecting data for moral psychology experiments, I am skeptical of the claim that participants in experiments such as these are always likely to rest content with their initial intuitions, even if such intuitions are produced immediately and reflexively on the basis of the information that is presented in an experimental scenario. I have often heard participants in my experiments express themselves vocally as they read through a scenario, making their reasoning processes absolutely transparent; and while the details of the deliberation differs from scenario to scenario, it typically runs approximately as follows:

> There is a trolley screaming down the tracks toward five people.–***Oh, that's bad!***– There is a side track onto which the trolley can be diverted.–***Well, I guess that seems better!***–However, there are three people on the side track–***Uh, this seems less good, but it might be ok…this sort of reminds me of that ethics class I took in college.***–They are all members of a terrorist organization.–***Wait a minute; what are they trying to get out of me here? I hope that my name doesn't end up on some list***…

Of course, even if this sort of reasoning occurs in the context of moral psychology experiments, this leaves open an important respect in which people could well be relying on reflexive intuitions that are produced as they read through a particular scenario. That is, each of the relevant aspects of the scenario may trigger a reflexive response that is produced in light of some particularly salient aspect of the scenario.

Indeed, the worries that I have advanced up until this point leave untouched an abductive argument for the claim that moral judgments are likely to be subserved by a reflexive moral faculty. The dominant assumption in social and cognitive psychology is that reflexive processes will dominate cognition unless a person meets three conditions: 1) she has some reason to mistrust her initial judgment; 2) she has some motivation to get the right answer; and, 3) she is aware of a better decision strategy (Talbot, 2010). Assuming that fast and automatic processes tend to be responsible for the production of evaluative judgments, those who hope to defend the claim that moral reasoning is employed in the evaluation of moral scenarios will have to establish that these three conditions are met. Indeed, there is good reason to suppose that precisely the conditions for overriding an immediate and reflexive intuition *would* be met for many of the participants in EMP studies. As psychologists have long noted, participants in psychology experiments often have a strong desire to be good participants. In the context of making moral judgments, participants may be motivated to get the 'right answer' so that they will be seen as having a virtuous moral character. But, the desire to get the right answer is likely to cause participants to reflect and even compare thought experiments to more familiar situations (perhaps in the guise of recent films, sermons, or even ethics classes). Indeed, many people are likely to have been exposed to criticisms of their reflexive moral judgments; frequently this happens in college ethics courses, but it also happens when political pundits and religious leaders (including the flamboyant televangelists who inform us that the common morality is Satan's morality) call the dominant social institutions into question. This sort of criticism plays a prominent role in our moral lives, and it would be surprising if it did not lead people to recognize that there is some reason to mistrust, or at least to double-check, their moral intuitions. This being the case, many participants might think that there is a better decision strategy than simply relying on their intuitions.

I concede that participants in moral psychology experiments may often begin from a set of initial intuitions about the moral status of various situations; and, I concede that these intuitions may even be driven by reflexive mechanisms (perhaps including a reflexive associationist mechanism that compares the current case with previously encountered cases). However, it seems unreasonable to assume, on the basis of the behavioral data that are collected for thought experimental prompts, that *most* of the participants in such studies will rest content with these initial intuitions. Perhaps more importantly, as we move outside of the laboratory to examine more ecologically valid contexts, it seems reasonable to suppose that each of the morally significant judgments that we make is likely to be embedded in social environments that are rich with corrective interpersonal feedback, as well as background assumptions about the social norms that are at play in our highly structured communities. In the absence of this sort of feedback, I contend that participants are likely to reflect upon and re-evaluate the plausibility of their initial judgments. However, recent data from the cognitive neurosciences seem to tell against this claim, again suggesting that moral intuitions are produced by reflexive mechanisms.

3. Neuroscientific data

Adopting an approach that builds on recent neuroscientific data, Joshua Greene and his colleagues (Cushman et al., in press; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) argue that alarm-bell emotions are triggered in those cases where participants are asked to consider the possibility of harming a person in a direct and personal way. In one well known study, Greene and his colleagues (2004) asked participants to read and respond to a series of moral scenarios while they were scanned using functional magnetic resonance imaging (fMRI). Greene and his colleagues found that moral dilemmas that required an agent to utilize direct and personal contact to intervene in a morally significant situation elicited increased activation in areas that are associated with emotional and affective processing; those scenarios that were more straightforwardly grounded on broadly utilitarian considerations of welfare-maximization, by contrast, tended to elicit increased activation in areas that are associated with controlled, cognitive processes such as working memory.

On the basis of these data, Greene and his colleagues argue that two distinct cognitive processes are responsible for implementing the moral mind. On the one hand, a system that is implemented by controlled, cognitive processes plays a critical role in evaluating ‘impersonal’ concerns about welfare. On the other hand, a system that is implemented by emotional and affective processes plays a critical role in the production of the characteristically deontological intuitions that lead us to avoid harming others. This later system is thought to employ alarm-bell like emotions to short-circuit *reflective* moral reasoning, and to increase attention to the prohibition against ‘personal’ harms to another human being. These alarm-bell like emotions have played a critical role in sustaining the norm of not harming others as a means to achieving a greater good, and they achieve this goal by reflexively blurting out a “That's wrong! Don’t do it!" representation any time we consider the acceptability of such harms (Cushman et al., in press; Greene, 2007; Greene & Haidt, 2002; Greene et al., 2004). Of course, this is not to say that these alarm-bell like emotions cannot be overridden with substantial cognitive effort; however, they are “not designed to be negotiable” (Cushman et al., in press). Moreover, the fact that utilitarian reasoning elicits increased activation in areas associated with controlled cognitive processing and working memory also suggests that we should retain a substantial role for moral reasoning where the alarm-bell like emotions are *not* triggered by an emotionally aversive harm. However, Greene and his colleagues (Cushman et al., in press) also contend that concern for human welfare is also likely to have originated in the integration of the affectively valanced intuition that 'harm is bad' with a general strategy of practical reasoning that leads us to want to minimize harm wherever possible; for these sorts of cases, emotional processing would continue to operate "in a currency-like manner that engages controlled reasoning processes".

This dual-process model has much to speak in its favor. The neurological data clearly reveal that distinct neurological circuits are differentially employed in responding to different sorts of dilemmas. Moreover, this model retains a substantial role for controlled cognitive processing; and, it is likely that multiple circuits for evaluating different sorts of situations implement most important facets of human psychology (Carruthers, 2009). Finally, converging evidence for the dual-process model of the moral mind is provided by recent neuropsychological data. Brain Koenigs and his colleagues (2007) examined patients with bilateral lesions to ventromedial prefrontal cortex (VMPFC), which yields flattened affect and diminished empathy. These patients were more likely to judge that it was acceptable to engage in an emotionally aversive harm if doing so will benefit a greater number of people; however, they tended to respond normally to impersonal moral scenarios that do not require engaging in an emotionally aversive harm. But we must tread cautiously in evaluating these neuropsychological data.

Although these data are typically reported as establishing that VMPfC patients make characteristically consequentialist judgments in all of the cases where controls do, these patients also offered characteristically deontological judgments in response to some emotionally aversive harms as well. In fact, every participant in this study agreed that it was impermissible to push your boss off of a roof just because you hate him. And, although some control participants expressed the intuition that it was permissible to push a fat man in front of an oncoming trolley to save five people, a significantly higher proportion of VMPFC patients judged that this action was permissible. What, then, are we to make of the intuitions that are expressed by VMPfC patients (Koenigs et al., 2007)? At the very most, these data establish that damage to VMPFC compromises a sensitivity to emotionally aversive harms that plays *some role* in producing 'normal' intuitions, but only for a narrow range of scenarios where utilitarian concerns conflict with the prohibition against using direct and forceful contact to harm a human being. This leaves open a number of points at which these emotional processes might play a role in producing these intuitions. The affective response could increase attention to the aversive harm while pulling attention away from the utilitarian trade-off; or the aversive harm could act as a filter, constraining the construction of the mental spaces that are used to evaluate these sorts of scenarios (Huebner et al., 2009). Unless there is some reason to prefer one of these options instead of the other, this will give us good reason to remain skeptical about the use of these neuropsychological data in constructing processing models for our moral intuitions.

So, even if there is a kernel of truth to the dual process model, we must ask *where* and *when* the relevant emotional process are likely to play their role. On this point, it is deeply strange to claim that the intuitions that are expressed in moral psychology experiments are implemented by alarm-bell like emotions. I contend that intuitions that are expressed in response to thought experimental scenarios, even in the context of an fMRI experiment, provide too blunt of a tool to provide clear data regarding the architecture of the moral mind. The experiment carried out by Greene and his colleagues

(2004) to establish that some moral intuitions are critically depend on alarm-bell like emotions used a blocked design to facilitate the collection of enough data to examine differences in cortical activity. This is important because participants would quickly work out the sort of question that they should expect to be asked after reading each scenario. Participants were always asked some version of the question "Is it acceptable to F"; so, as soon as they began to read a particular scenario, they could also begin to reason through the morally significant features of the scenario.

Crucially, participants in this experiment were allowed to read each of the scenarios at their own pace, only being excluded from the analysis if they exceeded an upper limit of 46 seconds to read a short scenario. However, in spite of the fact that there was a significant amount of time for scenario relevant processing to occur as people read through the narratives, some participants took a very long time to respond when they were presented with the question "Is it acceptable to F" (as long as 25.2 seconds to respond to 'impersonal' scenarios, mean=4.7 seconds, SD=2.98 seconds; and, as long as 22.8 seconds to respond to 'personal' scenarios, mean=5.2 seconds, SD=3.27 seconds). Indeed, there were some 'personal' dilemmas—the sort that are supposed to be evaluated by a system that employs alarm-bell like emotions to short-circuit *reflective* moral reasoning—for which participants took an average of *eight seconds* to offer their intuitions about whether it was acceptable for an agent to act in some morally relevant way. Assuming that personal scenarios are evaluated by a fast and automatic system, while impersonal scenarios are evaluated using controlled cognition, it is striking that there is very little difference in the average amount of time that it takes for these two systems to yield a moral intuition. It seems that there should be a noticeable difference in reaction times for intuitions produced by the two systems; but, *at best*, there is only a marginally significant difference in the amount of time required to respond to these different sorts of moral scenarios (Greene et al., 2009). With this in mind, I suggest that there is reason for skepticism regarding the claim that one of these systems is fast and automatic, while the other relies on controlled and conscious cognitive processes. But, this is the least of my worries.

Far more troubling is the fact that it takes so long for intuitions of either sort to be produced in a standard EMP experiment. It is implausible to claim that such experiments can establish that some kinds of moral intuitions are driven by *alarm-bell-like* emotions that are deployed in response to emotionally aversive harms; an evolutionarily plausible cognitive architecture is not likely to include alarm bells that will wait for 30-40 seconds to go off! But this worry generalizes to cover any intuition that is produced in the context of an experiment that requires this much cognitive processing. Survey methods operate at a near glacial pace relative to the speed at which information processing occurs in a human brain. Each moral intuition that is rendered in an EMP experiment requires nearly 30 seconds of reading followed by an additional 5-10 seconds to respond to an experimental prompt; so, it is unreasonable to claim that reflexive computations exhaustively explain these moral intuitions unless there is some reason to suppose that any additional processing will be superfluous or irrelevant. As I argued above, this assumption is not trivial. Thus, some alternative explanatory strategy must be employed to establish that the moral intuitions that are expressed in these experiments are the result of reflective computations.

Of course, there is an obvious sense in which a person could not lead a normal moral life if she did not have immediate affective responses to various morally relevant situations in the world around her. As philosophers (Hume, 1978; Nietzsche, 1887/1998;

Smith, 1759; Strawson, 1963) have long noted, many moral judgments are grounded on reactive attitudes (e.g., resentment and guilt). Moreover, as Haidt (personal correspondence) notes, affective responses like “being angered by an insult, or by an act of overt racism toward a third party” play a crucial role in our moral lives. Finally, affectively valenced biases drive many of our morally salient behaviors as we interact with others. Given that gut reactions may play an enormous, consistent, and completely undeniable role in our moral psychology, the story that is provided by Greene and his colleagues (Cushman et al., in press) could very well be right *as a genealogy* of the emergence of concern for human welfare. But, by itself, this is insufficient to establish that some form of 'fast and automatic' processing is responsible for the production of the intuitions that are elicited in the tasks that are typically employed in EMP.

4. The poverty of the moral stimulus

A final strategy for the establishing that moral intuitions are produced reflexively derives from a nativist hypothesis about the evolution of moral judgment. While the evidence for this hypothesis is still somewhat speculative, a few observations speak in its favor (cf., Dwyer et al., 2010). By 39 months of age, children begin to distinguish moral from conventional transgressions (Smetana & Braeges, 1990; Turiel, 1983). At approximately the same time, they start to ascribe the sorts of intentions that justify holding people responsible for their actions (Baldwin, Baird, Saylor, & Clark, 2001; Harris & Núñez, 1996); and they also treat unintended side-effects as intentional only when they result in a negative outcome (Leslie, Knobe, & Cohen, 1996). However, children are exposed to moral stimuli that seem to be insufficient to explain their facility with these moral rules (Dwyer, 1999; Mikhail, 2008b, forthcoming). While they are often corrected for misbehavior, the instruction they receive rarely distinguishes violations of etiquette from moral transgressions, and rarely specifies complex norms regarding intention. Indeed, it seems that young children are more likely to be corrected for violations of convention (e.g., “get your finger out of your nose”, “don't sit on that”, etc.) than for moral transgressions (e.g., “you shouldn't have stolen that”, or “you shouldn't have hit your brother”). Finally, competent adults appear to express moral intuitions that appear to be remarkably convergent across cultural and demographic variation, spontaneous, and confident (Banerjee, Huebner, & Hauser, in press).

Proponents of the ‘Linguistic Analogy’ argue that every typically developing child acquires the capacity to understand and produce a nearly infinite array of moral intuitions, while chimpanzees and family dogs raised in the same environments as human children do not, because moral intuitions are driven by the implicit computations of a distinctive moral faculty (Dwyer et al., 2010). Our tacit knowledge of moral principles is treated as an analog to the implicit grasp of the fundamental principles of grammar that underwrites our capacity to learn languages and make linguistic judgments; and these principles are thought to constrain the range of possible moralities in precisely the same way that implicit linguistic computations constrain the range of possible languages (Hauser, 2008). With this hypothesis in hand, the moral intuitions elicited in psychology experiments can be treated as the analogs of grammaticality judgments.

Since grammaticality judgments are a side effect of being able to use language for communication (Jackendoff, 2007a), judgments about ambiguous and ungrammatical sentences can be used to query the language faculty in the same way that ambiguous figures and visual illusions can be used to query the visual system (Jackendoff, 2007b, p.

7). Linguists attempt to find places where the linguistic processing can be perturbed with ungrammatical sentences, demonstrating the boundaries of a proper functioning linguistic system by showing where the system breaks down. Of course, grammaticality judgments can be distorted by theoretical commitments or overexposure to an ungrammatical sentence type. In such cases, ungrammatical sentences may be misperceived as grammatical. Moreover, linguists can sometimes fail to attend to the control sentences that would help them to explain where linguistic processing has failed, and so fail to explain why a particular sentence type is ambiguous or ungrammatical (Jackendoff, 2007b, p. 7). But, since the target of grammaticality judgments is the breakdown of linguistic processing, linguistic competence can serve as a background for formulating ungrammatical and ambiguous sentences that target the structure of the language faculty and its interfaces with non-linguistic systems. Unfortunately, experiments in moral psychology cannot target processing breakdowns in the same way.

To begin with, since making judgments about justice and morality *is* the point of moral competence, the explicit judgments that are offered in response to moral scenarios can only serve as evidence about how a properly functioning moral system might behave (Jackendoff, 2007a). However, unless there is an incredibly high degree of convergence in moral intuitions, or a story about where and why performance errors are likely to be introduced in moral processing, this data cannot serve directly as evidence for the structure of the moral faculty and its interfaces with non-moral systems. In an EMP experiment, the question is not "where does the moral faculty breakdown"; rather, such experiments assume that the intuitions offered by competent adults can directly target the proper functioning of the moral system. Of course, since it is not obvious what should be done in a moral dilemma (nor even whether there is a correct answer to the question of what ought to be done), convergence in the intuitions evoked by unfamiliar moral dilemmas can provide support for the claim that implicit moral principles act as constraints on moral intuitions. Indeed, this hypothesis could even allow for conscious reflection to enter into the evaluation of a moral scenario so long as such reflection is also constrained by the relevant kinds of implicit computational principles (Huebner, Hauser, & Pettit, in press). But, are moral intuitions really convergent enough to warrant such a conclusion?

One piece of evidence for the claim that they are comes from studies that elicit responses to moral scenarios using dichotomous measures. Across studies, between 80% and 90% of participants express the intuition that it is permissible to divert a trolley onto a side track, killing one person but saving five others; other scenarios that have focused on utilitarian trade-offs have yielded similar results (Greene et al., 2004; Huebner et al., in press). But, while this might provide a promising inroad to searching for convergence in moral intuitions, there is a plausible domain general explanation of such intuitions grounded on reasoning about welfare suggested by Greene and his colleagues (Cushman et al., in press). This hypothesis seems even more promising in light of the fact that moral dilemmas that stray beyond straightforward utilitarian trade-offs show far less convergence *across participants*. For example, 54% of participants express the intuition that it is acceptable for a person to smother a crying baby as a means of saving herself and the other people who are hiding with her in a basement (Greene et al., 2004); similarly, 43% of participants express the intuition that it is acceptable for a person to kill her oldest son in order to pacify an angry tribal leader, where doing so will save her husband and her other three children. But, it is difficult to know how we should we interpret this divergence in intuitions. Perhaps it provides evidence for a dimorphism in the moral faculty. But, if so, it is unclear precisely what this would mean from the perspective of the evolutionary hypothesis. To my knowledge, no one has offered a plausible story for why we should predict such a dimorphism, especially since it does not seem to track any theoretically interesting demographic variation (Banerjee et al., in press). So, perhaps these data suggest instead that moral dilemmas are genuinely ambiguous. Or, perhaps these data suggest only that a performance error is likely to be introduced by the fact that participants have a hard time converting their reflexive moral intuitions into one of the options with which they have been provided. Of course, there may be a plausible story to be told about the pattern of intuitions that is evoked by any particular scenario. However, in light of the fact that experiments using dichotomous measures do not tend to evoke between-participant convergence, the appeal to convergent intuitions must look elsewhere for support.

Other experiments employ scaled responses to detect reliable differences between the intuitions that are evoked by different scenarios. One common measure utilizes a 7-point Likert scale (1=forbidden; 4=permissible; 7=obligatory) to elicit intuitions about the permissibility of some action; and, the intuitions elicited by various scenarios are then analyzed using null-hypothesis significance tests (e.g., ANOVA) to demonstrate that differences in distributions are unlikely to have emerged by chance. Significant results in these experiments are taken to provide defeasible evidence that moral intuitions are sensitive to *some feature* of a scenario; and, by cataloging the intuitions elicited by various scenarios, researchers hope to discover *the precise features* of scenarios to which moral intuitions are sensitive. It is commonly assumed that rejecting the null hypothesis demonstrates that there is a clear and theoretically interesting sensitivity to some difference between scenarios. But, it is important to consider the shape of the data before accepting such a conclusion.

Across experiments, moral dilemmas elicit distributions of responses that 1) center on the ‘permissible’ ratings (between 3 and 5 on the scale above), and that 2) exhibit relatively wide standard deviations (between 1.5 and 2).[6] So, although different moral scenarios elicit significantly different distributions of responses, the difference between the mean responses is unlikely to be the most theoretically interesting aspect of these data. After all, considering only the mean response seems to suggest that participants in moral psychology experiments see every action that is described in a moral dilemma as relatively permissible (though some actions are slightly less permissible than others). However, the wide standard deviations in moral psychology experiments suggest that the responses to various dilemmas tend to be spread out across the scale. Without precise data on distributions, it is hard to know how to interpret these standard deviations. But two possibilities readily suggest themselves: the distribution of responses could be multi-modal; or, the distribution could be relatively normally distributed, with responses falling within three to four points of the mean (NB: this would mean that approximately 68% of participants offer responses that cover approximately half of the scale!). But, neither of these results offers a particularly compelling reason for supposing that there is a high degree of convergence in the responses to moral dilemmas.

[6] For example, the 30 genuine moral dilemmas used by Cushman et al (2006) each elicited a mean response between 2.61-5.08 (M = 4.21), with standard deviations between 1.54-2.00 (M = 1.79); a subset of 10 of these dilemmas translated into Dutch by Hauser et al (2009) yielded similar mean responses (2.63-4.78, M = 4.04), with similar standard deviations (1.3-1.63, M = 1.48). I exclude the control scenarios from Cushman et al 2006 and, the bystander case from Hauser et al (2009) as it had become familiar by this time and elicited nearly identical responses from every participant (M = 3.72, SD = .18). I return to this latter point below.

If responses are distributed multi-modally, then we once again face the sort of worry that I raised about the use of dichotomous scales. It is difficult to know whether to see this pattern of responses as evidence for a dimorphism in the moral faculty, as indicative of genuine moral ambiguity, or as merely a performance error introduced by the attempt to convert a moral intuition into a number along an unfamiliar scale. None of these options provides clear evidence for the existence of a dedicated moral architecture operating over implicit computational principles. By contrast, if responses are instead distributed normally around the mean, this would indeed suggest some degree of convergence in intuitions about moral dilemmas. However, the extent of this convergence would not be sufficient to license the claim that moral intuitions are produced reflexively in response to moral dilemmas. Indeed, on the assumption that such scenarios lead participants to construct mental spaces within which they can evaluate various counterfactual situations, we would predict just this distribution of responses. Each moral dilemma would provide a set of filters for narrowing the range of additional information that would be included in constructing a mental space; but, among other things, differences in prior experience and differences in presumptions of moral salience could lead participants to construct slightly different working memory representations of the narrative that is described with a particular moral dilemma. This being the case, moral intuitions could be relatively convergent, and an associationist and domain-general system constrained by the features of moral scenarios could drive these patterns of responses.

Based on these concerns, as well my concerns about moral dumbfounding and the selective accessibility of moral principles, I maintain that the intuitions that are elicited in moral psychology experiments are unlikely to provide clear evidence regarding the implicit principles of distinctively moral computations. But, there are further concerns about the claim that these intuitions are driven by an evolved moral architecture. Even if there is a plausible story about the evolution of morality, the relevant systems are unlikely to have evolved to evaluate moral dilemmas. There are plausible stories about the evolution of many of the heuristic and reflexive strategies that are employed in navigating various social situations. Racial stereotypes are likely to recruit innate mechanisms for social categorization (Faucher & Machery, 2009; Kinzler, Dupoux, & Spelke, 2007; Nosek et al., 2002). Some species of non-human primates posses rich psychological strategies for perceiving social hierarchies (Cheney & Seyfarth, 2007), and these capacities are likely to provide the evolutionary foundation for the human desire to conform with social norms. Furthermore, recent experiments have demonstrated that capuchin monkeys (S. Brosnan & de Waal, 2003) and (S. Brosnan, Schiff, & de Waal, 2005) refuse to participate in exchanges where an experimenter provides a smaller reward for doing so than has been provided to a conspecific who is carrying out the same task; and, a similar sort of aversions to inequity has been observed in canines (Range, Horna, Viranyi, & Hubera, 2009). Additionally, recent data suggest an evolutionary history for the capacity to utilize evaluations of both intentions and outcomes in normative reasoning; Chimpanzees show signs of frustration when experiments present and withhold food to tease them, but not when the experimenters are clumsy and drop the food (Call, Hare, Carpenter, & Tomasello, 2004); and, cotton-top tamarins are more willing to cooperate with individuals who intentionally gives them food than with individuals who only deliver food as a byproduct of otherwise selfish actions (Hauser, Chen, Chen, & Chuang, 2003). These behavioral capacities provide an evolutionary foundation for the more complicated normative capacities that underwrite the human ability to live a morally significant life, but they are not yet capacities for making judgments about moral dilemmas. Beyond the capacity for normatively significant behavior, explicit intuitions about moral dilemmas also require capacities for counterfactual and narrative reasoning, capacities which may critically implicate conscious and reflective processes in the production of the intuitions that are expressed in response to moral dilemmas.

5. Toward moral heterophenomenology

Where, then, does EMP stand? If you are like me, there are many cases in which you find yourself *stuck* with moral intuitions that you are unsure how to explain. Moreover, you might even find yourself with some intuitions that it is unclear that you would want to explain or justify even if you could. Critically, this often happens outside of the context of EMP as we read novels and newspapers, watch films and television, and listen to lectures and to people sitting next to us on airplanes. It happens as we walk around the cities in which we live; and, it happens as we find ourselves forced to make decisions about what sorts of lives we would like to live. From the perspective of the cognitive sciences, it is important to understand the computational mechanisms that are responsible for the production of such intuitions, and to understand the extent to which the production of these intuitions can be brought under conscious and reflective control. Even if the vast majority of our moral decisions are driven by reflexive and automatic computations, patterns of moral intuitions expressed in response to thought experimental prompts are do not provide the right sort of data for uncovering the computational mechanisms that are responsible for the production of these snap judgments. So, the resources that are provided by the methodology that is commonly employed in EMP are unlikely to provide the right sort of data for answering the questions for which this methodology was designed.

In a previous paper, my colleagues and I argued that the thought experimental methods that are commonly deployed within EMP are insufficient to distinguish between the various roles that might be played by affective and emotional processes in the production of moral intuitions (Huebner et al., 2009). Noting a series of infelicities in the behavioral, neuroscientific, and evolutionary data that have been offered in support of the claim that emotional and affective mechanisms are functional components of a distinctively moral faculty, we argued that these data fail to discriminate between:

1) Cognitive architectures in which affective mechanisms *modulate* responses to moral dilemmas by *increasing or decreasing the severity* of moral intuitions produced by a distinctively moral system;
2) Cognitive architectures in which affective mechanisms recruit selective attention, acting as filters that constrain the range of details that will be considered in evaluating morally significant scenarios; and,
3) Cognitive architectures in which affective mechanisms are genuinely functional components of the moral system itself.

This being the case, we argued that existing empirical methods were too coarse grained to allow researchers to uncover the precise role that is played by affective and emotional mechanisms in our moral psychology. On these bases, we argued that the increasingly common claim that emotional and affective processes play an integral role in the production of moral intuitions (See Prinz, 2009 for the clearest and most compelling defense of this claim) was not empirically well founded. Yet, we assumed that compelling reason remained for thinking that moral cognition would have to decompose into systems of fast, automatic, and unconscious processes operating over causal and intentional representations.

Since the publication of that paper, I have become increasingly skeptical of the claim that fast, automatic, and unconscious processes dominate the production of moral intuitions in moral psychology experiments. While I retain a broad commitment to thinking that much of social cognition is rapid and automatic, the claim that explicit moral intuitions are produced automatically has not (yet) been vindicated. In the same way that the role of affective processes in our moral psychology cannot be read off of moral psychology experiments, I suggest that current empirical methods provide too coarse grained of a tool to offer clear insight into the computational mechanisms responsible for the production of moral intuitions. In short, because it takes an enormous amount of time to read and respond to moral scenarios, such experiments are unable to pull apart the relative contributions of reflexive and reflective processes in the production of moral intuitions. From the perspective of EMP, it has simply been assumed that the intuitions offered in response to moral dilemmas can provide clear evidence about the architecture of the moral mind. However, these thought experimental methods fail to offer any means of distinguishing between:

1) Cognitive architectures in which moral intuitions are reflexively produced only after an entire narrative has been read and understood;
2) Cognitive architectures in which intuitions are reflexively produced at many points as participants read a moral scenario, then reflectively evaluated and consciously revised as new information is encountered; and,
3) Cognitive architectures in which all of the morally relevant processing is deliberative, but the nature of the thought experimental prompts leads participants to focus on only a narrow range of salient details in constructing a mental representation of the mental space that is described in a particular scenario.

Thus, the fact that relatively stable patterns of responses emerge *across experimental manipulations* cannot provide clear insight into the ways in which moral intuitions are produced, adjusted, and revised in light of the information that is presented in a moral psychology experiment. With this in mind, I suggest that the researchers who are currently working in EMP—myself included—should withhold judgment about the relative contributions of various reflexive and reflective mechanisms in the production of moral intuitions.

Fortunately, even on its strongest interpretation, this killjoy conclusion leaves open a great deal of space for further research in EMP; indeed, it leaves open space for research that relies on the same thought experimental methods of which I have been critical. However, to appreciate the value of this experimental paradigm, we also need to adopt a shift in theoretical perspectives. While I believe that experiments based on philosophical thought experiments are unlikely to tell us much about the properties of reflexive computational mechanisms, they can provide insight into the strategies that people tend to adopt in navigating unfamiliar moral dilemmas. A great deal of research in social psychology has already adopted this theoretical approach, even if only tacitly. As ‘the science of experience’ (Wegner & Gilbert, 2000), social psychology rarely needs to consider questions about the computational mechanisms that are responsible for the emergence of person-level behavior. For example, research carried by Haidt and his colleagues offers a rich range of resources for understanding the experience of moral justification; and, as further data on moral dumbfounding is examined, it may well yield intriguing insights into the import of this experience as it occurs in various domains of social evaluation. Furthermore, the wide array of research that has manipulated irrelevant emotional stimuli (e.g., Schnall, Haidt, Clore, & Jordan, 2008; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005), suggests a coarse grained strategy for examining the extent to which affective responses serve as information *at some point* in the evaluation of morally significant situations.

I contend that when these experimental results are understood properly, EMP can help us to understand the ways in which people experience moral dilemmas, as well as the ways in which they understand the intuitions that these dilemmas evoke. EMP, at least as it is conducted using thought experimental methods, cannot be used to justify claims about the computational states and processes that are responsible for the production of moral intuitions. From the examination of person-level behavior, to the cataloging of cross-cultural patterns of convergence and divergence in the expression of moral intuitions, EMP is a species of heterophenomneology (Dennett, 1978, 1987, 1992), the interpretive method of cataloguing overt speech acts, systematizing them as far as is possible, and then generating an account of how things are likely to hang together from the perspective of commonsense psychology. This means that the task of EMP should be to observe participant's responses to moral scenarios and to take them to be sincere expressions of their beliefs about how the moral phenomena seem to them. In addition, this means that moral psychologists must be prepared to discover that there are unresolved and undismissible disagreements about the way that they world seems to different participants (Dennett, 1978, p. 182); and, although standard demographic variables do not seem to explain the patterns of disagreement in moral psychology experiments, some recent data suggest important differences in the ability of people to resist their first intuitions in socially salient experimental contexts (cf., Frederick, 2005). Perhaps this is the sort of difference that matters for mapping the differences that emerge in the responses that people give to moral psychology experiments of various sorts. I contend that until moral psychologists have a better idea of what is driving differences *between participants*, they are unlikely to have a clear idea of what the patterns of agreement that are evoked by EMP experiments actually demonstrate.

So long as research in EMP does not attempt to go beyond heterophenomenological characterizations of human moral psychology, I have few general criticisms of survey methods. But, experiments designed with this end in view only provide a coarse grained tool for evaluating the structure of our moral psychology; and if we wish to develop a deeper understanding of the computational mechanisms that are responsible for the production of moral intuitions, then alternative experimental and conceptual resources must be developed for uncovering the architecture of the moral mind. I contend that even if we possessed a clear and complete account of how moral dilemmas are evaluated, this would leave much to be desired from the perspective of computational cognitive science. However, there are ways of moving forward.

One promising research program in this regard is suggested by recent work of John Mikhail (Mikhail, forthcoming, in press). Mikhail has attempted to extract systematic patterns from the Model Penal Code, the Restatement of Torts, and many other codified legal systems, including a careful analysis of legal norms regarding homicide in various geographically and culturally unrelated jurisdictions (Mikhail, in press). Although the research is still in its preliminary stages, Mikhail has found codified norms governing the criminalization of homicide across a wide range of cultures (the 204 member states of the United Nations), with these codes almost universally including a 'mental state' element in their definitions of criminal homicide (Mikhail, in press, p. 504). More strikingly, these codes also tend to include justifications and exculpations for killing another person as a result of (1) self-defense, (2) necessity, (3) insanity or mental illness, (4) duress or compulsion, (5) provocation, (6) intoxication, (7) mistake of fact, and (8) mistake of law (Mikhail, in press, p. 504). However, as Mikhail notes, there is also a great deal of heterogeneity in legal codes, suggesting that even if our moral psychology is governed by a set of implicit computational principles, there are likely to be a number of other variables that come into play in constructing and sustaining legal norms. Yet, where there are broad scale commonalities across widely disparate legal systems, this does provide some evidence for the existence of constraints on the underlying principles responsible for the structure of mature moral judgments. By itself, such data will not be able to distinguish between the broadly social constraints on moral judgments, and the broadly innate constraints on moral judgments. However, these data—like the heterophenomenological data that is collected by psychologists—offer nothing more, and nothing less, than a compelling foundation for constructing more robust computational models regarding the architecture of the moral mind. These models can then be tested with more targeted experimental investigations that attempt to examine the cases where our moral psychology breaks down in the way that linguistic processing breaks down in the case of ungrammatical sentences.

Finally, it is likely that research using Transcranial Magnetic Stimulation (TMS), and neuropsychological studies of persons with identifiable cognitive deficits can be used to target explicit hypotheses about the ways in which information processing is likely to occur in a typically functioning human brain. However, if these sorts of experiments are to be useful, they have to be grounded on much clearer ideas about the sorts of things to which moral minds like ours are likely to be sensitive. Given that moral scenarios are a coarse grained tool for investigating the architecture of the moral mind, even this will only provide data confirming *that some process is playing some role in the moral mind*, it cannot provide evidence for where and when that process is playing the role that it plays. If we wish to move beyond this coarse grain of analysis, and acutally target the computational mechanisms that constitute the architecture of the moral mind, then we are going to have to move beyond the use of thought experimental prompts. Unfortunately, questions about the methodology that we should prefer would lead us far beyond the scope of this paper.

6. Conclusion

In the attempt to make questions in moral psychology empirically tractable, cognitive scientists have attempted to target the computational principles that implement reflexive moral judgments in a narrow range of cases. But it would be a big surprise if this were the end of the story for moral cognition! Although certain aspects of our moral psychology operate automatically, generating judgments of right and wrong in the absence of deliberative reflection, this does not rule out the possibility that deliberative processes also play an important role in our moral psychology. Deliberative and conscious reasoning might lead us to adopt reflective moral positions that differ from our intuitive moral judgments; or, they might have little impact on the fast and automatic moral computations that allow us to navigate an ever-changing moral environment. But even if some moral judgments are rendered automatically under appropriate triggering conditions, the key question is whether conscious reasoning can *at least* contravene in a way that allows us to reject intuitive judgments that we do not consciously avow. I think that there is good reason for thinking that it can, at least as long as we remember that moral psychology *writ large* is a highly distributed process that is temporally extended and socially embedded. In the end, moral psychology is going to have to go beyond what's 'in the head' to show how psychological mechanisms and social structures mutually influence each other" (Haidt & Kesebir, 2010, p. 800). Still, while I have expressed serious worries about the methodology that is currently being employed in EMP, I do believe that it is possible to develop more promising methods for distinguishing the reflective and the reflexive processes that are employed in making moral judgments. Indeed, I believe that doing so is one of the most pressing and difficult tasks facing contemporary moral psychologists.

7. Works Cited

Annas, J. (2005). Comments on John Doris' Lack of Character. *Philosophy and Phenomenological Research, 73*, 636-647.

Baldwin, D., Baird, J., Saylor, M., & Clark, M. (2001). Infants parse dynamic action. *Child Development, 72*, 708-717.

Banerjee, K., Huebner, B., & Hauser, M. (in press). Intuitive moral judgments are robust across demographic variation in gender, education, politics, and religion: A large-scale web-based study. *Journal of cognition and culture*.

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

Brosnan, S., & de Waal, F. (2003). Monkeys reject unequal pay. *Nature 425*, 297-299.

Brosnan, S., Schiff, H., & de Waal, F. (2005). Tolerance for inequity may increase with social closeness in chimpanzees. *Proceedings of the Royal Society of London, Series B, 1560*, 253-258.

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). 'Unwilling' versus 'unable': Chimpanzees' understanding of human intentional action. *Developmental Science, 7*, 488–498.

Carruthers, P. (2009). An architecture for dual reasoning. In J. Evans & K. Frankish (Eds.), *Two Minds: dual processes and beyond*. Oxford: Oxford University Press.

Cheney, D., & Seyfarth, R. (2007). *Baboon Metaphysics: The Evolution of a Social Mind*. Chicago: University of Chicago Press.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, Ma.: MIT Press.

Cushman, F., Greene, J., & Young, L. (in press). Our multi-system moral psychology: Towards a consensus view In *Oxford handbook of moral psychology*.

Cushman, F., Young, L., & Hauser, M. (2006). The Role of Reasoning and Intuition in Moral Judgments:Testing three principles of harm *Psychological science, 17*.

Dennett, D. (1978). *Brainstorms*. Cambridge, MA: Bradford Books.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: Bradford Books.

Dennett, D. (1992). *Consciousness explained*. New York: Penguin.

Doris, J. (2002). *Lack of Character*. Cambridge: Cambridge University Press.

Doris, J. (forthcoming). *A natural history of the self*. Oxford: Oxford University Press.

Doris, J., & Stich, S. (2006). Moral psychology: empirical approaches. *Stanford Encyclopedia of Philosophy*, from http://plato.stanford.edu/entries/moral-psych-emp/

Dwyer, S. (1999). Moral Competence. In K. Murasugi & R. Stainton (Eds.), *Philosophy and Linguistics* (pp. 169-190). Boulder, CO: Westvew Press.

Dwyer, S., Huebner, B., & Hauser, M. (2010). The linguistic analogy: motivations, results, and speculations. *Topics in cognitive science., 2*(3), 486-510.

Faucher, L., & Machery, E. (2009). Racism: Against Jorge Garcia's Moral and Psychological Monism. *Philosophy of the Social Sciences, 39*, 41-62.

Fauconnier, G., & Sweetser, E. (1996). *Spaces, Worlds, and Grammar*. Chicago: University of Chicago Press.

Fauconnier, G., & Turner, M. (1998). Conceptual Integration Networks. *Cognitive science, 22*(2), 133-187.

Fauconnier, G., & Turner, M. (2003). *The Way We Think*. New York: Basic Books.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*, 25–42.

Gendler, T. (2007). Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium. *Midwest studies in philosophy, 31*, 68-89.

Greene, J. (2007). Why are VMPFC patients more utilitarian?: A dual-process theory of moral judgment explains. *Trends in cognitive science, 11*, 322-323.

Greene, J., Cushman, F., L., S., Lowenberg, K., Nystrom, L., & Cohen, J. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*, 364-371.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in cognitive science, 6*, 517-523.

Greene, J., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(389-400).

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science, 293*, 2105-2108.

Haidt, J. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of personality and social psychology, 65*, 613-628.

Haidt, J. (2001). The emotional dog and its rational tail. *Psychological Review, 108*, 814-834.

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity.* (Vol. 2, pp. 181-217). Cambridge, MA: MIT Press.

Haidt, J., Bjorklund, F., & Murphy, S. (n.d.). Moral dumbfounding: When intuition finds no reason. University of Virginia.

Haidt, J., & Hersh, M. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31*, 191-221.

Haidt, J., & Kesebir. (2010). Morality. In S. Fiske, D. Gilbert & L. Gardner (Eds.), *Handbook of Social Psychology*. Hoboken, NJ: Wiley.

Haney, C., Banks, W., & Zimbardo, P. (1973). A study of prisoners and guards in a simulated prison. *Naval Research Review, 30*, 4-17.

Harris, P., & Núñez, M. (1996). Children's understanding of permission rules. *Child Developmen, 67*, 1572-1591.

Haslam, N. (2006). Dehumanization: an integrative review. *Personality and Social Psychology, 10*(3), 252-264.

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans and nonhumans in three cultures. *Social Cognition, 26*, 248-258. .

Hauser, M. (2008). *The seeds of humanity*. Paper presented at the The Tanner lectures of human value.

Hauser, M., Chen, M., Chen, F., & Chuang, E. (2003). Give unto others: genetically unrelated cotton-top tamarin monkeys preferentially give food to those who

altruistically give food back. *Proceedings of the Royal Society of London, Series B, 270*, 2363–2370.

Hauser, M., Tonnaer, F., & Cima, M. (2009). When moral intuitions are immune to the law: A case study of euthanasia and the act-omission distinction in the Nederlands. *Journal of Cognition and Culture, 9*, 149-169.

Hauser, M., Young, L., & Cushman, F. (2007). Reviving Rawl's linguistic analogy. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 1: The evolution of morality). Cambridge, MA: MIT Press.

Hollos, M., Leis, P., & Turiel, E. (1986). Social reasoning in Ijo children and adolescents in Nigerian communities. *Journal of Cross-Cultural Psychology, 17*, 352-376

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in cognitive science*.

Huebner, B., Hauser, M., & Pettit, P. (in press). When are intentional, means based harms morally permissible? . *Mind and language*.

Huebner, B., Lee, J., & Hauser, M. (2010). The moral-conventional distinction in mature moral competence. *Journal of cognition and culture, 10*(1-2), 1-26.

Hume, D. (1978). *A treatise of human nature*: Oxford University Press.

Jackendoff, R. (2007a). *Language, Consciousness, Culture: Essays on Mental Structure (Jean Nicod Lectures)*. Cambridge, MA: MIT Press.

Jackendoff, R. (2007b). Linguistics in cognitive science: The state of the art. *The Linguistic Review, 24*, 347-401.

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

Kinzler, K., Dupoux, E., & Spelke, E. (2007). The native language of social cognition. *PNAS, 104*, 12577-12580.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446*, 908-911.

Leslie, A., Knobe, J., & Cohen, A. (1996). intentionally and the side–effect effect: ‘Theory of mind’ and moral judgment.

. *Psychological science, 17*, 421-427.

Mikhail, J. (2007). Universal Moral Grammar. *Trends in cognitive science, 11*, 143-152.

Mikhail, J. (2008a). Moral cognition and Computational theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The Neuroscience of Morality: Emotion, Disease, and Development*. Cambridge, MA: MIT PRESS.

Mikhail, J. (2008b). The poverty of the moral stimulus. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: The evolution of morality*. Cambridge, MA: MIT Press.

Mikhail, J. (forthcoming). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press.

Mikhail, J. (in press). Is the Prohibition of Homicide Universal? Evidence from Comparative Criminal Law. *Brooklyn Law Review, 497*.

Milgram, S., & Sabini, J. (1978). On maintaining urban norms: A field experiment in the subway. . In A. Baum, J. Singer & S. Valins (Eds.), *Advances in environmental psychology* (pp. 31-40). New York: Erlbaum.

Nichols, S. (2004). *Sentimental rules: On the natural foundation of moral judgment*: Oxford University Press.

Nietzsche, F. (1887/1998). *On the Genealogy of Morality: A Polemic* (A. J. Swensen & M. Clark, Trans.). Indianapolis: Hackett Publishing.

Nisan, M. (1987). Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology, 23*, 719-725.

Nosek, B., Banaji, M., & Greenwald, A. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group dynamics, 6*(1), 101-115.

Nucci, L., & Turiel, E. (1993). God's word, religious rules, and their relation to Christian and Jewish children's concepts of morality. *Child Development, 64*, 1475-1491.

Nucci, L., Turiel, E., & Encarnacion-Gawrych, G. (1983). Children's social interactions and social concepts in the Virgin Islands. *Journal of Cross-Cultural Psychology, 14*, 469-487.

Paxton, J., & Greene, J. (in press). Moral Reasoning: Hints and Allegations. *TopICS*.

Prinz, J. (2009). *The emotional construction of morals*. New York: Oxford University Press.

Range, F., Horna, L., Viranyi, Z., & Hubera, L. (2009). The absence of reward induces inequity aversion in dogs. *PNAS, 106*, 340-345.

Schnall, S., Haidt, J., Clore, G., & Jordan, A. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096-1109.

Scholl, B. (2008). *Two kinds of experimental philosophy, and their methodological dangers*. Paper presented at the SPP Workshop on Experimental Philosophy.

Smetana, J., & Braeges, J. (1990). The Development of Toddlers' Moral and Conventional Judgments. *Merrill-Palmer Quarterly, 36*, 329-346.

Smith, A. (1759). *The theory of moral sentiments*: Cambridge University Press.

Strawson, P. (1963). Freedom and Resentment. In J. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility*. Ithaca, NY: Cornell University Press.

Talbot, B. (2010). *The Irrelevance of Folk Intuitions to the "Hard Problem" of Consciousness*. Paper presented at the Second Annual Online Consciousness Conference. from http://consciousnessonline.wordpress.com/2010/02/19/the-irrelevance-of-folk-intuitions-to-the-%E2%80%98hard-problem%E2%80%99-of-consciousness/

Turiel, E. (1983). *The development of social knowledge: morality and convention*. Cambridge: Cambridge University Press.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological science, 17*, 476-477.

Valian, V. (1999). *Why so slow*. Cambridge, MA: MIT press.

Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: MIT press.

Wegner, D., & Gilbert, D. (2000). Social psychology–the science of human experience. In H. Bless & J. Forgas (Eds.), *Subjective experience in social cognition and behavior* (pp. 1-9). Philadelphia: Psychology Press.

Wheatley, T., & Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science, 16*, 780-784.