Do emotions play a constitutive role in moral cognition?

Bryce Huebner, Department of philosophy, Georgetown University

Behavioral experiments have revealed that the presence of an emotion-eliciting stimulus can affect the severity of a person's moral judgments, while imaging experiments have revealed that moral judgments evoke increased activity in brain regions classically associated with emotion, and studies using patient populations have confirmed that damage to these areas has a significant impact on the ability to make moral judgments. To many, these data seem to suggest that emotions may play a robustly *causal* or perhaps even a *constitutive* role in moral cognition (Cushman, Young, & Greene 2010; Greene et al. 2001, 2004; Nichols 2002, 2004; Paxton & Greene 2010; Plakias 2013; Prinz 2007; Strohminger et al. 2011; Valdesolo & DeSteno 2006). But others have noted that the existing data are also consistent with the possibility that emotions operate *outside of* moral cognition, 'gating' off morally significant information, or 'amplifying' the output of distinctively moral computations (Decety & Cacioppo 2012; Huebner, Dwyer, & Huaser 2009; Mikhail 2011; Pizarro, Inbar, & Helion 2011). While it is commonly thought that this debate can be settled by collecting further data, I maintain that the theoretical foundations of moral psychology are themselves to blame for this intractable dispute, and my primary aim in this paper is to make a case for this claim.

My argument for this claim builds up slowly. In the first four sections, I provide a critical review of data that have seemed to support the hypothesis that emotion plays a critical role in moral cognition. In each case, I argue that the existing data cannot rule out plausible alternative hypotheses that do not rely on emotional processing. My arguments in these sections will be largely critical, but they also have a larger purpose. They give me the theoretical and empirical tools to diagnose the source of ongoing disputes about the role of emotion in moral cognition. So in the final section, I argue that folk-taxonomic categories such as 'emotion' and 'judgment' are unlikely to provide the correct vocabulary for studying moral cognition; and I outline an approach to moral cognition that relies on *predictive* and *evaluative* mechanisms, rather than affective and cognitive mechanisms. This paper is not the place to provide a thoroughgoing defense of this approach. But I hope to show why it promises to yield novel insights about the architecture of moral cognition.

1. The role of emotion in moral cognition

Norm transgressions often evoke expressions of contempt, anger, and disgust, which shape our understanding of what we ought to do (Rozin 1997; Rozin et al. 1999); expressions of disappointment and anger nudge us toward mutually beneficial forms of cooperation (Baumard et al. 2013; Cova et al. 2013); and, feelings of disgust and fear sustain dehumanizing assumptions about who deserves moral consideration (Haslam 2006; Pizarro et al 2006; Sherman & Haidt 2011). Against this backdrop, our moral convictions must often be bolstered by emotional reactions if we are to overcome the debilitating effects of real and imagined sanctions (Milgram & Sabini 1978; Skitka & Wisneski 2011). It would be a mistake to deny that emotions play an important role in every ordinary moral life, especially where interpersonal coordination is concerned. But how, precisely, are emotions and moral judgments connected?

Several recent studies have revealed that changes in emotion reliably cause changes in moral judgment (Haidt 2001; Horberg et al. 2009; Nichols 2004; Wheatley & Haidt 2005): people make harsher moral judgments when they are seated at a dirty desk near a greasy pizza box, and when they are exposed to the noxious odor of fart spray (Schnall et al. 2008a); judgments about personal harm are made more severe by listening to aggressive Japanese noise music and drinking bitter tasting beverages (Eskine et al. 2011; Seidel & Prinz 2013); and people think it is more permissible to push a fat man in front of a moving train after watching funny film clips or listening to comedians (Strohminger et al 2011; Valdesolo & DeSteno 2006). Such data suggest that emotions play a critical role in the production of moral judgments. But these data are simply too coarse-grained to identify precisely what role emotion plays in the production of moral judgments. These data are consistent with the hypothesis that emotions are causally or constitutively involved in the production of moral judgments; but they are also consistent with the hypothesis that moral judgments are rendered by a dedicated cognitive system, and then

triangulated against other active and potentially salient goals (Huang & Bargh in press); and they are consistent with the hypothesis that emotions merely direct our attention to the morally salient features of a situation (Bless & Fiedler 2006; Huebner, Dwyer, & Hauser 2009, 3).

The harsher moral judgments revealed by negative affect induction paradigms may suggest that people calibrate their existing moral judgments against other potentially relevant goals. For example, disgust triggers the goal of avoiding or rejecting an offensive stimulus (Han, Lerner, & Zeckhauser 2012), and calibrating a moral judgment against this goal may facilitate the negative processing of morally salient features. Strikingly, this effect is moderated when people are allowed to wash their hands after viewing a disgusting film clip. While such hand washing may reduce disgust (Schnall et al. 2008b, 1221), it also primes thoughts about cleanliness, and enhance the accessibility of purity concepts (Schnall et al. 2008b, Experiment 1; Holland et al. 2005). So perhaps people are calibrating their moral judgments against many other goals, including the desires to avoid or reject offensive stimuli, as well as the goal of promoting cleanliness.[1] Similarly, anger induction may trigger increased vigilance, or lead to the production of antisocial and aggressive goals, impacting judgments about personal harm by modulating existing moral judgments. When people view an abusive interaction, they report feeling outraged, and they make harsher and more punitive judgments about subsequent unrelated cases (Goldberg et al 1999). But perceiving this violent interaction is likely to evoke a reflexive appraisal of the situation, thereby triggering the production of at least two corresponding goal-representations: *someone must be held responsible for this wrongdoing* and *the injustice must be redressed*. Where people calibrate their moral judgments against these potentially relevant and currently active goal-representation, they may offer harsher and more punitive judgments related to *personal harm*. Indeed, *the belief* that justice has been served is able to deactivate *the goal* of seeking retribution, moderating this effect (Goldberg et al. 1999, 783).

A similar account is available for mirth induction, which is likely to broaden attention, increase the need for cognition, and widen of the range of factors considered in evaluating a situation (Fredricksen & Branigan 2005, 315). Positive affect induction often leads people to consider a broader range of contextual factors, and to be more flexible in trying different strategies for solving problems that antecedently interest them (Isen 2001); it can also moderate anchoring effects, and increase the willingness to revise judgments (Asby, Isen, & Turken 1999; Estrada, Isen, & Young 1997; Isen, Rosenzweig, & Young 1991). Importantly, mirth induction also increases dopamine levels in the prefrontal cortex [PFC] and the anterior cingulate [ACC], which are likely to facilitate working memory, enhance executive attention, and modulate the selection of cognitive perspectives (Ashby et al. 1999, 2002). Of course, mirth induction can also lead to pleasant feelings, which are sustained by the opioid circuits in the forebrain (Berridge 1996, Berridge & Kringelbach in press). But the modulation of attention and working memory would be sufficient to cause changes in moral judgment. People with a high 'need for cognition' focus on considerations of aggregate welfare when responding to moral dilemmas (Bartels 2008), and if mirth induction triggers a goal like 'broaden-and-build', this could yield an increased role for considerations of aggregate welfare.[2]

I contend that it is not at all obvious that the best way to approach these priming data is by appealing to the distinction between emotion and cognition. But the existing debates in moral psychology have often missed the importance of alternative explanations because the research questions tend to be framed in terms that

[1] Alternatively, the desire to avoid or reject disgusting stimuli may fade after distraction, and Schall et al. (2008b) don't control for the time it takes people to wash their hands. Alternatively, as an anonymous referee notes, an embodied cognition approach may better accommodate these data. But as Rupert (forthcoming) argues, the most plausible approach to embodied cognition is likely to rely on multiple parallel processes, yielding a view that is roughly in accord with the view that I articulate in the main text.

[2] Strohminger and colleagues (2011) also found that people who listened to selections from *Chicken soup for the soul* were less likely to think it permissible to pursue aggregate welfare in moral dilemmas. I am unsure what goal listening to these clips would trigger, but it is conceivable that it would increase the salience of some values, which may lead to privileging norms against battery. The effects of mood induction are situationally variable, and Hunsinger, Isabell, and Clore (2012) propose that positive affect is more likely to facilitate situationally dominant attitudes. However, the fact that the number of lives saved is presented as a situationally relevant should lead people to be (slightly) more accepting of an option that takes this into account, which is consistent with the suggestion I develop in the main text.

highlight this distinction. Unfortunately, the structure of the computations responsible for variations in behavior is often opaque from the standpoint of behavioral research. So even where emotional mechanisms increase the salience of some considerations, or decrease the salience of others, this cannot establish that it is the emotions that are playing a causal or constitutive role in the production of moral judgments—unless, that is, there is some way to rule out the possibility that emotion-processing occurs upstream, downstream, or even in parallel to moral processing. But once we reject the hypothesis that the moral judgments must be the result of information flowing linearly through a single feed-forward system, we are forced to entertain difficult empirical questions about the computational roles that are played by various competing influences, operating in parallel, and collectively affecting behavior. Moral judgments may arise through "the integration of all the currently active representations, including the primed concept, knowledge about the self and the target person of the action, and other aspects of the situation" (Schröder & Thagard 2013). But given that variations in behavior are often the result of integrating multiple competing goal-influences, the strength of various desires and associations, as well as the inhibitory and excitatory relations between representational systems, become crucial variables in interpreting behavioral experiments (Huang & Bargh in press; Rupert in press; Huebner & Rupert in press). Seeing things in this way helps to explain why priming paradigms are often fragile, and it helps to explain why subtle variations in priming stimuli often affect the success of a manipulation.[3]

Of course, it is common knowledge that multiple systems are involved in the production of moral judgments. *Pace* my suggestions thus far, Josh Greene maintains that this fact is evidence for the moderate claim that some moral judgments *depend* on affective mechanisms. Greene and his colleagues (2004) found that dilemmas in which aggregate welfare conflicts with direct physical harm yield increased activity in the ACC, which is consistent with a conflict between cognitive systems that evaluate considerations of aggregate welfare, and affective systems that produce the aversion to direct intentional harm. I agree that the existence of a system that tracks aggregate welfare gains support from the fact that a numerical search task selectively affects response times for judgments based on such considerations (Greene et al. 2008); it also gains support from the fact that people with a high 'need for cognition' are more likely to focus on considerations of aggregate welfare in high-conflict moral dilemmas (Bartels, 2008). Finally, time pressure also decreases the proportion of judgments based on aggregate welfare, and people are more willing to revise their judgments to favor welfare considerations when they reconsider their original judgments (Cummins & Cummins 2012; Suter & Hertwig 2011). In short, the fact that cognitive load and 'need for cognition' selectively affect judgments based on aggregate welfare strongly suggests that accepting such an option is effortful, and requires overriding the output of another system.

Building on the affect induction experiments discussed above, such data may seem to suggest that emotional mechanisms are constitutively involved in judgments about direct personal harm at the very least. But as I noted above, these data are too coarse grained to establish *what systems* are involved in the relevant competitions. As John Mikhail (2011, forthcoming) notes, high-conflict dilemmas tend to conflate affective valence with battery; so the data in these experiments are consistent with the hypothesis that moral judgments rely on two rule-based systems: one that evaluates aggregate welfare, and one that reflexively tracks the intricate representational structure of norms against battery. Further support for this hypothesis comes from the fact that increased stress leads to a higher proportion of judgments based on the aversion to harm, and it does so without having any affect on positive or negative emotions (Starcke et al 2012; Youssef et al. 2012). Even minor stress can modulate the levels of dopamine in the PFC (Deutch & Roth 1990),

[3] This fact was driven home to me during a failed attempt to replicate the 'fart spray' study. I was working in a public space, and could not control for idiosyncrasies of personal history, or for stray assumptions about the experimental situation. But one of my participants asked me, in a thick Dorchester accent, "You know why it smells funny around here? It's usually nice, but it kinda smells like garbage". The experiment was originally run on the campus of Stanford University, and I was running it on the Boston Common, in an urban area where noxious odors are common. My participants may have reflexively downgraded the salience of the desire to avoid and reject offensive stimuli, which would moderate the effect of the induction. Of course, it's hard to know whether my failure derived from differences in the experimental context, difference in participants, or whether it was an artifact of the complex nature of moral judgment. But this is precisely the problem I want to call attention to.

and such a shift cold impact decision-making by biasing people toward reflexive patterns of reasoning (Miller & Cohen 2001; Crockett 2013). So, perhaps the stress triggered by increased task demands and time pressure could affect moral judgments by decreasing the resources allocated to considerations of aggregate welfare, and increasing attention to the aversive and negative features of a high-conflict dilemma (Chajut & Algom 2003).[4]

It may be assumed that this suggestion is inconsistent with well-known behavioral data, which seem to speak decisively in favor of the hypothesis that emotions sometimes play a causal or constitutive role in the production of moral judgments. As the data are typically reported, people who are hypnotized to feel disgust when they hear neutral words (take; often) offer harsher moral judgments for some scenarios that include these words (Wheatley & Haidt 2005). There are compelling reasons to think that disgust could draw attention to the morally salient features of these scenarios (Huebner, Dwyer, & Hauser 2009; May forthcoming; Prinz 2006, 31), and that the desire to reject disgusting stimuli could amplify existing moral aversions, yielding harsher responses. But it is also worth noting that these data are more difficult to interpret than the received wisdom suggests. Participants placed a hash mark along an unbroken line anchored at *Not at all morally wrong* and *Extremely morally wrong*; these responses were converted to numerical values between 1-100; and they were analyzed using seven t-tests (one for each scenario, one for the pooled mean). These analyses revealed a significant effect of disgust for two of the six scenarios (one about second cousins with a sexual relationship; one about a congressman who accepted bribes), and for the pooled mean.[5]

The decision to use a series of t-tests (including one for the pooled mean), rather than an ANOVA and post-hoc tests, increases the likelihood of discovering a significant effect even where there isn't one. Since these tests were not independent, it would be reasonable to use a Bonferroni corrected-α = .007 in evaluating the significance of these tests.[6] But, this would leave just one significant effect: second-cousin incest. I maintain that it is not particularly surprising that heightened disgust increases the salience of incest aversion, especially where a marginal case of 'incest' is concerned. More interestingly, this case of incest aversion is the only case where disgust causes judgments to cross the mid-line of the scale (M=43.29; M=67.63), suggesting a shift from treating this action as not particularly wrong, to treating it as kinda-sorta wrong. But even if the bribery case and the pooled mean are legitimate effects, the descriptive data reveal that they were only shifted slightly in the predicted direction, amplifying what was already a *moral* judgment. So even the most charitable interpretation of these data only justifies the hypothesis that disgust can modulate moral judgments, and can make them *a bit* harsher (May Forthcoming; Pizarro, Inbar, & Helion 2011).

In light of the arguments that I have offered in this long section, I maintain that the existing behavioral data cannot, on their own, sustain any claim about the role of emotional and cognitive processes in moral cognition. While there are many debates in moral psychology about the role of emotion in these experiments, framing research questions in this way may obscure alternative possibilities that do not rely on emotional processing at all. But of course, most

[4] A distraction task that requires participants to rate the pleasantness of a house decreases the likelihood of endorsing actions that use personal force or cause intentional harm, while counting the number of windows in that house has no noticeable effect on these judgments. Cummins & Cummins (2012) see this as evidence that emotion plays a role in moral cognition, but this difference is more plausibly explained by appeal to the enhanced goal of reporting reflexive evaluations after being told to make a reflexive judgment about a house. Since they did not use a manipulation check to confirm the presence of affect, these data cannot support the hypothesis that emotions are causally or constitutively implicated in the production of moral judgments. (Thanks to Nina Strohminger for helpful discussion of this experiment).

[5] In a second experiment they found no statistically significant effects for the 6 scenarios, and an effect for the pooled mean. I confine my discussion to the first experiment, though the extension to the second should be clear enough.

[6] The probability of finding a statistically significant effect as a matter of chance for a single test at α=.05 is approximately 1/20 in this case. Without adjusting for the number of levels in a factor, the probability of finding a statistically significant effect merely as a matter of chance for seven tests at α=.05 is $(1-(0.95)^7)$, or approximately a 30% chance of a false discovery. That said, the Bonferroni method is a highly conservative correction, and it can obscure real effects. I offer this correction as nothing more than a reminder that the data *as analyzed* cannot provide unambiguous support for the claim that emotion plays a causal role in the production of moral judgments.

researchers working in moral psychology know that more is required to establish that emotions are causally or constitutively involved in the production of moral judgments. This is why many of them have turned to the data collected in cognitive neuroscience and neuropsychology, which appear to offer more robust data about the causal and temporal role that is played by emotional mechanism in the production of moral psychology.

2. Toward a more mechanistic hypothesis

There is broad consensus, however, that such data provide compelling support for a moderate constitution hypothesis. As Greene (2009) puts the point, a dual-process theory "which emphasizes both emotional intuition and controlled cognition, is supported by multiple fMRI studies using different behavioral paradigms, multiple behavioral studies of neurological patients, and a variety of behavioral studies using both experimental manipulations and individual difference measures". In general, moral judgments recruit a network of mechanisms in mPFC, precuneus, and posterior superior temporal sulcus [pSTS]—especially at the right temporal parietal junction (Greene 2009; Greene et al 2001, 2004; Young & Saxe 2008). These areas are reliably correlated with mentalizing tasks (Saxe et al 2004), and they are often thought to be part of the default network that underlies perspective taking, planning, episodic remembering, and counterfactual thinking (Buckner, Andrews-Hanna, & Schacter 2008). But regions associated with emotion also tend to be active in tasks where moral judgments are made. Judgments about direct personal harm are correlated with activity in the vmPFC, insula, and amygdala (Greene et al 2001, 2004; Harenski & Hamann 2006, Heekeren et al. 2003; Moll et al 2001, 2003, 2005), while high-conflict dilemmas selectively increase activity in the ACC (Greene et al. 2004).[7]

Unfortunately, studies examining *moral judgments* move at a glacial pace relative to the speed of neural processing. Reading a moral dilemma, and coming to a decision about its permissibility, may take a long time. The time scale of these experiments should trigger numerous cognitive processes, which will interact at various time-scales, and without any clear behavioral evidence about when particular processes are being carried out; this means that the "associated brain activity is not event-related in the classical sense of the term, as its duration and profile are independent of the physical and statistical characteristics of the stimulus eliciting a given reasoning episode" (Papo 2013). So the problem with such data is that it is simply unclear when participants make their judgments. Many studies assume that judgments will only arise in a brief window surrounding the actual response. But I maintain that it is unclear whether people wait until all the information is in, or instead begin to evaluate morally salient information much sooner, making initial judgments and revising them as more data comes in (Huebner 2011).

Consider the familiar experiment by Greene and his colleagues (2004), which relied on a blocked design where participants read 60 practical dilemmas and responded to roughly the same question for each one ("Would you Φ"). Participants were allowed up to 46s to read the dilemmas, and some waited as long as 25s afterwards to provide their response.[8] But the blocked design would allow participants to form expectations about the upcoming questions, and to begin evaluating morally salient information long before offering a response; if participants make moral judgments all along, the activity in emotional circuits could either be a component of moral processing, or it could indicate a response to moral judgments that have already been rendered.

Such worries are well known, and studies using high-density event-related potentials have been used to examine the temporal organization of these mechanisms. One such study used a series of images to present a moral situation, revealing an initial burst of activity over right pSTS (62 ms after stimulus onset), followed by a cascade of activity in the amygdala (122-180 ms), and finally activity in the vmPFC (182ms). One way of interpreting these data is to see them as demonstrating that information about intentional harm is processed first, yielding a moral judgment, which is passed along to affective

[7] Some moral judgment tasks also reveal activity in other parts of the orbitofrontal cortex [OFC}, but fMRI introduces distortions near the sinuses, making it difficult to get clear data for this area; that said, it would not be surprising if OFC were active in the production of moral judgments (Cf. Landreth 2008)

[8] Impersonal up to 25.2 seconds, M=4.7, SD=2.98; personal up to 22.8 seconds, M=5.2, SD=3.27

circuits that act "as a gain antecedent to moral judgment by alerting the individual of the moral salience of a situation" (Decety & Cacioppo 2012, 10pdf).[9] But it is also plausible to think that moral *judgments* require integrating intention information with affective information, and that this cascade of activity supports the hypothesis that moral judgments causally and constitutively depend on emotional mechanisms. Nothing in the data requires accepting one hypothesis over the other; even worse, the only way to address this issue would be to develop a non-question-begging account of how long it takes *the brain* to arrive at a moral judgment, as well as a computational model that could distinguish between gain-circuits that amplify the strength of a moral judgment, and distinctively moral-circuits. No one has addressed either issue, and to my mind it is not clear how one would do so. These data cannot distinguish between architectures in which emotional mechanisms are constitutively involved in moral processing, and architectures in which emotional mechanism simply modulate the output of distinctively moral systems; as such, they provide further grounds for skepticism about the tendency to frame experiments in ways that the role of emotional processing in moral cognition.

3. Psychopathy and moral judgments

Of course, my pessimism about this research is likely to be met with apprehension, for the causal role of mechanisms in the vmPFC and amygdala should be discoverable using neuropsychological tasks. One way of examining the role of these mechanisms would be to examine psychopathic individuals, who have deficits in emotional capacities and display structural and processing deficits in both the vmPFC and amygdala. Strikingly, they also classify moral transgressions, and make judgments about morally salient scenes and moral dilemmas that are indistinguishable from controls (Aharoni, Sinnott-Armstrong, & Kiehl 2012; Glenn et al 2009a; Harenski et al. 2010). They seem to know the difference between right and wrong, but they don't seem to care (Cima, Tonnaer, & Hauser 2010). This is consistent with the hypothesis that the mechanisms in the vmPFC and amygdala play a modulatory role in moral judgment, being activated only after moral judgments are rendered. But before accepting this hypothesis, it will pay to consider what these mechanisms are doing.

Processing in the amygdala has long been seen as essential to fear-conditioning and Pavlovian learning. But recent computational models suggest that it processes a wide variety of unpredictable and ambiguous stimuli, coding for evaluations of biological salience and for the need to gather information (Adolphs 2011). Increased amygdala activity strongly correlates with ambiguity aversion (Hsu et al. 2005), and amygdala lesions both reduce loss aversion and increase risk-taking and social curiosity (Demartino, Camerer, & Adolphs 2010). These lesions also inhibit the production of *expected* reward signals in the mPFC, modulating choice behavior in spite of the fact that the capacity to monitor rewarding and punishing feedback remains intact (Hampton et al 2007). The amygdala-mPFC circuit that is compromised in psychopathy thus appears to play an significant role in computing expected reward values; and a diminished capacity to compute such values might explain why psychopathic individuals act immorally while knowing the difference between right and wrong. They may process many kinds of morally salient information, but fail to compute the expected reward value of acting on the basis of these moral representations.

An increase in dlPFC activity is also observed when psychopaths make moral judgments, which suggests that they may be using an alternative decision-making strategy (Glenn et al 2009b). Building on data from fMRI, EEG, and TMS experiments, D'ardenne and her colleagues (2012) propose that the dlPFC plays a critical role in updating working memory representations in light of contextually salient information. This proposal helps to make sense of a number of data in decision-science, and it points the way forward to a better understanding of the computational tasks that are carried out in the production of a moral judgment.

Increased dlPFC activity is observed when people resist the urge to make unfair offers in the ultimatum game (Spitzer et al 2007), while and rTMS over the right dlPFC yields risky decision-making and increases the likelihood of accepting

[9] Another recent study using Lateralized Readiness Potentials and a Go/No-Go task suggests that moral information is processed prior to, and independently of, disgust information. (Yang et al. 2013).

such unfair offers (Knoch et al. 2006a, b).[10] There is emerging consensus that accepting unfair offers in these games requires effortful goal-maintenance and the suppression of the impulse to punish unfair offers (Koenigs & Tranel 2007; Kirk et al 2011; Sanfey et al 2003). Here, the goal of acting cooperatively must be held in working memory to inhibit the impulsive desire to punish. Increased dlPFC activity is also observed when people make judgments about criminal responsibility or the appropriateness of punishment, suggesting that initial impulses may be triangulated against the contextually relevant goal of making fair and impartial decisions (Buckholtz et al 2008). Finally, disrupting the right dlPFC *increases* welfare-based judgments for some types of moral dilemmas—specifically for third-personal moral judgments, but not for high-conflict subjective dilemmas (Tassy et al 2012). Greene's moderate approach would predict the effect for high-conflict dilemmas, but the effect on third-personal moral judgments would seem surprising if these judgments depended solely on controlled cognition; but if the familiar pattern of responses requires up-regulating the goal of not causing direct physical harm, as would be predicted by Mikhail's (2011) rule-based model of moral cognition, this pattern of behavior becomes predictable.

In each case, activity in the dlPFC is correlated with the inhibition of impulsive desires, which can also be conceptualized as an up-regulating of goals in light of current task demands; and this is consistent with the hypothesis that mechanisms in the PFC organize ongoing behavior in light of goal-representations stored in working memory (D'ardenne et al 2012; Miller & Cohen 2001; Knoch et al. 2008). This suggests a potential correlation between the increased dlPFC activity when psychopaths make moral judgments and an up-regulation of the task-relevant goal of pleasing the experimenter. As with typically functioning participants who make judgment about criminal responsibility (Buckholtz et al 2008), psychopathic individuals may simply be triangulating their initial impulses against the contextually relevant goal of making the right judgment. Unfortunately, behavioral measures are unlikely to be able to confirm this hypothesis, given the nature of the experimental population. Regardless of how this turns out, however, it is not clear that it is emotional deficits that explain the patterns of judgments offered by psychopathic individuals. Again, we find a case where existing debates in moral psychology are likely to miss the importance of alternative explanations because they have framed their research questions in terms that highlight the distinction between emotion and cognition.

[10] In the ultimatum game, one player proposes a way to divide some money, while a second player decides whether to accept this proposal. If the second player rejects the offer, neither gets anything. People routinely offer fair splits (50/50), and routinely reject offers that they are too unfair (typically around 20-30%). Cathodal tDCS over right dlPFC also reduces the likelihood of punishing someone who makes an unfair offer (Knoch et al. 2008) by inhibiting neural excitability, and decreasing the likelihood that this population of neurons will fire in response to the relevant stimulus.

4. Looking to the medial PFC

What, then, are we to say of the fact that violations of social norms also routinely trigger activity in vmPFC (Berthoz et al. 2002), an area that seems to play an important role in the acquisition of knowledge about social norms (Moll et al 2005), and an area that is commonly thought to play an important role in integrating affective and cognitive representations. Early damage to vmPFC yields robust patterns of anti-social and abusive behavior (Anderson et al 1999), and later damage leads to impulsive decision-making, along with a generalized flattening of affect that is accompanied by exaggerated displays of anger in social situations that involve frustration or provocation (Koenigs & Tranel 2007). Surprisingly, however, the capacity for making moral judgments is largely preserved following damage to this area. People with lesions in vmPFC tend to think it permissible to act in ways that bring about aggregate welfare in 'impersonal' dilemmas, and they tend to reject the use personal force to intentionally cause physical harm. But they are more likely than controls to endorse the use of physical force to bring about aggregate welfare (Koenigs et al. 2007; Ciaramelli et al. 2007), and they are more likely to endorse bringing about aggregate welfare by harming family members or loved ones (Thomas, Croft, & Tranel 2011).

Activity in vmPFC is often observed in experiments involving decisions under risk or uncertainty, and mechanisms in this area appear to translate the reward and avoidance values computed by other systems into a common currency that can be used in behaviorally significant decision-making (Levy & Glimcher

2011; Montague & Berns 2002). High-conflict dilemmas pit the avoidance value of battery (or harming friends and family members) against the reward value of aggregate welfare, and people with damage to vmPFC may be unable to convert these representations into a common currency—and they may reflexively prefer the positive value of aggregate welfare regardless of the accompanying costs. Consistent with this hypothesis, Molly Crockett (in press) maintains that the model-based representations required for interpreting high-conflict dilemmas as involving an instance of battery might not be up-regulated in light of the experimental task-demands in people with damage to vmPFC. She argues that Pavlovian and model-free systems that depend on reinforcement histories might thus come to dominate judgment strategies in these patients, leaving the aversion to causing direct physical harm and the preference for aggregate welfare unscathed, but compromising the ability to integrate these distinct representational resources.

Importantly, people with vmPFC lesions make high-risk decisions in the face of economic loss, and display a marked tendency to accept risky bets even where the odds of winning are known to be vanishingly small (Clark et al. 2008; Saver & Damasio 1991). They also tend to accept a higher proportion of unfair offers in ultimatum games (Koenigs & Tranel 2007). In general, the preference for smaller immediate rewards over larger long-term rewards is highly correlated with the likelihood of rejecting unfair offers in the ultimatum game (Crockett et al. 2008; Crockett et al 2010a). So assuming that the rejection rates result from an impulsive failing to adjust the value of accepting unfair offers against the value of punishing someone who behaves unfairly should lead us to expect higher rejection rates as the result of the inability to triangulate the value of punishment against the immediate value of monetary gain.[11]

I contend that the activity of the mPFC during moral decision-making tasks cannot establish that emotions play a role in moral cognition. There are too many alternative hypotheses, which rely on the triangulation of affective and intentional information, but on the integration of multiple evaluative representations that are not adequately captured by these folk-taxonomic terms. On the hypothesis that the vmPFC converts distinct value-representations into a common currency to allow for comparative judgments, we have no clear reason to posit emotional processing as part of moral cognition—though I return to this point in the concluding section of this paper.[12]

4. The insula, the basal ganglia, and more neuroeconomics

Although the insula has not typically been targeted by experiments in moral psychology, the broader judgment and decision-making literature has frequently revealed activity in this area. The insula is an evolutionarily old structure that receives projections from somatosensory cortex, and the anterior insula is commonly activated by disgusting and painful stimuli (Singer, Critchley, & Preuschoff 2009). Many people working in moral psychology are also likely to be aware that insula activity is correlated with emotional experience, and that the insula is sometimes thought to be the neural substrate for the aversive somatic markers (Damasio et al. 2000; Paulus et al. 2003). The insula is part of a distributed network

[11] I cannot do justice to this interesting literature here. However, the depletion of serotonin triggers both increased impulsivity and an increase in the rate at which unfair offers are rejected. Serotonin innervates a network that includes the vmPFC, insula, and amygdala (Crockett et al. 2012), and both experimental data and computational models suggest that this network plays a critical role in comparisons between anticipated reward-values over various time delays (Daw, Kakade, & Dayan 2002; Rogers 2011); more importantly, it appears to play a role in pruning decision trees by eliminating options expected to lead to aversive outcomes (Crockett et al. 2011; Crockett 2013). Serotonin does not so much promote self-control, as modulate the effect of predictions about aversive stimuli (Crockett et al. 2009; Dayan, 2008). This is important because depleting serotonin modulates the impact of punishment-related and aversive signals, and yields an exaggerated aversion to stressful or threatening stimuli (Cools et al. 2008), and increasing serotonin triggers an increase in subjective disapproval for actions that bring about aggregate welfare in high-conflict dilemmas (Crockett et al. 2010b). In this case, serotonin seems to increase the salience of battery avoidance by up-regulating this task-relevant evaluative representation.

[12] It is often noted that frontotemporal dementia [FTD] yields motivational deficits, flattened affect, and an increase in the proportion of judgments based of aggregate welfare (Mendez et al. 2005). But FTD yields numerous processing deficits and widespread damage to both the frontal and parietal cortex, and this makes it unclear whether the flattening of affect is causally implicated in this pattern of judgments, or whether it is just highly correlated with it. Furthermore, the fact that disrupting the right dlPCF yields a similar pattern of judgments makes it incredibly difficult to interpret these data.

that includes the striatum, mPFC, parietal cortex, and amygdala, which is frequently activated in tasks that require decision-making under risk or uncertainty (Clark et al 2008). So it should come as no surprise that a recent imaging task has revealed bilateral activity in the anterior insula, along with activity in the basal ganglia (caudate), an area associated with processing reward values, when people judge an action to be morally wrong (Schaich Borg et al 2011).

The insula is commonly seen as an interface between model-free and Pavlovian learning systems that play a critical role in reward processing (e.g., basal ganglia, OFC, and amygdala), and structures that are more commonly associated with goal-based and value-based cognition (e.g., the ACC and the prefrontal-parietal network), both of which are frequently activated by tasks requiring risky decision making (Bechara, 2001; Moll et al 2006). People with insula lesions are less sensitive to risk, and they fail to adjust the magnitude of their bets against known chances of winning. In this respect, they differ from people with vmPFC lesions, who engage in impulsive betting behavior but adjust their bets in accordance with information about their chances of wining (Clark et al. 2008). Imaging studies have revealed anterior insula activation when people *anticipate* making risk-averse choices, and increased activity in this area also predicts safer choices following a loss in a double-or-nothing game (Kuhnen & Knutson 2005; Paulus et al 2003). Finally, activity in both the dlPFC and anterior insula are highly correlated with the magnitude of unfair offers in ultimatum games (Safaney et al 2003), and complex neuroeconomic tasks reveal activity in the basal ganglia that tracks considerations of subjective utility (caudate nucleus) and aggregate utility (putamen), while ongoing activity in the anterior insula correlate with perceived inequality (Hsu et al 2008).

Together, these data support the hypothesis that the anterior insula plays a role in signaling the probability of aversive outcomes. Doing so in a way that produces both risk-signals and risk-prediction-error signals that can facilitate both learning and on-line updating in light of decisions made on the fly (Preuschoff 2006, 2008; Quartz 2009). Put differently, the mechanisms in the anterior insula produce signals that predict the likelihood of aversive outcomes, as well as signals that correlate with inaccuracies in these predictions; this is what allows for the adjustment of online behavior when risk-predictions turn out to be wrong. So, why does judging an action to be morally wrong evoke activity in both the caudate nucleus and the anterior insula? One plausible hypothesis, which is now beginning to gain some support, is that a suite of domain-general learning and valuational mechanisms play a critical role in both neuroeconomic tasks and in tasks examining moral judgments (Crockett 2013; Cushman in press; Huebner in prep; Quartz 2009; Railton forthcoming).

5. An alternative hypothesis

At the end of this long review of existing data, we are now in a position to consider an alternative approach to interpreting these results, an interpretation that sidesteps the distinction between affective and cognitive circuits. Investigations using a variety of different experimental techniques have converged on the hypothesis that multiple evaluative representations are deployed in parallel when people respond to moral dilemmas. It is clear that some systems are assigning positive value to aggregate welfare, and that others are assigning negative value to direct physical harm; it is also clear that people compute the expected positive utility of engaging in an action, as well as the expected costs of engaging in various actions, and that these judgments sometimes depend on goals and values; but where are the emotional systems in this process?

Recall the data collected by Decety & Cacioppo (2012). These data are consistent with the hypothesis that representations of intentional harm are computed in the STS, while the circuit linking the amygdala and vmPFC allows for the computation of expected reward values for the action that is being considered. Something similar is suggested by an experiment carried out by Amitai Shenhav & Josh Greene (2010), who used moral dilemmas that varied in the number of lives saved and the likelihood of success. They found activity in right anterior insula and ventral striatum that was strongly correlated with individual sensitivity to the number of lives saved and lost; and they found activity in the vmPFC/mOFC that was highly correlated with the expected value of gains and losses (i.e., the interaction between the number of lives lost and the probability of success). Finally, Jana Schaich Borg and colleagues (2011) found that judgments about more controversial situations also recruited

mechanisms in the vmPFC, posterior cingulate, and the TPJ. These data suggests that people may be comparing the expected value of actions against models of alternative possible outcomes, up-regulating or down-regulating evaluative representations in light of task-demands or previously encoded values.

This brings us to my core positive claim. The existing data in moral psychology suggest that moral judgments rely on a complex network of interconnected mechanisms that carry out things like utility-assessment, risk-assessment, and counterfactual reasoning, and they suggest that initial impulsive judgments are sometimes triangulated against other sorts of goals and values. Many of these processes may be affective in a broad sense of the term, but none of them are emotions. Indeed, I contend that the neuroscientific confirm a set of fairly traditional assumptions about the considerations people employ in making moral judgments. It is the speed and automaticity of these computations that goes beyond anything that could be expected *a priori*—and this points the way toward a more plausible approach to the study of moral cognition. As Fiery Cushman (in press) notes, everyone in moral psychology must now acknowledge that moral judgments and morally significant behavior depend on the "motivational force derived from value representations, as well as computational processing over a representation of the action in question".

Recent approaches to computational modeling in neuroscience suggest an intriguing alternative approach to learning and decision-making that flouts the distinction between affect and cognition, suggesting that these folk-taxonomic categories fail to capture the evaluative structure of neural computation (Quartz 2009). The key insight is that the capacity for evaluative decision-making is likely to have evolved to guide behavior in ways that allow organisms to successfully 'recharge their batteries' in a world where energy is limited and its distribution is uncertain (Montague 2006). Put less figuratively, every strategy that an organism can adopt for finding food and finding mates is risky, and the cost of failure is infinitely high; so biological cognition typically depends on adaptive decision-making capacities that can be updated in light of new information and subjective expectations about the distribution of risks and benefits in the environment. In line with this hypothesis, neuroscientists have discovered a wide variety of evaluative mechanisms dedicated to things like reward-prediction, risk-prediction, and other forms of valuation (Huebner 2012). Somewhat surprisingly, these evaluative mechanisms are found both in regions that were classically seen as centers of emotion and in regions that were classically thought to facilitate working memory and controlled cognition.

For example, the basal ganglia are part of the midbrain dopaminergic system that dominates discussions of reinforcement learning (Montague et al. 1996; Schultz 1998, 2010). Mechanisms in this region compute prediction-error signals for expected rewards, and implement a bi-directional teaching signal that tracks the extent to which outcomes are better or worse-than-expected (specifically, spiking rates in the basal ganglia increase when rewards are better-than-expected, decrease when they are worse-than-expected, and are unaffected when rewards are accurately predicted; Montague et al. 1996). Specifically, mechanisms in the ventral caudate produce 'fictive error' signals that allow for comparisons between actual outcomes and models of the way things might have been; these signals allow organisms to update their expectations in light of imagined rather than real feedback (Lohrenz et al. 2007). Parallel mechanisms in the ventral striatum compute expectations where the distribution and likelihood of rewards is uncertain. In concert with mechanisms in the anterior insula, the ventral striatum facilitates the evaluation of risk, and computing risk-prediction-error signals (Preuschoff et al. 2006, 2008; Quartz 2009); a final component of this system, centering on the orbitofrontal cortex including vmPFC, represents a variety of distinct reward values, relying on inputs from the basal ganglia to facilitate decision making on the basis of the probability of a positive outcome given recent patterns of gains and losses (Frank & Claus 2006; Shenhav & Greene 2010).

The representations employed by these systems point in two directions: they indicate the way that world is (or might be) and they motivate us to pursue (or avoid) the things they depict (Millikan 1995). They implement both the learning signals and the motivational "umph" required for Pavlovian, model-free, and model-based learning (Rangel et al. 2008; Liljeholm & O'Doherty 2012). The simplest Pavlovian systems rely on associative computations that link values to fixed features of the world, producing motivations to approach or avoid biologically significant stimuli.

Model-free systems are more complex. They assign value on the basis of reinforcement history, but they also compute prediction-error signals for future risks and rewards; these signals are used to dynamically update the value assigned to a stimulus by adjusting behavior when things go better or worse than expected. Finally, the most computationally expensive model-based systems generate forward looking decision trees that can be used to represent distributions of possible values for various actions and outcomes; importantly, their assignments of value can depend on goals, rules, and policies that aggregate both potential and actually experienced outcomes.

It is commonly hypothesized that these mechanisms collectively compute polysensory and multimodal signals to guide attention, learning, and action-selection in ways that will maximize valuable outcomes. In many organisms, evaluative signals are only computed for primary rewards such as food, but these signals can be attuned to almost any reward-predicting stimuli. Indeed, there is growing evidence that these mechanisms can even facilitate cultural attunement by treating norm compliance as rewarding and norm violation as aversive (Klucharev et al., 2009, 2011; Montague, 2006).

The links between these models and moral cognition have only recently come to fore. [13] For example, Molly Crockett (2013) has proposed an approach to moral psychology that builds on the distinction between Pavlovian, model-free, and model-based systems. She maintains that the outputs of these systems need not converge, and that the interactions between these systems are likely to yield variations in judgment and behavior. Specifically, she suggests that model-based systems can generate structural descriptions of action-outcomes pairs, creating decision trees that can be searched for the best possible option, while model-free systems simultaneously assign value to the described actions on the basis of reinforcement histories, yielding aversions to things like causing direct physical harm or causing harm to friends and loved ones. The outputs of these systems must then be integrated with Pavlovian habits and triangulated against motivations to pursue or avoid various situations. Behavioral choices would then emerge as a result of the combined influence of these various systems. There is much to recommend this hypothesis, though I would add that differences between risk-predictions and reward-predictions are also relevant.

If this hypothesis is approximately correct, then factors such as the strength of different evaluative representations, as well as the inhibitory and excitatory relations between systems become crucial variables that must be taken into account in moral psychology—much as I suggested above in my review of affective priming studies. There is a lot more to say in this regard (see Cushman in press). But for my current purposes, the most relevant thing to notice is that this approach begins from the assumption that moral judgments are likely to depend on integrated networks of action-guiding systems, which evolved to guide *behavior* in dangerous and unpredictable environments. These systems have clearly been re-purposed to respond to social-normative phenomena, but most of them retain their action guiding character. They assign values to outcomes and predictions, trigger the production of appetites, and motivate us to pursue various actions. So moral judgment are always infused with *valuation*, and intimately bound up with behavioral motivation (cf., Greene et al. 2004, 397). This makes it all the more striking that this approach to moral cognition opens up a novel approach to the study of moral universals. I cannot develop this hypothesis fully here, but I would like to close by sketching a strategy for interpreting the existing data in a way that can fund a minimalist approach to universal moral grammar (Crockett 2013, Mikhail 2007, 2012).

John Mikhail (2011) maintains that an *exclusive* focus on neurocogitive and neurobiologiocal phenomena is likely to obscure patterns of stability that emerge only at higher levels of description. Beyond this, an exclusive focus on judgments about moral dilemmas and other borderline moral phenomena might obscure

[13] There are obvious similarities between this computational approach and dual-process theories of judgment and decision-making. But whether the relationship is one of implementation will depend on the precise commitments of the dual process theory under consideration. Many dual-process models assume that the slow-processes operate consciously, while the fast-processes operate subconsciously and associatively. By contrast, model-based, model-free, and Pavlovian systems can all operate reflexively, and their outputs are often integrated in guiding behavior. Of course, there are dual-process accounts that allow for reflexive rule-based processing. I have a preference for the neuro-computational approach as it is more rigorously and formally articulated.

robust patterns of moral judgments by focusing attention on the types of cases where people are more likely to disagree about what should be done. Building on methodological insights from corpus linguistics, Mikhail (2009) adopted the novel strategy of analyzing the penal codes of the 204 member states of the United Nations, and the Rome Statute of the International Criminal Court. In doing so, he uncovers a narrow set of overlapping principles governing laws against intentional killing, and a relatively narrow range of variation in the types of consideration that are thought to justify intentional killing (93% of the examined legal systems treated self-defense and mental illness as exculpating factors, while other considerations showed more cultural variability). A similar sort of analysis of suggests that the representation of battery is similarly robust across cultures.

Like Mikhail, I maintain that the most plausible approach to interpreting such data is to assume that a structural description is extracted from a target scenario, specifying arguments for the event such as the agent who acts and her mental state, the patient(s) who will be affected (and the number of patients affected), as well as other relevant features about causal and temporal organization of the events described. While it may be tempting to assume that this structural description is fully computed and evaluated by a single moral system, the values of these arguments are unlikely to be given a full interpretation at this point; instead, they serve to constrain a range of coherent structural organizations for morally evaluable actions (roughly, a syntactic structure). To take one example of how this might work, the distinction between intentional and accidental harms might require structural descriptions that specify a value of [-] or [+] for [accidental], with this feature requiring input from mentalizing systems in the rTPJ to generate evaluable representations of intentional harm (Young & Saxe 2008). Interpreting the relevant features yields a decision tree, which can then evaluated for moral significance; but intriguingly, this process of evaluation is likely to depend on parallel constraints imposed by model-based, model-free, and Pavlovian systems.

Model-based systems can employ stored representations of values and goals (which could either be innate or learned) to generate an aversion to battery, an aversion to unfair treatment in economic transactions, and a preference for promoting aggregate welfare. This is consistent with the lesion data as well as the rTMS data discussed above, and it is consistent with the fact that increased stress is inversely correlated with the preference for aggregate welfare; as Crockett (2013) notes, even minor stress leads people to abandon computationally expensive model-based reasoning and to deploy computationally cheap model-free mechanisms. Model-free systems are sensitive to individual differences in learning history and reinforcement; and the boundaries that they place on acceptable amounts of risk and reward are likely to be sensitive to individual differences in impulsivity and risk-aversion—yielding variance in the patterns of judgments that people provide in response to moral scenarios. Finally, Pavlovian systems may be responsive to the outputs of model-based mechanisms, and may play an important role in pruning trees to settle on an evaluation of a structural description.

On the assumption that outputs from these diverse computational systems must be integrated to produce moral judgments, we should expect to see some patterns of stability that emerge within a broader sea of cultural and individual differences. Specifically, we should expect to find—as we do—decreased patterns of interpersonal agreement in judgments about high-conflict dilemmas as well as dilemmas that involve unfamiliar and unpredictable situations (Greene et al. 2004; Huebner, Huaser, & Pettit 2011). These types of judgments require triangulating predictions made by context-sensitive, model-free systems against (often underspecified) evaluative representations produced by model-based mechanisms. Where cultural pressures lead to a convergence between model-based and model-free systems, we should expect to find stable patterns in moral judgments—and it would be unsurprising if this were to occur in the case of norms against battery and unjustified intentional killing. However, where a difficult trade off must be evaluated (for example, in a Sophie's Choice type case), we should expect to find individual differences in the impact of model-based as opposed to model-free systems on the resulting judgment.

Understanding judgments about morally salient situations is likely to require careful attention to the interactions between the core moral systems that produce structural descriptions of actions, and the various evaluative

mechanisms that drive goal-directed behavior. Of this much I am reasonably certain. But worrying about the role of emotion in this process is unlikely to be rewarding.[14] So, I propose to stop thinking about the role of emotion in moral psychology.

Works cited:

Aharoni, E, W Sinnott-Armstrong, & K Kiehl (2012). Can Psychopathic Offenders Discern Moral Wrongs? J Abnorm Psychol 121, 2, 484-497.

Anderson S, A. Bechara, H Damasio, D Tranel, A Damasio (1999) Impairment of social and moral behavior related to early damage in human prefrontal cortex. Nat Neurosci 2: 1032-1037.

Ashby, F, A Isen, & A Turken (1999). A neuropsychological theory of positive affect and its influence on cognition. Psychol Rev, *106*, 529-550.

Ashby, F, V Valentin, A & Turken (2002). The effects of positive affect and arousal on working memory and executive attention. In S Moore & M Oaksford (Eds.), *Emotional cognition* (pp. 245-287). Amsterdam: John Benjamins Publishing Company.

Bartels D (2008) Principled moral sentiment and the flexibility of moral judgment and decision-making. Cognition 108:381-417

Baumard, N, J André, and D Sperber (2013) A Mutualistic Approach to Morality, *Behav Brain Sci*, 36, 1: 59-122.

Bechara, A (2001). Neurobiology of decision-making. Semin Clin Neuropsychiatry 6: 205–216.

Berridge, K. (1996) Food reward: Brain substrates of wanting and liking. Neurosci Biobehav Rev, 20, 1-25, 1996.

Berridge, K & M Kringelbach (in press). Neuroscience of affect: brain mechanisms of pleasure and displeasure. Curr Opin Neurobiol.

Berthoz S, J Armony, R Blair, & R Dolan (2002) An fMRI study of intentional and unintentional (embarrassing) violations of social norms. Brain 125:1696-1708.

Bless, H, & K Fiedler (2006). Mood and the regulation of information processing and behavior. In J. P. Forgas (Ed.), Affect in social thinking and behavior (pp. 65−84). New York: Psychology Press.

Buckholtz, J, C Asplund, P Dux, D Zald, J Gore, O Jones, & R Marois (2008). The Neural Correlates of Third-Party Punishment. *Neuron*, Vol. 60, pp. 940-950.

Buckner, R, J Andrews-Hanna, & D Schacter (2008) The brain's default network. Ann NY Acad Sci, 1124: 1-38.

Chajut, E, & D Algom. (2003). Selective attention improves under stress. J Pers Soc Psychol, 85(2), 231–248.

Ciaramelli E, Muccioli M, Ladavas E, di Pellegrino G (2007) Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. Soc Cogn Affect Neur 2:84-92

Cima, M, F Tonnaer, & M Hauser (2010). Psychopaths know right from wrong but don't care. Soc Cogn Affect Neur, 5, 59-67

Clark, L, A Bechara, H Damasio, M Aitken, B Sahakian, & T Robbins (2008). Differential effects of insular and ventromedial prefrontal cortex lesions on risky decision-making. Brain, 131, 5, 1311-1322

Cools, R, A Roberts, & T Robbins (2008). Serotoninergic regulation of emotional and behavioural control processes. Trends Cogn Sci, 12: 31–40.

Cova, F, J Deonna, & D Sander (2013). The emotional shape of our moral life: anger-related emotions and mutualistic anthropology. *Behav Brain Sci, 36*(1), 86-87.

Crockett M, L Clark, G Tabibnia, M Lieberman, T Robbins (2008) Serotonin modulates behavioral reactions to unfairness. Science 320:1739.

Crockett M, L Clark, & T Robbins (2009). Reconciling the role of serotonin in punishment and inhibition in humans. J Neurosci, 29, 38: 11993-11999.

Crockett, M, L Clark, M Lieberman, G Tabibnia & T Robbins (2010a) Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion*, 10, 6:855-62.

Crockett, M, L Clark, M Hauser & T Robbins (2010) Serotonin selectively influences moral judgment and behavior through effects on harm aversion. PNAS, 107, 40**:** 17433-8.

Crockett, M, L Clark, J Roiser, O Robinson, R Cools, H Chase, H den Ouden, A Apergis-Schoute, D Campbell-Meikeljohn, B Seymour, B Sahakian, R Rogers, & T Robbins (2012). Converging evidence for central 5-HT effects in acute tryptophan depletion. *Mol Psychiatr,* 17, 2:121-3.

Crockett, M (in press). Models of morality. *Trends Cogn Sci.*

Cummins, D & R Cummins (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology: Emotion Science*, 3, 1-16.

Cushman, F (in press). "Action, outcome and value: A dual-system framework for morality and more." Pers Social Psych Rev.

Cushman, F, L Young & J Greene (2010). Our mutli-system moral psychology. In The Oxford Handbook of Moral Psychology, J. Doris et al (Eds). Oxford University Press.

D'Ardenne K, N Eshel, J Luka, A Lenartowicz, L Nystrom, & J Cohen (2012) Role of prefrontal cortex and the midbrain dopamine system in working memory updating. PNAS, 109:19900–19909.

Damasio A, T Grabowski, A Bechara, H Damasio, L Ponto, J Parvizi & R Hichwa (2010). Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat Neurosci. 3:1049-1056.

Daw, N, S Kakade & P Dayan (2002). Opponent interactions between serotonin and dopamine. Neural Networks, 15 (4-6), 603-616.

Dayan P (2008). The role of value systems in decision making. In Engel C & W Singer (Eds) *Better than Conscious?* Frankfurt, Germany: MIT Press, 51-

Decety, J, & S Cacioppo (2012). The speed of morality. J Neurophysiol, 108, 3068–3072.

Demartino, B, C Camerer & R Adolphs (2010). Amygdala lesion eliminates loss aversion, PNAS 107 (8) 3788-3792.

Deutch, A & R. Roth (1990). The determinants of stress-induced activation of the prefrontal cortical dopamine system. Prog Brain Res, 85, 367–403.

Eskine, J, A Kacinik, & J Prinz (2011). "A bad taste in the mouth," Psychol Sci, 22: 295–299.

14 For a very different argument for a similar conclusion, see Sinnot-Armstrong (2011).

Estrada, C, A Isen & M Young (1997). Positive affect facilitates integration of information and decreases anchoring in reasoning among physicians**.** Organ Behav Hum Dec, 72(1), 117-117

Frank M & E Claus (2006). Anatomy of a decision. Psycholical Review, 113, 300–326

Fredricksen B & C Branigan (2005). Positive emotions broaden the scope of attention and thought-action repertoires. Cognition Emotion, 19(3), 313-332.

Glenn, A, A Raine & R Schug (2009a). The neural correlates of moral decision-making in psychopathy. Mol Psychiat*, 14,* 5-6.

Glenn, A, A Raine, R Schug, L Young, & M Hauser (2009b). Increased DLPFC activity during moral decision-making in psychopathy. Mol Psychiat*, 14,* 909-911.

Goldberg J, J Lerner & P Tetlock (1999). Rage and reason. Eur J Soc Psychol, 29: 781–795.

Greene, J (2009) The cognitive neuroscience of moral judgment, in *The Cognitive Neurosciences IV*, M.S. Gazzaniga, Ed. MIT Press, Cambridge.

Greene, J, R Sommerville, L Nystrom, J Darley, & J Cohen (2001). An fMRI investigation of emotional engagement in moral Judgment. Science, Vol. 293, Sept. 14, 2001, 2105-2108.

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

Greene, J, S Morelli, K Lowenberg, L Nystrom & J Cohen (2008) Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107, 1144-1154

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

Hampton A, R Adolphs, M Tyszka & J O'Doherty (2007) Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex. Neuron 55:545–555.

Han S, J Lerner & R Zeckhauser. Disgust promotes disposal. J Risk Uncertainty**.** 44(2):101-113.

Harenski C & S Hamann (2006) Neural correlates of regulating negative emotions related to moral violations. Neuroimage 30:313-324

Harenski, C, K Harenski & K Kiehl (2010). Aberrant neural processing of moral violations in criminal psychopaths. J Abnorm Psychol, 119, 863– 874.

Haslam, N (2006). Dehumanization. *Pers Soc Psychol, 10*(3), 252-264.

Heekeren, H, I Wartenburger, H Schmidt, H Schwintowski & A Villringer (2003). An fMRI study of simple ethical decision-making. Neuroreport 14, 1215–1219.

Holland, R, M Hendriks & H Aarts (2005). Smells like clean spirit. Psychol Sci, 16(9), 689-693.

Horberg, E, C Oveis, D Keltner & A Cohen (2009). Disgust and the moralization of purity. J Pers Soc Psychol, 97: 963–976.

Hsu M, C Anen, S Quartz (2008) The right and the good. Science 320:1092-1095.

Huang, J & J Bargh (in press). The Selfish Goal. Behav Brain Sci

Huebner, B (2011). Critiquing empirical moral psychology. Philos Social Sci**,** 41, 1: 50-83.

Huebner, B (2012). Surprisal and Valuation in the Predictive Brain. Front Psychol*,* 3, 415.

Huebner, B, S Dwyer & M Hauser (2009). The role of emotion in moral psychology. *Trends Cogn Sci, 13* (1), 1-6.

Huebner, B., M Hauser & P Pettit (2011). How the Source, Inevitability and Means of Bringing About Harm Interact in Folk-Moral Judgments. Mind Lang, 26 (2): 210-233.

Huebner, B & R Rupert (in press). Massively representational minds are not always driven by goals, conscious or otherwise. Behav Brain Sci.

Hunsinger, M, L Isbell & G Clore (2012). Sometimes happy people focus on the trees and sad people focus on the forest. Pers Soc Psychol B*, 38*, 220-232.

Isen, A (2001). An Influence of Positive Affect on Decision Making in Complex Situations. J Consum Psychol 11, 2: 75-85.

Isen, A, A Rosenzweig & M Young (1991). The influence of positive affect on clinical problem solving. Medic Decis Making, 11(3), 221.

Kelly, D. (2011). Yuck! Cambridge: MIT Press.

Kirk, U, JDownar & P Montague (2011). Interoception Drives Increased Rational Decision-Making in Meditators Playing the Ultimatum Game. Front Neurosci **5:49**

Klucharev V, K Hytönen, M Rijpkema, A Smidts, G Fernández (2009). Reinforcement learning signal predicts social conformity. Neuron 61, 140–151.

Klucharev V, M Munneke, A Smidts, G Fernández (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. J Neurosci 31, 11934–11940

Knoch D, A Pascual-Leone, K Meyer, V Treyer & E Fehr (2006a) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314:829-832.

Knoch, D, L Gianotti, A Pascual-Leone, V Treyer, M Regard, M Hohmann & P Brugger (2006b). Disruption of Right Prefrontal Cortex by Low-Frequency Repetitive Transcranial Magnetic Stimulation Induces Risk-Taking Behavior. J Neurosci, 26(24): 6469-6472.

Knoch, D, M Nitsche, U Fischbacher, C Eisenegger, A Pascual-Leone & E Fehr (2008). Studying the Neurobiology of Social Interaction with Transcranial Direct Current Stimulation. Cereb Cortex,18 (9):1987-1990.

Koenigs M & D Tranel (2007) Irrational economic decision-making after ventromedial prefrontal damage. J Neurosci 27:951-956.

Koenigs M, L Young, R Adolphs, D Tranel, F Cushman, M Hauser & A Damasio (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446: 908-911.

Kuhnen, C & B Knutson. The neural basis of financial risk taking. Neuron, 47 (2005), pp. 763–770

Landreth, A. (2008). Emotion and the neural substrate of moral judgment. Fact and Value in Emotion, 12, 49, 157–179

Levy D & P Glimcher (2011) Comparing apples and oranges. J Neurosci **31**:14693–14707.

Liljeholm M & J O'Doherty (2012). Contributions of the striatum to learning, motivation, and performance. Trends Cogn Sci, 16, 467–475.

Lohrenz T, K McCabe, C Camerer, P Montague (2007). Neural signature of fictive learning signals in a sequential investment task. PNAS 104, 9493–9498.

May, J (forthcoming). Does Disgust Influence Moral Judgment? Australas J Phil.

Mikhail, J (2007). Universal Moral Grammar. *Trends Cogn Sci, 11*, 143-152.

Mikhail, J (2009). Is the Prohibition of Homicide Universal? Brooklyn Law Review, 497.

Mikhail, J (2011). Elements of Moral Cognition. Cambridge: Cambridge University Press.
Mikhail, J (forthcoming). Any Animal Whatever? Ethics
Milgram, S & J Sabini (1978). On maintaining urban norms. In A Baum et al. (Eds), Advances in Environmental Psychology. Hilsdale, New Jersey: Lawrence Earlbaum.
Miller E & J Cohen (2001) An integrative theory of prefrontal cortex function. Ann Rev Neurosci, 24:167–202.
Millikan, R (1995). Pushmi-Pullyu Representations. Philosophical Perspectives 9:185-200
Moll J, P Eslinger & R Oliveira-Souza (2001) Frontopolar and anterior temporal cortex activation in a moral judgment task. Arq Neuropsiquiatr 59:657-664.
Moll J, R de Oliveira-Souza, P Eslinger (2003). Morals and the human brain. Neuroreport 14, 3: 299-305.
Moll, J, R Zahn, R de Oliveira-Souza, F Krueger & J Grafman (2005). The neural basis of human moral cognition. Nat Rev Neurosci, 6(10), 799-809.
Moll J, F Krueger, R Zahn, M Pardini, R de Oliveira-Souza & J Grafman (2006) Human fronto-mesolimbic networks guide decisions about charitable donation. PNAS 103:15623-15628
Montague P (2006). Why Choose This Book? New York: Dutton
Montague P & G Berns (2002). Neural economics and the biological substrates of valuation. Neuron **36**:265–284.
Montague P, P Dayan & T Sejnowski (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.
Nichols, S (2002). Norms with Feeling Cognition 84, 2: 221–236.
Nichols, S (2004). *Sentimental rules.* Oxford University Press.
Papo, D (2013). Time scales in cognitive neuroscience. Front Physiol.
Paulus, M, C Rogalsky, A Simmons, J Feinstein & M Stein (2003) Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. Neuroimage, 19, 1439–1448.
Paxton, J & J Greene (2010). Moral reasoning. Topics Cogn Sci, 2, 511-527.
Pizarro, D, B Detweiler-Bedell & P Bloom (2006). The creativity of everyday moral reasoning. In J. C. Kaufman & J. Baer (Eds.), Creativity and reason in cognitive development. Cambridge: Cambridge University Press, 81–98.
Pizarro, D, Y Inbar & C Helion (2011). On Disgust and Moral Judgment, Emotion Rev 3, 3: 267–268.
Plakias, A (2013). The Good and the Gross. Ethical Theory & Moral Practice 16/2: 261–278.
Preuschoff K, P Bossaerts, S Quartz (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51, 381–390.
Preuschoff K, S Quartz & P Bossaerts (2008). Human insula reflects risk predictions errors as well as risk. J. Neurosci. 28, 2745–2752.
Prinz, J (2006). The Emotional Basis of Moral Judgments. Philos Explor 9 (1): 29-43.
Prinz, J (2007). The Emotional Construction of Morals, Oxford University Press.
Quartz, S (2009). Reason, Emotion, and Decision-Making. Trends Cogn Scis. 13(5). 209-215.
Railton, P (forthcoming). The affective dog and its rational tail. Ethics.
Rangel A, C Camerer & P Montague (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. 9, 545–556.
Rogers, R (2011). The Roles of Dopamine and Serotonin in Decision Making. Neuropharmocology, 36, 1, 114-132.
Rozin, P (1997) Moralization and becoming a vegetarian. *Psychol sci*, 8, 67–73.
Rozin, P, L Lowery, S Imada & J Haidt (1999). The CAD triad hypothesis. J Pers Soc Psychol*,* 76, 574–586.
Rupert, R (forthcoming). Embodiment, Consciousness, and the Massively Representational Mind. Philos Topics.
Sanfey A, J Rilling, J Aronson, L Nystrom & J Cohen (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755-1758
Saver, J & A Damasio (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. Neuropsychologia, 2912
Saxe, R, D Xiao, G Kovacs, D Perrett & N Kanwisher (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. Neuropsychologia*, 42*, 1435-1446.
Schaich Borg, J, W Sinnott-Armstrong, V Calhoun, & K Kiehl (2011). The Neural Basis of Moral Verdict and Moral Deliberation. Soc Neurosci, iFirst, 1–16
Schnall, S, J Haidt, G Clore & A Jordan (2008a). Disgust as Embodied Moral Judgment, Pers Soc Psychol B 34/8: 1096–1109.
Schnall, S, J Benton, & S Harvey (2008b). With a Clean Conscience, Psychol sci 19/12: 1219–1222.
Schröder, T & P Thagard (2013). The Affective Meanings of Automatic Social Behaviors. Psychol Rev, 120, 1, 255–280.
Schultz W (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiol. 80, 1–27.
Schultz W (2010). Dopamine signals for reward value and risk. Behav Brain Funct 6, 24.
Seidel, A & J Prinz (2013). Sound morality. Cognition, 127 (1): 1–5.
Shenhav, A & J Greene (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron, 67, 667-677.
Sherman, G & J Haidt (2011). Cuteness and disgust. Emotion Rev, 3, 245-251. 84
Singer, T, H Critchley, & K Preuschoff (2009). A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci, 13*(8), 334–340.
Sinnot-Armstrong, W. (2011). Emotion and Reliability in Moral Psychology. Emotion Rev, 3, 3, 288 - 289
Skitka, L & D Wisneski (2011). Moral conviction and emotion*. Emotion Review, 3,* 328 - 330.
Spitzer M, U Fischbacher, B Herrnberger, G Grön & E Fehr (2007).The neural signature of social norm compliance. Neuron, 56: 185-96.
Starcke, K, C Polzer, O Wolf & M Brand (2011). Does stress alter everyday moral decision-making? Psychoneuroendocrino 36, 210—219.
Strohminger, N., R Lewis, & D Meyer (2011). Divergent effects of different positive emotions on moral judgment. Cognition 119, 295–300.
Suter, R & R Hertwig (2011). Time and moral judgment. Cognition 119, 454–458.
Tassy, S, O Oullier, Y Duclos, O Coulon, J Mancini, C Deruelle, S Attarian, O Felician & B Wicker (2012).

Disrupting the right prefrontal cortex alters moral judgement. *SCAN,* 7, 3, 282-288.
Thomas, B, K Croft & D Tranel (2011). Harming Kin to Save Strangers. J Cogn Neurosci 23(9), 2186-2196.
Valdesolo, P & D DeSteno (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychol sci, 17*, 476-477.
Wheatley, T & J Haidt (2005). Hypnotic Disgust Makes Moral Judgments More Severe, Psychol sci 16/10: 780–784.
Yang Q, L Yan, J Luo, A Li, Y Zhang, X Tian & D Zhang (2013) Temporal Dynamics of Disgust and Morality. PLoS ONE 8(5): e65094.
Young. L & R Saxe (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*(4), 1912-1920.
Youssef F, K Dookeeram, V Badeo, E Francis, M Doman, D Mamed, S Maloo, J Degannes, L Dobo, P Ditshotlo, & G Legall (2012). Stress alters personal moral decision making. Psychoneuroendocrino 37, 491–498.