In Press: Social and Personality Psychology Compass

Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm

Ryan Miller & Fiery Cushman
Brown University

Many studies attest to the critical role of affect in the condemnation of harmful actions, but few attempt to identify the precise representations underlying this affective response. We propose a distinction between two potential sources of affect: an aversion to the negative outcomes of an action versus an aversion grounded in the action itself. Whereas previous models have focused on outcome-oriented processes (e.g. empathy and victim perspective-taking), we argue that moral judgment is also strongly influenced by action-based aversions. Specifically, we propose that individuals engage in a process of 'evaluative simulation' when judging others, imagining how much it would bother them to perform the same action. Furthermore, we present evidence that this aversion can be based in superficial sensory or motor properties of the action. We consider how such 'action aversions' might be acquired, and we highlight important areas for future research.

Keywords: Moral judgment, harm, emotion, aversion, empathy

Emotion animates our moral lives. Compassion inspires sacrifice, rage incites reckless violence, guilt cries for forgiveness, and shame can compel destructive self-harm. Thus, our understanding of these behaviors cannot be complete without an intimate knowledge of their affective basis. Just as with moral behavior, research indicates a key causal role for emotion in moral judgment, the process of determining which behaviors are right or wrong in the first place (Greene, 2008; Greene & Haidt, 2002; Haidt, 2001). Understanding the precise affective contributions to moral judgment is therefore equally important and yet, in some respects, surprisingly underdeveloped. Which specific emotions contribute, how are they triggered, and how do they influence judgment processes?

We explore these questions in the context of one particular subset of morally relevant behaviors: harmful actions. We propose a distinction between two basic sources of affect: the undesirable outcomes of a harmful action, and features of the action itself. To get a sense for this distinction, consider the aversion you would feel toward punching your own mother in the face. On the one hand, it might depend on a representation of its negative outcomes, such as your mother's pain; we call this *outcome aversion*. Models of moral judgment that emphasize the importance of empathy—i.e., concern for the victim of a transgression—depend on the concept of

outcome-based aversion (e.g. Eslinger, Moll, & De Oliveira-Souza, 2002; Hoffman, 1987; Pizarro, 2000). A second potential source of affect, one perhaps less obvious, is the action itself. For instance, the mere thought of intentionally swinging your fist at your mother might elicit an aversive response without requiring you to consider the harm that it causes; we call this *action aversion*. Our aim is to understand whether action aversion exists, how it is acquired, and the scope of its influence on moral judgment and behavior.

Affect and Judgment

The role of emotion in moral judgment is an area of some debate. Although it is widely agreed that emotions *follow* from moral judgment (i.e., if I judge that your behavior was immoral, it makes me angry), it is still disputed whether emotions play a causal role in the judgment process (Huebner, Dwyer, & Hauser, 2009; Pizarro, Inbar, & Helion, 2011). From a certain perspective it would be remarkable if emotions did not influence moral judgments, simply because of the overwhelming evidence for their pervasive influence on judgments of other kinds (Bargh & Chartrand, 1999; Greenwald & Banaji, 1995; Murphy & Zajonc, 1993; Wilson, 1979; Zajonc, 1980). Numerous studies have demonstrated that individuals often rely on "emotional signposts" and gut reactions rather than purely rational analysis or deliberation when making a wide array of decisions (Damasio, 1994; Finucane, Alhakami, Slovic, & Johnson, 2000; Loewenstein, Weber, Hsee, & Welch, 2001; Schwarz & Clore, 1983),

There is additional evidence that emotion contributes to moral judgment and decision-making, specifically (Haidt, 2001), and individuals may even *prefer* moral decisions to be based on emotion rather than reason (Merritt & Monin, 2011). When it comes to research on harm, the literature is dominated by hypothetical moral dilemmas where harmful action directed towards one person can be used to save many others. The undisputed patriarch of such dilemmas is the trolley problem (Foot, 1967; Thomson, 1985). In the *switch* version of the dilemma, a runaway trolley is en route to kill five individuals but can be diverted onto a sidetrack where it will kill only one. Most people judge that it is permissible to divert the trolley. In the *footbridge* version, one must instead push a large man off a bridge and onto the tracks in order to stop the trolley and save the five people. In contrast to the *switch* version, most people judge pushing the man to be impermissible. How can we explain this discrepancy? Based upon neuroimaging and behavioral evidence, Greene and colleagues (2004; 2001) argue that the *footbridge* case (and other similar 'personal' moral dilemmas) generates a much stronger, negative emotional response due to the "up close and personal" nature of its violent action, whereas the *switch* case (and related 'impersonal' dilemmas) generates a weaker emotional response because the harm is more impersonal and removed. Subsequent studies of individuals with neurological and neurocognitive disorders support this interpretation. Frontotemporal dementia patients (Mendez, Anderson, & Shapira, 2005), psychopaths (Koenigs, Kruepke, Zeier, & Newman, 2012), and individuals with damage to the ventromedial prefrontal cortex (Ciaramelli, Muccioli, Làdavas, & Di Pellegrino, 2007; Koenigs et al., 2007) all exhibit socio-emotional and empathic deficits, and all show increased approval of the harmful action in personal moral dilemmas.

If the level of affect one experiences directly influences moral judgment, then manipulating affect prior to judgment should lead to changes in the perceived

permissibility of an action. Specifically, inducing negative affect should make an action seem less attractive and lead to greater condemnation, whereas blocking it—or inducing positive affect—should make the action seem less aversive and lead to greater endorsement. Precisely this pattern is observed across several studies. Individuals placed under acute stress report higher levels of anxiety (negative affect) than controls, and they more severely condemn harmful actions in personal moral dilemmas (Starcke, Ludwig, & Brand, 2012; Youssef et al., 2012). A similar effect occurs following increased serotonergic activity. The neurotransmitter serotonin is thought to play a role in inhibiting aggressive behavior by encoding the aversiveness of future outcomes, like harm (Dayan & Huys, 2009). Consistent with this account, pharmacologically increasing serotonergic activity leads to increased condemnation of harm (Crockett, Clark, Hauser, & Robbins, 2010). Conversely, inducing positive affect by means of a funny video (Valdesolo & DeSteno, 2006) and blocking negative affect through administration of an antianxiety drug (Perkins et al., 2012) both lead to greater approval of harm in personal moral dilemmas.

In sum, several lines of convergent evidence indicate that affect contributes to moral judgments of harm. However, very few studies have attempted to identify either the source of this affective response, or the psychological mechanism that generates it. We closely examine this gap in the literature, and explore two potential ways of filling it.

Identifying the source of affect in moral judgment

When we condemn canonically harmful actions, like pushing a man off a footbridge, what is the source of the emotion underlying this judgment? Asked another way, why does it feel worse to kill a man by pushing him than by flipping a switch[1]?

As a starting point, we can look to existing component models of moral judgment that specify the necessary ingredients for wrongness and blame. To a first approximation, people consider it morally wrong to intentionally harm another person who is innocent, absent any justification (Darley & Shultz, 1990; Guglielmo, Monroe, & Malle, 2009; Heider, 1958; Weiner, 1995). Can our emotional response to trolley-type dilemmas be understood in terms of this basic framework? In the *footbridge* version of the trolley dilemma, one must *intentionally* kill a man in order to save five lives, whereas in the *switch* version the man's death is merely an unfortunate side effect (although it is foreseen). And, there is good evidence that intentional harms generate stronger emotional responses than commensurate unintentional harms (Alicke, 2000; Decety & Cacioppo, 2012; Gray & Wegner, 2008; Russell & Giner-Sorolla, 2011; Weiner,

[1] By referring to action and outcome aversion as potential sources of *emotion*, we are not suggesting that they require the type of rich, cognitive construal thought to accompany more complex emotions, nor do we suggest that action and outcome aversion are themselves specific types of emotion. Rather, we view them primarily as sources of negatively valenced affect or arousal (i.e. aversion) that can make an action seem morally worse. We use the word here to connect our ideas to previous research on personal moral dilemmas, where the relevant effects that we are discussing and attempting to explain have been repeatedly (and loosely) referred to as 'emotional.'

1995). Thus, the difference in emotional intensity between the *push* and *switch* cases may depend in part on differences in the perceived intent of the perpetrator (Cushman & Young, 2011; Mikhail, 2007).

Yet, it is unlikely that mental state attributions specifically—or standard models of moral judgment more generally—can fully explain our aversion to, and condemnation of, harmful actions. Consider another variant of the trolley problem: the *trapdoor* case. It is identical to *footbridge*, except rather than directly pushing the man off the bridge, the agent can flip a switch that swings open a trapdoor, dropping the man onto the tracks. Though the agent's intent is identical in both cases, individuals judge using a trapdoor to be significantly more acceptable (Greene et al., 2009). Research has identified several such factors, including the presence of physical contact (Cushman, Young, & Hauser, 2006) and use of personal force (Greene et al., 2009), that influence moral judgment. Yet, these factors do *not* appear to influence ascriptions of intent or causation (Cushman & Young, 2011).

This review aims to explain why features such as the physical manner of harm matter. For simplicity, we will limit our discussion to intentional harms where we already have good evidence that moral judgment is sensitive to the physical nature of the action (Greene et al., 2009); whether the same factors influence judgments of unintentional or accidental harms is an interesting topic that requires further research. Thus, we ask: When we judge some intentional harms to be more aversive—and therefore morally worse—than others, what is the origin of this aversion?

Outcome Aversion

Because harmful actions have clear victims, one natural possibility is that our affective response to intentional harm is grounded in consideration of the victim's suffering. In other words, we experience empathy for the victim, and our aversion to their pain increases the perceived wrongness of the action. When considering one's own behavior, outcomes other than victim suffering might also be relevant: the prospect of punishment, for instance, or a loss of social standing. However, our present concern is the judgment of *third-party* behavior, and here we consider empathy for a suffering victim to be the most likely source of *outcome aversion.*

According to this interpretation, we can best characterize the emotional response evident in studies of personal moral dilemmas (i.e. those involving 'up close and personal' harm) as a type of concern for the proximate victim of the action. This entails that features of a dilemma that are ostensibly about the action, such as physical contact with the victim or the use of personal vs. mechanical force, may actually exert their influence on moral judgment by facilitating empathy for the victim. For instance, pushing a man off a bridge may seem more concrete and 'psychologically near' than flipping a switch, and these characteristics may allow for a richer empathic response (see Liberman, Trope, & Stephan, 2007 on the relationship between psychological distance and empathy). Non-utilitarian judgments in cases like the *footbridge* dilemma may then derive from an enhanced concern with a nearby and salient victim, compared with more distant and less salient alternative victims.

Several lines of evidence support the importance of outcome aversion. Psychologists working from diverse perspectives have suggested that empathy plays a critical role in proper moral development (Blair, 1995; Hoffman, 1982,

2001), and philosophers have posited that it is integral to the very meaning of morality (e.g. Blum, 1994; Murdoch, 2001/1970). It has also been directly argued that empathy occupies a causal role in the process of moral judgment (Pizarro, 2000). The possibility that we identify with the victim of a transgression during moral judgment is further bolstered by evidence suggesting that victim perspective-taking is spontaneous. Specifically, individuals who passively viewed one person harming another exhibited increased brain activation in the same "pain" regions that are activated when participants are explicitly instructed to adopt the victim's perspective (Decety, Michalska, & Akitsuki, 2008; Decety & Porges, 2011). If individuals are simulating the victim's pain even when *not* asked to make moral judgments, there's a good chance that they're engaging in the same simulation when they are. It is also notable that the effect of serotonin on moral condemnation discussed earlier (Crockett et al., 2010) only occurred in participants who were high in empathy. If serotonin increases the aversiveness of others' pain—that is if it works by modifying outcome aversion—then it is unsurprising that we find its greatest effect in people who are highly sensitive to the suffering of others. Finally, individuals with psychopathy exhibit decreased neural activity in empathy and pain regions when viewing suffering victims (Decety, Skelly, & Kiehl, 2013), and this may contribute to their increased willingness to endorse harm in personal moral dilemmas (Koenigs et al., 2012).

Action Aversion

Although there is strong evidence in favor of outcome aversion, there are good reasons to think that the condemnation of harm is also grounded in an emotional aversion to the action, independent of its negative consequences. This hypothesis entails at least two distinct, but related, components. First, people may make moral judgments of others by assessing their own aversion to performing the action in question. This would imply that judging the wrongness of a harmful action involves putting oneself in the *agent's*, rather than (or in addition to) the victim's, shoes. We call this 'evaluative simulation', because a moral evaluation depends upon the simulation of an action. Others have drawn a similar connection between personal aversions and moral judgment, noting the relatively low cost (and potential strategic advantage) of endorsing norms that prohibit actions one would prefer not to do anyway (Tybur, Lieberman, Kurzban, & DeScioli, 2013). A second, related component of our hypothesis is that the aversion to harmful action can be triggered directly by an action's intrinsic properties, and need not depend contingently upon the outcomes that are expected to follow. These intrinsic action properties may include, for instance, sensory or motor representations that become aversive through associative learning. Thus, certain "canonically" violent actions, like stabbing or shooting, may acquire an aversive quality that other less typically harmful actions lack. We consider each of these components in turn.

Evaluative Simulation

The evaluative simulation hypothesis receives its strongest support from a very specific corner of the moral domain: judgments of consensual sibling incest. Lieberman and Lobel (2012) found that moral disapprobation of third-party incestuous behavior was predicted by one's own family structure (number of siblings, their relative ages, etc.), and that this effect was mediated by one's own aversion to engaging in sibling incest. In other words,

having a sibling makes you more averse to committing incest yourself, which in turn makes you condemn others who do it. Furthermore, there is reason to believe that the source of this emotional aversion is the action itself rather than any perceived negative consequences. Even when individuals believe that there are, in fact, harmful outcomes (or appeal to such outcomes as the reason for their condemnation of the action), several studies have shown that such beliefs are poor predictors of moral judgment in cases of incest and related purity violations (Haidt, Bjorklund, & Murphy, 2000; Haidt, Koller, & Dias, 1993). Rather, perceptions of wrongness in such cases seem to be largely determined by the individual's own disgust sensitivity (Horberg, Oveis, Keltner, & Cohen, 2009).

We suggest that evaluative simulation is not limited to incest but is instead a common process in the moral sphere. If so, then we should observe broad connections between the motivational and affective mechanisms that influence first-person behavior, including the inhibition of harm, and the endorsement of commensurate moral norms. Recent research has provided evidence for both types of connections.

At a general level of description, two primary motivational systems appear to underlie human behavior: a Behavioral Inhibition System (BIS) that drives avoidant behavior and prevents actions associated with negative consequences, and a Behavioral Activation System (BAS) that governs appetitive behavior and encourages actions that lead to rewards (Gray, 1990). If first-person motivational processes are intimately tied to third-party moral judgment, as our account suggests, then we might expect the BIS to be most strongly related to proscriptive moral norms ("thou shalt nots") that require the inhibition of harmful actions, and the BAS to be most related to prescriptive norms ("thou shalts") that promote prosocial behavior. Consistent with this prediction, Janoff-Bulman, Sheikh, and Hepp (2009) found that experimentally priming the BIS led to increased generation of proscriptive norms, like "people should not hurt others" and "people should not steal," whereas priming of the BAS led to increased generation of prescriptive norms, like "people should help others" and "people should be kind." Furthermore, they found that individual differences in the sensitivity/strength of the BIS and BAS correlated with the subjective moral weight assigned to proscriptive and prescriptive norms, respectively.

In addition to basic motivational relationships, we should also expect congruence between the types of affect that influence our own behavior and the types of affect involved in condemning others. This relationship holds for purity violations: Disgust encourages the avoidance of contaminants or situations that might render one impure (Rozin, Haidt, & McCauley, 1993), and it selectively increases the perceived wrongness of third-party purity violations (Horberg et al., 2009). Is there evidence of such "affect congruence" in violations specifically involving harm? Yes, in the form of anxiety. The negative arousal associated with anxiety influences behavior by increasing the perceived aversiveness of harmful actions (Perkins et al., 2012) and inhibiting instrumental aggression (Haller & Kruk, 2006; Raine, 1996). Accordingly, several studies show that anxiety/stress inductions also increase moral condemnation of harmful actions (Perkins et al., 2012; Starcke et al., 2012; Youssef et al., 2012). Further research is necessary to determine whether a similar congruency is observed between other types of affect and related moral concerns.

We have seen evidence suggesting that the motivational and affective systems

that guide one's own behavior also contribute to the moralization of third-party behavior, consistent with the notion of evaluative simulation. This is an important step, but questions remain about the precise nature of these behavioral aversions. We now examine evidence for the second component of our hypothesis: that our first-person aversions to particular actions can be grounded in properties of the *actions,* independent of their negative consequences.

The Aversive Properties of Actions

If the performance or simulation of canonically harmful actions is accompanied by aversive arousal, what is the source of this aversion? It is likely due in part to outcome aversion, yet there is reason to doubt that this is the whole story. Cushman et al. (2012) monitored physiological changes in participants while asking them to perform "pretend" harmful actions (e.g. pull the trigger of a fake gun aimed at an experimenter), witness someone else perform these actions, or carry out motorically-controlled neutral actions (e.g. squirt water from a spray bottle). If the aversion to performing an action is derived solely from the action's potential to cause harm, then the 'perform' and 'witness' conditions should be identical: they both contain the same potential for "harm." Contrary to this prediction, signs of aversive arousal were greater in the perform condition than either the witness or control conditions, suggesting that the aversion associated with harmful actions is not reducible to concern for a victim. A similar conclusion can be drawn from a study by Navarette and colleagues (2012) that examined arousal associated with harmful actions vs. harmful omissions. In the context of the *switch* version of the trolley dilemma, participants exhibited heightened arousal when they actively diverted the trolley onto the track with only a single individual (action) compared to when the trolley was already headed toward the individual (omission). Because the outcomes were identical in both cases, the heightened arousal associated with flipping the switch cannot simply index the aversiveness of the harmful outcome.

If the aversion associated with an action cannot be explained entirely in terms of victim harm, what are other potential contributors? Self-oriented concerns like fear of retaliation, punishment, or condemnation are likely candidates; after all, personally causing harm renders one vulnerable to a slew of negative consequences that merely observing harm does not. Nevertheless, we argue that attempts to explain aversiveness solely in terms of outcome-based considerations—whether oriented toward the victim or the self—are insufficient. Rather, we propose that part of an action's aversiveness can be non-contingent, with triggers including low-level sensory or motor features of the action itself.

Possibly, sensitivity to such triggers may be innate. For instance, Greene and colleagues (2004) proposed that up-close, direct violent action—behaviors commonly available to our primate ancestors—might trigger an evolved aversive response in order to inhibit counterproductive aggression and maintain social order. We are not aware of further research identifying the cognitive properties or neural substrates of such an innate response, and this remains an important area for further study.

Another possibility is that the relevant triggers are learned. Blair (1993, 1995) offers one such model. Drawing on observations of violence inhibition among non-human animals, Blair proposes that humans have an innate aversion to the distress of others (e.g. crying, yelling). When an aggressive act leads to harm, the

distress cues exhibited by the victim trigger a Pavlovian (i.e. instinctive) withdrawal response in the aggressor, stopping the behavior. Crucially, any mental representation of the behavior that is active in the aggressor's mind at the time that the withdrawal response is initiated will, through conditioning, acquire an aversive quality. Consequently, the mere thought of an action that commonly leads to harm, like pushing, hitting, or kicking, will become sufficient to produce negative affect. It is important to note that these aversions can be learned vicariously or through simulation and need not be learned first-hand, as evidenced in studies of Pavlovian conditioning outside the moral domain (Li, Delgado, & Phelps, 2011; Olsson, Nearing, & Phelps, 2007).

In contrast to Greene and colleagues' (2004) nativist proposal, Blair's model predicts that actions that do not involve personal force, like shooting a gun, can nonetheless become infused with negative affect by virtue of the fact that they are so often associated with harm. We found some support for this prediction in a recent study (Dillon & Cushman, in prep). Participants were asked to perform pretend versions of two classes of actions: those which might be relatively "typical" ways of harming someone in the real world (e.g. pointing a gun at someone and pulling the trigger, hitting someone in the foot with a hammer) and closely matched versions of these actions that were executed in a more atypical manner (though readily understandable and simple, e.g. pulling the trigger of the gun using a string, or dropping the hammer onto the person's foot using a rope and pulley). At the beginning of each trial, participants were carefully shown the relevant apparatus and asked to imagine performing the associated action. Participants exhibited greater increases in blood pressure (a physiological indicator of aversion) when anticipating performing the typical actions than the atypical actions. Because both conditions involved identical "imagined" harms, the greater aversiveness of the typical actions is likely due to sensorimotor properties of actions that have been routinely associated with harm in the past.

Obviously, however, actions like cutting a watermelon with a knife or taking a swing at a punching bag are minimally aversive, despite their similarity to stabbing or punching a person. Thus, the values attached to an action must be associated with particular environmental circumstances, or "states". This point is well-recognized in both formal and psychological models of reinforcement learning (e.g. Dayan & Niv, 2008; Sutton & Barto, 1998). When the state includes an apparent human target, stabbing and punching will possess different values than when the state includes a watermelon or punching bag target. Nevertheless, some of our experiments suggest that the identification of states can depend on superficial features of the environment rather than explicit knowledge of its structure. For instance, individuals report feeling averse to kicking a realistic-looking baby doll or stabbing a life-like mannequin even though the potential for harm is absent (Miller & Cushman, 2013).

Action Aversion and Third-Party Condemnation

We are now in a position to integrate the two dimensions of our hypothesis. First, we have seen evidence that our own behavioral aversions shape our moral judgments of third parties (evaluative simulation). Second, one source of behavioral aversion appears to be the intrinsic properties of harmful action (action aversion). Putting these ideas together, we should be able to show that individual

differences in action aversion predict moral judgments of third-party harm. Adopting this logic, Miller, Hannikainen, and Cushman (2013) developed a questionnaire to independently measure action aversion and outcome aversion. The 'action' items asked participants how much it would upset them to perform "typically" harmful actions that had been rendered harmless and socially acceptable, like stabbing a fellow actor in the neck with a fake knife during a play. These items were designed to gauge first-person action aversions divorced from concerns about harm. The 'outcome' items asked participants how much it would upset them to witness others endure pain, like a broken leg or cut finger. These items were used to assess outcome aversion without the confound of a potentially aversive action—and, predictably, they correlate well with a standard measure of empathic concern (Interpersonal Reactivity Index; Davis, 1980, 1983). Participants then judged several moral dilemmas, indicating the wrongness of a third party's decision to violently harm one individual in order to save the lives of several others.

Three notable findings emerged. First, most participants exhibited some degree of aversion to the 'action' items, confirming that simulated harmful actions still possess residual aversion when stripped of harm and placed in a socially acceptable context. Second, individual differences in action aversion were consistent predictors of the tendency to condemn harmful actions in personal moral dilemmas, even after controlling for outcome aversion, additional measures of empathy (IRI; Davis, 1980, 1983), and various demographic variables. This result provides support for the qualitative uniqueness of action aversion. If 'action' items were aversive solely because they triggered implicit representations of harmful outcomes, we would expect this variance to also be captured by outcome aversion and empathy. Rather, it appears that the relationship between action items and moral dilemmas is due to the surface properties of the aversive actions described in each. Third, outcome aversion was not itself a reliable predictor of moral judgment. Although interesting, this finding may have more to do with the idiosyncratic structure of personal moral dilemmas than with the irrelevance of outcome aversion: Because harmful outcomes occur regardless of which choice is made (e.g. either the one man dies, or the five people die), outcome aversion may not recommend either course of action over the other. Taken together, these results suggest that the process of judging third-party harmful behavior in the context of personal moral dilemmas involves asking yourself how you would feel performing the same behavior, and part of this feeling is best characterized as an aversion to particular features of the action.

To explore the generality of this effect, Miller, Hannikainen, and Cushman (2013) assessed the relationship between action/outcome aversion and the moralization of mercy killings. One advantage of examining mercy killings is that harm is limited to a single individual, affording a cleaner test of the importance of outcome aversion. Participants were assigned to one of three groups and asked to imagine twenty-three different methods of mercy killing that might be specifically requested by a dying individual, such as giving him pills, suffocating him with a pillow, and shooting him in the head with a shotgun. Those in the *action* group were told to imagine that they were actors in a movie, and that the mercy killing was therefore entirely fake and part of the movie plot. Furthermore, they were told to imagine that proper precautions had been taken and no real harm was possible. They were then asked to rate (1 to 10) how upsetting it would be to perform fake

versions of each mercy killing, thus targeting action aversion. Those in the *outcome* group were told to imagine that a third party was actually carrying out a real mercy killing and were asked to rate how much suffering they thought each method would cause, thus targeting outcome aversion. Finally, those in the *moral judgment* group were also told to imagine that a third party was carrying out a real mercy killing, and they rated how morally wrong it would be to agree to kill someone using each method. Ratings were averaged across subjects within each group to provide a mean rating for each item, and correlations between the groups were then computed. Unlike the previous study, *both* action aversion *and* outcome aversion were very strong, unique predictors of moral judgment, together explaining approximately 70% of the variance. Strikingly, action and outcome ratings did not significantly correlate with each other—a result that would not be expected if action aversions were simply implicit or automatic representations of harm. In addition to providing clearer support for the importance of outcome-based concerns in moral judgment, these findings strengthen our confidence in the power of action aversion as a unique predictor, suggesting a robust relationship between first-person behavioral aversions and third-party condemnation.

We have reviewed evidence that (i) negative affect can become associated with superficial features of an action, (ii) this affect is triggered when one imagines performing the action, (iii) when someone else performs the action, the same aversion can be triggered through a process of evaluative simulation, and (iv) this contributes to moral condemnation of the third party's action. Because this model of moral judgment has been tested primarily in artificial, high-conflict dilemmas, it is important to consider its applicability to the broader moral domain. We speculate, for instance, that in simple cases of obvious moral transgression it may not contribute meaningfully. Imagine you hear that Bill murders John, an innocent stranger, in order to rob him. How morally wrong was Bill's action? It seems unlikely that you need to either engage in evaluative simulation or consult your affective state to answer this question. We have strong, explicit prohibitions against murder, and being aware of these prohibitions is sufficient to judge the action as extremely wrong. This may explain why, despite substantial emotional deficits, psychopaths are often able to make normal moral judgments. Thus, it may be that both evaluative simulation and action aversion (or, for that matter, outcome aversion) play their largest role in moral judgment when the status of a potential transgression is uncertain.

Conclusion and Future Directions

We have drawn attention to a gap in the literature on emotion and moral judgment, and have proposed a way to fill it. Specifically, we have argued that the negative affect associated with moral condemnation may be differentiated according to two potential sources: aversive features of the action and negative features of the outcome. Furthermore, we have suggested that action aversion is facilitated by evaluative simulation, and may play a larger role in the moral judgment of harmful actions than previously recognized.

This approach opens up many potential avenues of research. For instance, what factors moderate the impact of emotion on moral judgment? Research suggests that some individuals rely more on their intuitions and gut feelings than others (e.g. Pacini & Epstein, 1999); might action aversion more heavily influence moral judgment in those with an intuitive vs.

rational thinking style? How are action- and outcome-based value representations acquired and implemented in the brain? Researchers studying learning and decision-making outside of the moral domain have developed dual-process models that distinguish between the values of actions versus outcomes, and such models have provided a fruitful way of understanding both animal and human behavior (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Dayan & Niv, 2008; Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995); could these same models also provide a way of understanding action and outcome aversion in the moral domain (see Cushman, 2013)? As these questions are explored and answered, we hope to paint a more accurate portrait of the emotional systems that color human morality.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

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

Blair, R. . J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, *57*, 1–29.

Blair, R. J. R. (1993). *The development of morality* (Unpublished PhD). University of London, London.

Blum, L. A. (1994). *Moral perception and particularity*. Cambridge University Press.

Ciaramelli, E., Muccioli, M., Làdavas, E., & Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*, 84–92.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 17433–17438.

Cushman, F. (2013). *Action, outcome and value: a dual-system framework for morality and more*. Manuscript submitted for publication.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*, 2–7.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*, 1052–1075.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science*, *17*, 1082 –1089.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain* (1st ed.). New York: Putnam.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*, 525–556.

Davis, M. H. (1980). A multi-dimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113–126.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215.

Dayan, P., & Huys, Q. J. M. (2009). Serotonin in affective control. *Annual review of neuroscience*, *32*, 95–126.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 185–196.

Decety, J., & Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of Neurophysiology*, *108*, 3068–3072.

Decety, J., Michalska, K. J., & Akitsuki, Y. (2008). Who caused the pain? An fMRI investigation of empathy and intentionality in children. *Neuropsychologia*, *46*, 2607–2614.

Decety, J., & Porges, E. C. (2011). Imagining being the agent of actions that carry different moral consequences: an fMRI study. *Neuropsychologia*, *49*, 2994–3001.

Decety, J., Skelly, L. R., & Kiehl, K. A. (2013). Brain response to empathy-eliciting scenarios involving pain in incarcerated individuals with psychopathy. *JAMA psychiatry (Chicago, Ill.)*, *70*, 638–645.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, *23*, 197–206.

Dillon, K., & Cushman, F. (in prep). [Typical vs. atypical harm]. Manuscript in preparation.

Eslinger, P., Moll, J., & de Oliveira-Souza, R. (2002). Emotional and cognitive processing in empathy and moral behavior. *Behavioral and Brain Sciences*, *25*, 34–35.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making*, *13*, 1–17.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Reviews*, *5*, 5–15.

Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition & Emotion*, *4*, 269–288.

Gray, K., & Wegner, D. M. (2008). The sting of intentional pain. *Psychological Science*, *19*, 1260–1262.

Greene, J. D. (2008). The secret joke of Kant's soul. In *Moral psychology, vol 3: The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). Cambridge, MA, US: MIT Press.

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

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*, 517–523.

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

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105 –2108.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes,

self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry*, *52*, 449–466.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, *108*, 814–834.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). *Moral dumbfounding: when intuition finds no reason*. Unpublished manuscript.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*, 613–628.

Haller, J., & Kruk, M. R. (2006). Normal and abnormal aggression: human disorders and novel laboratory models. *Neuroscience & Biobehavioral Reviews*, *30*, 292–303.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hoffman, M. L. (1982). Development of prosocial motivation: Empathy and guilt. In N. Eisenberg (Ed.), *The development of prosocial behavior* (pp. 281–313). New York: Academic Press.

Hoffman, M. L. (1987). The contribution of empathy to justice and moral judgment. In N. Eisenberg & J. Strayer (Eds.), *Empathy and its development* (pp. 47–80). New York: Cambridge University Press.

Hoffman, M. L. (2001). *Empathy and Moral Development: Implications for Caring and Justice*. New York: Cambridge University Press.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*, 963–976.

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

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*, 521–537.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*, 708–714.

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

Li, J., Delgado, M. R., & Phelps, E. A. (2011). How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences*, *108*, 55–60.

Liberman, N., Trope, Y., & Stephan, E. (2007). Psychological distance. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles (2nd ed.)* (pp. 353–381). New York: Guilford Press.

Lieberman, D., & Lobel, T. (2012). Kinship on the Kibbutz: coresidence duration predicts altruism, personal sexual aversions and moral attitudes among communally reared peers. *Evolution and Human Behavior*, *33*, 26–34.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*, 267–286.

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*, 193–197.

Merritt, A. C., & Monin, B. (2011). The trouble with thinking: People want to have quick reactions to personal taboos. *Emotion Review*, *3*, 318–319.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*, 143–152.

Miller, R., & Cushman, F. (2013). [Action aversion and moral judgment]. Unpublished raw data.

Miller, R., Hannikainen, I., & Cushman, F. (2013). *Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm*. Manuscript in preparation.

Murdoch, I. (2001). *The Sovereignty of Good. 1970*. London: Routledge.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*, 723–739.

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem." *Emotion*, *12*, 364–370.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*, 3–11.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*, 972–987.

Perkins, A. M., Leonard, A. M., Weaver, K., Dalton, J. A., Mehta, M. A., Kumari, V., … Ettinger, U. (2012). A Dose of Ruthlessness: Interpersonal Moral Judgment Is Hardened by the Anti-Anxiety Drug Lorazepam. *Journal of Experimental Psychology. General*. doi:10.1037/a0030256

Pizarro, D. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30*, 355–375.

Pizarro, D., Inbar, Y., & Helion, C. (2011). On Disgust and Moral Judgment. *Emotion Review*, *3*, 267–268.

Raine, A. (1996). Autonomic nervous system factors underlying disinhibited, antisocial, and violent behavior: Biosocial perspectives and treatment implications. *Annals of the New York Academy of Sciences*, *794*, 46–59.

Rozin, P., Haidt, J., & McCauley, C. R. (1993). Disgust. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 575–594). New York: Guilford Press.

Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion*, *11*, 233–240.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*, 513–523.

Starcke, K., Ludwig, A.-C., & Brand, M. (2012). Anticipatory stress interferes with utilitarian moral judgment. *Judgment and Decision Making*, *7*, 61–68.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). Cambridge University Press.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*, 1395–1415.

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*, 65–84.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*, 476 –477.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: Guilford Press.

Wilson, W. R. (1979). Feeling more than we can know: Exposure effects without learning. *Journal of Personality and Social Psychology*, *37*, 811–821.

Youssef, F. F., Dookeeram, K., Basdeo, V., Francis, E., Doman, M., Mamed, D., … Legall, G. (2012). Stress alters personal moral decision making. *Psychoneuroendocrinology*, *37*, 491–498.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*, 151–175.