Cognition 150 (2016) 232–242

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Can model-free reinforcement learning explain deontological moral judgments?

Alisabeth Ayars

University of Arizona, Dept. of Psychology, Tucson, AZ, USA

ARTICLE INFO

Article history:
Received 16 September 2015
Revised 1 February 2016
Accepted 3 February 2016
Available online 23 February 2016

Keywords:
Moral judgment
Model-free
Model-based
Dual-system
Reinforcement learning

ABSTRACT

Dual-systems frameworks propose that moral judgments are derived from both an immediate emotional response, and controlled/rational cognition. Recently Cushman (2013) proposed a new dual-system theory based on model-free and model-based reinforcement learning. Model-free learning attaches values to *actions* based on their history of reward and punishment, and explains some deontological, non-utilitarian judgments. Model-based learning involves the construction of a causal model of the world and allows for far-sighted planning; this form of learning fits well with utilitarian considerations that seek to maximize certain kinds of outcomes. I present three concerns regarding the use of model-free reinforcement learning to explain deontological moral judgment. First, many actions that humans find aversive from model-free learning are not judged to be morally wrong. Moral judgment must require something in addition to model-free learning. Second, there is a dearth of evidence for central predictions of the reinforcement account—e.g., that people with different reinforcement histories will, all else equal, make different moral judgments. Finally, to account for the effect of intention within the framework requires certain assumptions which lack support. These challenges are reasonable foci for future empirical/theoretical work on the model-free/model-based framework.

1. Introduction

Dual-system frameworks for explaining moral judgments have received much attention. These frameworks propose that there are two systems, such as a quick/automatic/intuitive/emotional system alongside a slow/controlled/rational system that govern our moral judgments (Greene & Haidt, 2002; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001). Dual-systems frameworks are invoked to explain a set of apparently conflicting moral intuitions involving sacrificing one person to save five others, known as the trolley problem (Foot, 1967). The trolley problem specifies that a runaway trolley is headed toward five people, who will all be killed if it strikes them. In the *switch* version, one can flip a switch that will divert the trolley to a different track in which it will kill only one person. In the *footbridge* version, the tradeoff is the same, except one must push a heavy man off a footbridge whose mass will stop the trolley in order to save the five. Most people think that it is permissible to flip the switch in *switch*, but wrong to push the man off the footbridge to his death in *footbridge* (Greene et al., 2001; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007; Mikhail, 2000; Petrinovich, O'Neill, & Jorgensen, 1993). This is in some sense puzzling, because in both cases, the utilitarian calculation is the same: one person can be sacrificed to save five others. Why then, is it permissible to flip the switch, but not to push the man off the bridge?

Dual-system frameworks propose an automatic emotional reaction to pushing the person off the *footbridge* explains why *footbridge* is deemed morally worse (Greene, 2007; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001). For instance, Greene proposes that because pushing a man is a personal, up-close, and direct form of harm, it triggers a negative affective reaction that subverts the rational, controlled cognition involved in reasoning that saving five lives is worth sacrificing one. In *switch*, because the act involved (flipping a switch to divert the train) is non-emotional, the utilitarian calculus dominates. Keeping with the literature, I will label the choice to sacrifice one person to save five others "utilitarian," and the refusal to sacrifice a person for this end "deontological."

Recently, Cushman (2013) argued that the distinction between emotion and controlled cognition is too crude and fails to capture certain features of moral attitudes. For instance, both the utilitarian and deontological choices carry affective weight. People do not merely calculate that five lives is *greater* than one, but that saving five people is *better* than saving one (Cushman, 2013). The

E-mail address: alisabetha@email.arizona.edu

http://dx.doi.org/10.1016/j.cognition.2016.02.002
0010-0277/

conflict arises in virtue of the value we place on one outcome (i.e., saving five lives) versus the disvalue we place on certain actions (i.e., pushing the man off the bridge). The proper distinction is not between emotion and controlled cognition, but rather, different targets for valuation: outcomes versus actions.

In light of these considerations, Cushman proposed a new dual-system framework. The framework exploits a distinction from computational neuroscience between model-free and model-based reinforcement learning, two forms of reinforcement learning with different structural targets of valuation. Model-free learning attaches values to actions and is responsible for the deontological choice in *footbridge*. Model-based learning, which constructs a causal model of the world and maximizes reward based on expected outcomes, is responsible for the pull of the utilitarian choice in both dilemmas.[1]

The model-free/model-based framework is an exciting new proposal through which to analyze our psychological responses to moral dilemmas. The view has already made a significant impact (e.g., Chakroff, Dungan, & Young, 2013; Greene, 2014; Rand et al., 2014; Ross, Bartels, Bauman, Skitka, & Medin, 2009; Shenhav & Greene, 2014; Sloman & Lagnado, 2015). The theoretical statement of the view was recently awarded the Daniel M. Wegner Theoretical Innovation Prize for making the most innovative contribution in social/personality psychology in 2013. Given its impact and breadth, the proposal is worthy of careful attention and evaluation.

In this paper, I present three criticisms of the use of model-free reinforcement learning to explain deontological moral judgment. First, many actions that we consider aversive we do not consider to be wrong—the proposal is thus an incomplete account of moral judgment. Second, there is a paucity of evidence for central predictions of the reinforcement account—e.g., that people with different reinforcement histories will, all else equal, make different moral judgments, and that typical actions are judged to be worse than atypical actions. Finally, the model-free learning explanation for the influence of intention on moral judgment depends on certain key undefended assumptions.

1.1. Model-free vs. model-based reinforcement learning

In model-free reinforcement learning, the cognitive system encodes values for actions based on their history of leading to rewards or punishments. For instance, if a certain action (like pressing a lever) consistently leads to a reward (like food), the model-free system will come to value the action. And, if an action repeatedly leads to punishment, it will devalue it. Actions that are valued by the model-free system become intrinsically rewarding. (The opposite is true for actions devalued by the model-free system.) Hungry rats that are trained to press a lever for food pellets will, even after being sated such that they no longer desire food, continue to press the lever if the brain is lesioned such that the model-based system is prevented from guiding behavior (Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995).

Decision-making using model-free learning is simple: the organism simply chooses the action with the highest value from the array of actions under consideration. Model-free decision making can be thought of as "short-sighted" and, in some sense, "dumb." Because a model of the world is not employed, model-free decision making takes into account only the immediate array of action options, without regard to the further options a particular choice will make available.

Two forms of learning are utilized by the model-free system to update action values. One is prediction-error learning. In prediction error learning, the cognitive system adjusts the value for actions according to the discrepancy between predicted and actual reward. If there is a discrepancy between the reward expected upon undertaking a certain action and the reward received, the value for the action is adjusted according to a function. (Thus, it is not the reward per se that promotes learning, but the discrepancy between predicted and actual reward.) Much evidence for prediction error learning has been uncovered, implicating the involvement of dopaminergic neurons in the midbrain (Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Schultz & Dickinson, 2000). The second form of learning, temporal difference learning, explains how the value of actions can be influenced by outcomes that occur much later in time.

Importantly, model-free learning need not operate only over actions defined by their lower-level properties ("stabbing," "kicking," etc.). The model-free system can also attach values to actions defined abstractly by their consequences (Cushman, 2013). For instance, candidate actions for evaluation by the model-free system include "save a person" or "harm a person" (p. 281).[2]

Model-free learning can also be hierarchical—a point that will be relevant to my third criticism. Consider the task of making a sandwich (Cushman, 2013, p.281). Making a sandwich, says Cushman, "involves a nested sequence of hierarchically dependent goal-directed actions: putting cheese on the bread within making the sandwich, obtaining cheese within putting cheese on the bread, opening the refrigerator within obtaining cheese, and so forth" (p. 281). Model-free learning can treat the selection of subgoals as "actions" in the context of a superordinate goal (a "state"), and associate appropriate subgoals (e.g., "select goal: obtain cheese") with reward in certain goal-states (e.g., "goal: make sandwich").

The model-free system can acquire action representations not only from observing the consequences of one's *own* actions, but by observing the consequences of others' actions as well (Bellebaum, Jokisch, Gizewski, Forsting, & Daum, 2012; Cooper, Dunne, Furey, & O'Doherty, 2012).[3] This is a crucial element of Cushman's theory, since a great deal of the actions that we morally condemn we have never performed ourselves.

Model-based learning, in contrast to model-free learning, involves the construction of a causal model of the world. The model, which can be envisioned as a decision tree, contains information about the probability of entering into certain states conditional on choosing a particular action, and the actions (and rewards) made available by entering into each state. Model-based learning guides action selection by searching through the decision tree to compute the action sequence that is most likely

[1] A highly similar proposal was put forth by Crockett (2013), although my critique will focus on Cushman (2013). Despite that my critique will only engage Cushman (2013), the first two criticisms may be considered objections to Crockett's (2013) account as well. The third criticism is less applicable because Crockett puts forth a different explanation for the effect of intention on moral judgment than Cushman (2013).

[2] This element of Cushman's theory is important; it is required to explain why certain actions are deemed wrong even when they do not involve lower-level action properties that have been punished, e.g., killing a person by putting poison in his drink, in order to prevent the deaths of five other people. (Unlike pushing, putting a substance in a drink is not an action that plausibly had a bad reinforcement history. If the categorical prohibition against such actions is to be explained by negative model-free values, the values must be associated with a more abstractly construed action like [poison a person] or [kill a person]).

[3] As of now, the evidence for model-free observational learning is tentative. Prediction error in the dorsal striatum is thought to be associated specifically with the learning of associations between actions and rewards (Cooper, Dunne, Furey, & O'Doherty, 2012; Delgado, Jou, & Phelps, 2011; O'Doherty et al., 2004; Tricomi, Delgado, & Fiez, 2004). While some studies like Cooper, Dunne, Furey, and O'Doherty (2012) found prediction error signals in the dorsal striatum, others, like Burke, Tobler, Baddeley, and Schultz (2010) did not. Furthermore, as Cushman acknowledges, prediction error signals in the striatum for observational learning, when found, seem to be attenuated; Bellebaum et al. (2012) conclude that "the striatum and orbitofrontal cortex thus appear to link reward stimuli to own behavioural reactions and are less strongly involved when the behavioural outcome refers to another person's action" (p. 241).

to lead to the best outcomes (Crockett, 2013). Model-based learning is more flexible than model-free learning, because the model can be continuously updated (based on what the organism learns about its environment) and allows for far-sighted planning that will maximize overall reward. However, it is computationally expensive, because it must search over the entire space of possible action combinations to arrive at a decision.

1.2. The model-free/model-based framework and trolley intuitions

Model-free and model-based decision-making is plausibly connected to deontological and utilitarian judgment, respectively. Utilitarian judgment involves a calculation of the expected outcome of each action under consideration, and the value of these outcomes. For instance, in the *switch* dilemma, one calculates that flipping the switch will lead to the death of one person but save five others, and doing nothing will lead to the death of five people but save one. Given that one values lives, the greatest expected reward is obtained by flipping the switch. This fits with model-based learning in that it involves a search for the best possible action given one's goals. An advantage to analyzing the utilitarian choice as the product of model-based reinforcement learning rather than simply "controlled cognition" is that it explains the *affective* value of saving five lives over one (Cushman, 2013). If the pull toward the utilitarian option is the output of learning involved in reward-maximization, then the affective weight of the utilitarian choice is explained—a virtue of Cushman's account.[4] I will not discuss the involvement of model-based reinforcement learning the utilitarian choice further, since my critique will only address the involvement of model-free learning in deontological judgment.

Model-free decision making is prima facie well-suited to deontological judgment because deontological judgment is insensitive to the relative value of the states of affairs brought about by each possible choice, and considers only the intrinsic value of the choices at hand. With respect to *footbridge*, Cushman argues that the model-free system, oblivious to the expected (high-value) outcome of pushing the man off the bridge (viz., that five people will be saved), signals that the act of harming someone directly and intentionally has low value—since, in the past, it has tended to produce bad outcomes, like victim distress (Cushman, 2013, p. 283). This produces a visceral aversion to the action, prompting condemnation of *footbridge*. In *switch*, the action involved (flipping a switch in order to divert a train) has no negative model-free value attached to it, as this action generally does not lead to negative consequences: "a model- free system might not assign much value at all to flipping a switch because it does not typically lead to a negative outcome", Cushman, 2013, p. 282. Thus model-based valuation of the utilitarian choice dominates.

1.3. Evidence for the effect of the model-free representations on moral judgment

Cushman (2013) presents a wide array of evidence for his account. Cushman first describes an empirical result that supports the operation of model-free values in human cognition. People are averse to performing pretend harmful actions, such as shooting a person with a fake gun, even when these actions do not lead to bad consequences (Cushman, Gray, Gaffey, & Mendes, 2012; Hood, Donnelly, Leonards, & Bloom, 2010; King, Burton, Hicks, & Drigotas, 2007). In Cushman et al. (2012), participants were asked to pretend to perform various harmful actions (e.g., smacking a baby doll on a table, pointing a fake gun at an experimenter). These actions led to symptoms of aversive reactivity like vasoconstriction—despite the fact that participants were fully aware they were pretend. Additionally, the actions were more aversive to *perform* than to simply witness. This suggests that the reason for the aversion was not simply a function of imagining the typical outcome (which would occur both during performance and observation), but rather, of executing the actions. This result supports the existence of intrinsic aversion to certain actions that is independent of the expected outcome of the actions—the mark of model-free learning.

Then, Cushman defends the claim that personal aversion to performing actions is a basis for making moral judgments of third parties. This step is important, since the link between personal aversion to actions and moral condemnation of them is not obvious (the weakness of this link is, in fact, the subject of my first criticism). The authors propose that individuals employ a process of "evaluative simulation" when assessing others in which they imagine how much it would aggravate them to perform the same action, which effects moral judgment (see Miller & Cushman, 2013).

In support of this claim, Miller, Hannikainen, and Cushman (2014) developed two scales to dissociate action aversion from outcome aversion. The action aversion scale assessed aversion to actions divorced from their normally harmful consequences, such as stabbing a fellow actor in the neck using a fake knife during a stage play. The outcome aversion scale assessed aversion to outcomes, like seeing a football player break a leg during a football game. The action aversion scale, but not the outcome aversion scale, was predictive of non-utilitarian moral judgment in moral dilemmas. This supports the claim that moral judgment of third parties is partly grounded in personal aversion to performing such actions.

Additionally, the authors note that people judge victimless crimes, such as incest, to be morally wrong, even when these actions do not lead to bad outcomes (Graham, Haidt, & Nosek, 2009; Haidt, Koller, & Dias, 1993). There is some evidence that one's personal aversion to incest predicts condemnation of it (Lieberman & Lobel, 2012) (p. 276).

This does not exhaust the evidence for Cushman's proposal. But it is a large chunk, and will suffice for the present purposes.

1.4. Two components of the footbridge action, and their reinforcement history

How exactly does the action in *footbridge* acquire its negative model-free association? Cushman (2013) usefully specifies the precise features of the *footbridge* action that are thought to trigger its condemnation (compared to *switch*). These features are the personal force involved the status of the harm as intentional:

> Two features... have been repeatedly demonstrated to trigger deontological response. The first is the manner of physical interaction between the agent and the victim (Cushman, Young, & Hauser, 2006; Greene et al., 2009). When the agent directly transfers his or her bodily force onto the victim (as in the push case) this elicits reliably higher levels of moral condemnation than when no such transfer of "personal force" occurs (as in the switch case). The second is... the status of harm as a means to saving others versus a side-effect of saving others (Cushman et al., 2006; Foot, 1967; Greene et al., 2009; Mikhail, 2000; Royzman & Baron, 2002; Thomson, 1985).
>
> [Cushman, 2013, p. 274–275]

Cushman argues that peoples' aversion to both these features can be traced to their history of producing bad outcomes. With respect to personal force (viz., pushing), Cushman says:

[4] A potential complication here is that sometimes utilitarian-seeming responses actually appear to reflect a lack of emotional engagement or low levels of empathetic concern (see Bartels & Pizarro, 2011; Gleichgerrcht & Young, 2013; Kahane, Everett, Earp, Farias, & Savulescu, 2015).

> A model-free system might assign negative value to "pushing," for instance, because it typically led to negative outcomes such as harm to the victim, punishment to the perpetrator, and so on. That is, most of the time that a person has personally pushed another (e.g., on the playground) or has witnessed one person push another (e.g., in a movie), this action lead to negative consequences (282).

The second component of the *footbridge* action is that it is intentional. In the *switch* dilemma, the harming of the person on the tracks is not necessary for saving the five. On the other hand, in the *footbridge* version, the harm to the man on the bridge is the means to saving the five (since his death is required to stop the train). Intentionally harming other people in service of other goals, claims Cushman, also typically leads to negative outcomes like victim distress, and thus would be attached to a negative model-free value as well:

> . . . consider cases where harm is used as a means to an end. This requires the cognitive action "select subgoal: harm a person." A model- free system will associate the execution of the subgoal with subsequent rewards or punishments. Generally, executing "select subgoal: harm a person" leads to aversive outcomes such as victim distress, reprimand, and so forth. Thus, a model-free system will tend to associate negative value with executing subgoals of the form "harm a person." By contrast, it will tend not to associate negative value with executing subgoals of the form "divert a train" or, more abstractly, "divert a threat," "save several lives," and so on, because these subgoals are not typically associated with aversive outcomes.
>
> [Cushman, 2013, p. 283]

It is helpful to contrast Cushman's account of the effect of intention with a more traditional account. An alternative explanation for the effect of intention is that people are responding to an internally represented rule known as the Doctrine of Double Effect (Foot, 1967). The Doctrine of Double Effect specifies that sometimes it is permissible to bring about a harm as a side-effect (i.e., "double effect") even if the harm would be wrong to intend, if the harm is an unavoidable outcome of bringing about a good result, and the good result "outweighs" (as it were) the bad outcome. While this explanation has fallen out of favor, Nichols and Mallon (2006) argue that rules that reflect the Doctrine of Double Effect play at least some role in trolley judgments (see Nichols & Mallon, 2006, for empirical results supporting the role of rule-based representations).

Rule-based accounts face the burden of explaining how it is that people come to represent complex rules like the Doctrine of Double Effect, which is sensitive to not only whether a harm is intentional but also defines a tradeoff function specifying when foreseen-unintended harm is permissible. Given the presumably impoverished data available to children from which to glean these complex norms (see, e.g., Wright & Bartsch, 2008), the declaration that people represent such rules arguably commits one to the existence of a moral module in which the rules are innately specified (Dwyer, 1999; but see Nichols, Kumar, Lopez, Ayars, & Chan, 2015 for a contrary view). An advantage of Cushman's account is that it provides a way of accounting for features of moral judgment employing relatively uncontroversial features of human cognition such as (1) representations of harm (including whether the harm was intentional, whether it was caused by the agent or merely allowed, etc.) and (2) model-free and model-based learning algorithms. The minimization of the assumed features of cognition needed to account for judgment makes Cushman's account highly attractive—but as I hope to show, there are limitations to the current formulation.

2. Action aversion and moral judgment

Now that I have laid out Cushman's proposal in detail, I will turn to my concerns. My first concern involves the proposed link between action aversion and moral judgment. Cushman claims that the moral attitude toward *footbridge* is grounded in our own aversion to performing the *footbridge* action. If this is to be a complete explanation, it must be the case that negative model-free values (and the associated aversion) are sufficient to prompt moral judgment in cases like *footbridge*, even when the outcome produced by the action is good (i.e., five lives are saved). Of course, even if action aversion is not a complete explanation for the *footbridge* judgment and other intuitions, it still may be a partial explanation—a proposal I will turn to at the end of this section.

There is prima facie reason to doubt that mere aversion or disinclination toward an action produced by model-free learning can cause moral condemnation of it. The most direct source of skepticism regarding the role of model-free values/action aversion in moral judgment is simply that there are many actions that we find aversive but not morally suspicious. Doing errands that we previously found highly aversive. Returning to a place in which something unpleasant, embarrassing, or traumatic occurred in the past. Getting behind the wheel of a car again after having a car accident. Cutting ourselves off from an addiction. We may perform these actions for good reasons, but they are uncontroversially unpleasant. But, with respect to these unpleasant actions, there is no temptation to consider them *morally* wrong.

Even actions which are tremendously aversive as a result of model-free updating are usually not moralized. Consider the following scenario:

> Sam goes to his favorite coffee shop every morning for coffee. Sam has no expectation of anything dangerous occurring at the coffee shop; coffee shops are, after all, safe places. But one morning, a horrendous incident occurs. As Sam is sipping his cup of coffee and gazing out the window, an armed robber enters. The robber waves his gun in the air and threatens to shoot the patrons of the coffee shop. After demanding money from the cashier, the robber leaves.

In this scenario, Sam's model-free system should be extremely active. Sam made, after all, a huge prediction error. Sam predicted that the action "having coffee at a cafe" would be relatively rewarding—it turned out, however, to be extremely unpleasant. Sam's model-free system would plausibly (based on prediction error learning) adjust the value of this action downward, to the point which going to cafes is a significantly and intrinsically aversive to Sam from then on. In fact, Sam might never attend a cafe again. (If one is concerned that such a dramatic model-free adjustment could occur after a single incident, then simply tweak the example such that Sam encounters multiple incidents of armed robberies at coffee shops).

However, it seems unlikely that Sam, from then on, would *moralize* attending cafes. It would be an odd and unusual reaction, for example, for Sam to look with moral suspicion upon patrons of cafes, or feel guilt the next time he attended a cafe. It seems rather that Sam's aversion would produce or constitute a strongly negative but non-moral attitude about attending cafes.

The contention that attitudinal strength and extremity is insufficient for and distinct from *moral* conviction is supported by research by Skitka and colleagues indicating that not all strongly held attitudes—i.e., attitudes that are extreme, important, and held with a high degree of certainty—are moralized (Skitka, Bauman, & Sargis, 2005) Given the distinction between moral and non-moral attitudes, it is unclear why we should expect action aversion to prompt *moral* judgment, as opposed to strongly negative but non-moral attitudes toward the relevant actions.

Consider the harmless but extremely aversive actions performed by participants in Cushman et al. (2012). These actions included shooting a fake gun at an experimenter, drawing a (dull) knife against a victim's throat, or smacking a baby (doll) against a table. The participants experienced significant physiological responses to performing the actions, as per the researchers' hypothesis. By the authors' own lights, these are paradigmatic examples of actions encoded with a negative model-free value. However, it is surely not the case that participants judged the harmless actions to be *morally* wrong (since, of course, they did not cause harm).

One might contend that I have unfairly dismissed a range of observations that suggest that deontological judgment is indeed grounded in our own aversion to performing the actions involved. For instance, people morally condemn "victim-less" crimes like consensual sibling incest, even when it is specified that no harm follows from them (Haidt, 2001). However, the condemnation of victimless crimes like incest need not be the sole result of aversion to them. It is possible that moral condemnation of incest relies on the recognition that incest violates a moral rule. Indeed, incest violates certain ideals of purity common to many moral frameworks (Haidt & Graham, 2007). The response to incest may be derived (partially or wholly) from the detection of the purity violation it entails, rather than from simply an emotional reaction to the act.[5]

These considerations should make one suspicious of the notion that aversion or disinclination toward an action can be the whole story in explaining deontological judgments like *footbridge*. Most of the actions that we find aversive we would label *unpleasant* or even *bad*, but not *wrong*. Another way to put the objection is that model-free values are sufficient to prompt *non-moral* evaluations (e.g., "I dislike eating broccoli", "broccoli is icky"), but moral evaluations (e.g., "no one should eat broccoli", "it is wrong to eat broccoli") require something more—e.g., an additional factor that spurs the agent to interpret this aversion as indicative of the normative status of some act.

Consonant with Nichols and Mallon (2006), a plausible candidate for the additional factor required to prompt moral condemnation of an action is the recognition that it violates a moral norm. Consider a likely explanation for why the participants in Cushman et al. (2012) did not consider the harmless but aversive actions like smacking a baby doll against the table to be morally wrong: they identified the actions to be harmless in that context, and therefore permitted despite the existence of a rule prohibiting intentionally harming others (a rule which *would* prohibit the *mimicked* actions—e.g., shooting someone). In other words, because the actions did not *in fact* violate this (or any other) moral rule (despite being suggestive of actions that *would*), the negative model-free values attached to the actions were insufficient to prompt an all-things-considered judgment of moral wrongness, suggesting that the detection of a rule-violation is necessary to prompt moral judgment.[6] In support of this, Greene et al. (2009) found that the effect of personal force on trolley judgments does not emerge unless the harm involved is intentional. Whatever influence model-free values for personal force have on the *footbridge* judgment is apparently negated in the absence of the recognition that the actions *really were* instances of intentionally harming in these contexts.

Of course, that it matters for moral judgment in these contexts that an action is genuinely an instance of intentional harm does not in itself demonstrate that people represent a rule that categorically prohibits intentional harm. For instance, it may be—as Cushman claims—that the model-free values for personal force and intentional harm simply combine to produce judgment of *footbridge*. I submit, however, that this proposal fails as a complete explanation for the *footbridge* judgment and other deontological judgments because, as I have argued, action aversion—even extreme aversion—is insufficient to prompt moral judgment.

Even if detection of a rule violation is *necessary* to prompt moral judgment, this does not mean that it is *sufficient*. It might be that the detection of a rule violation must be accompanied by model-free aversion (or other emotional reaction) that corroborates the significance of the norm violation to prompt moral judgment. This view is similar to Nichols' view that distinguishing features of moral judgments (like belief in authority independence) emerge when the rules implicated are affect-backed (Nichols, 2002), with the addition that at least sometimes the affect involved is model-free action aversion. Applied directly to trolley intuitions, this hybrid theory would specify that the *footbridge* intuition is produced by the recognition that the action violates a prohibition against intentionally harming others, combined with a visceral aversion to the personal force involved derived from the negative model-free value for personal force or "pushing"—a proposal friendly to Cushman's intended project of demonstrating the involvement of model-free values in these judgments. Even if model-free learning is only a partial explanation of deontological moral judgments, this does not mean model-free values are unimportant to moral judgment nor diminish the significance of work supporting their involvement. It is important to keep in mind, however, that a major attraction of Cushman's account mentioned earlier—i.e., the reduction of the seemingly complex machinery required for moral judgment (i.e., a moral module) to uncontroversial features of human cognition (model-free values and representations of harm)—would be removed if rule-representations must be trafficked in.

It is also important to note that the concerns raised in the section regarding the sufficiency of the proposed emotional/intuitive/automatic system in prompting moral judgment are not specific to Cushman's account. They have been raised explicitly against other dual-system theories as well. For example, Nichols and Mallon (2006, p. 532) made similar criticisms against Greene's view that peoples' aversion to the "personal-ness" of the harm in *footbridge* prompts their moral condemnation of it.[7]

The criticisms in the subsequent sections, however, apply more narrowly to Cushman's proposal, because they focus on the proposed distal source of these deontological moral judgments: model-free *learning*. It is to these criticisms I now turn.

3. Personal force and reinforcement history

A reinforcement learning account of deontological moral judgment predicts that, all else equal, moral judgments will vary

[5] Condemnation of victim-less crimes is not the only evidence Cushman provides. Miller et al. (2014) found that one's own aversion to performing aversive but harmless actions, as measured by an action aversion scale, is predictive of deontological judgment. However, there is reason to interpret this result cautiously. The relation between action aversion and moral judgments was only correlational. Thus, it is possible that an additional variable—a "third variable"—could explain both judgments of wrongness and action aversion. For instance, people who are sensitive to or condemning of violations of deontological moral rules such as "Do not harm" or "Do not kill" may be more likely to rate actions that violate these rules as morally wrong, *and* to be averse to actions associated with these rule violations (such as stabbing, shooting, or cursing) as assessed by the action aversion scale. This would explain the correlation between action aversion and judgments of wrongness.

[6] Note that it is unlikely that it was merely the recognition that the actions did not produce a bad outcome (as opposed to that they did not violate a moral rule) that preempted judgment: if satisfactory outcomes were sufficient to negate the effect of model-free values on judgment, then we should expect people to judge the action in *footbridge* to be permissible in virtue of its good outcome.

[7] Specifically, the authors note that "some acts of self-defense, war, and punishment are plausibly personal and emotional, but regarded as permissible nonetheless. For instance, many people think that spanking their own child is permissible, even though it is obviously personal and emotional" (Nichols & Mallon, 532).

according to different model-free reinforcement histories—at least according to the component of the history that is proposed to influence these judgments. Is this prediction borne out? In this section I discuss a few reasons to worry that it is not.

It is worth observing at the outset that condemnation of the *footbridge* action is robust; about 90% of people think that the act in *footbridge* is impermissible (Hauser, Young, & Cushman, 2008; although see Lanteri, Chelini, & Rizzello,2008 for a lower figure), and it is immune to many context effects that other moral intuitions (like *switch*) are subject to (e.g., Lanteri et al., 2008). If this is to be explained by model-free learning, it must be plausible that such a large percentage of people have undergone the requisite reinforcement history with personal force. Furthermore, it must be plausible that the necessary conditioning is undergone early in development; children as young as 3 condemn *footbridge* (Pellizzoni, Siegal, & Surian, 2010). That the vast majority of children have undergone the specified reinforcement history by this age may strike many as dubious.

This worry is enhanced by considering the specific component of the reinforcement history thought to make a difference. Cushman identifies the factor of personal force as particularly relevant to the *footbridge* judgment. When the agent in a trolley dilemma directly transfers her bodily force onto the victim, as in *footbridge*, this invokes higher levels of moral condemnation than when no such transfer of bodily force is employed, as in the *switch* case (Greene et al., 2009). According to Cushman, this can be explained by the fact that personal force is associated with negative consequences in the reinforcement history:

> A model-free system might assign negative value to "pushing," for instance, because it typically led to negative outcomes such as harm to the victim, punishment to the perpetrator, and so on. That is, most of the time that a person has personally pushed another (e.g., on the playground) or has witnessed one person push another (e.g., in a movie), this action lead to negative consequences (282).

Prima facie, we might expect variation in this component of the reinforcement history considering variation in the amount of violence children are permitted to watch on television and movies, the extent to which they are exposed to force at home and at school, and their own tendencies toward violent behavior.

Of great value would be to actually examine the moral intuitions of individuals with different reinforcement histories with personal force. A clear prediction from the reinforcement learning account is that, all else equal, people with a "worse" reinforcement history with personal force (e.g., for who have engaged in it with great frequency, or for whom it has resulted in particularly negative consequences) should be more condemning of *footbridge*.

I will discuss two reasons to worry that the prediction is not borne out, by providing examples of populations that plausibly differ in reinforcement history but not patterns of deontological judgment. The first comparison is between males and females, who plausibly differ in their experience with personal force. For females, pushing (and other forms of personal violence), even as schoolchildren, is less normative than for males. Boys are much more likely to use physical means and inflict physical pain as children than girls, a finding that is robust across cultures (Lansford et al., 2012). Additionally, regarding the observation of violence in movies, television, and video games, boys spend more time playing violent video games (Gentile, Lynch, Linder, & Walsh, 2004) and watching violent television than girls. Given the gender difference in experience with and observation of violence, one might expect a gender difference in *footbridge* intuitions if moral attitudes toward *footbridge* depend on reinforcement history, but a gender difference has not been found. In a study of 8778 participants who volunteered to respond to moral dilemmas online, Banerjee, Huebner, and Hauser (2010) found a slight gender difference in the opposite direction, with men giving slightly more utilitarian judgments than women in *footbridge*-style dilemmas.

Because there is no random assignment to gender, there might be other differences between the populations or countervailing pressures (e.g., innate tendencies or cultural factors) that can account for why the difference in reinforcement history does not lead to different moral responses—for instance, perhaps girls are innately more averse to violence and therefore are both inclined to condemn personal force (as in *footbridge*) *and* refrain from participating in it. However, this is just to say that for girls the *footbridge* intuition does not rely on model-free learning but instead on innate dispositions—which arguably displaces the reinforcement account itself. Proposals to account for data potentially inconsistent with the account must be cautious not to undermine the very claims the account makes about the source of deontological judgment.

The second comparison is individuals of different cultures and socioeconomic backgrounds. Little or no cultural differences in trolley intuitions have been obtained- people from all cultures, religions, and educational levels appear to think *footbridge* is morally wrong (Banerjee et al., 2010). People from different cultures, religious backgrounds, and educational levels differ systematically in their exposure to instances of personal force. For instance, people of low socioeconomic status are exposed to significantly more violence than people of high socioeconomic status (e.g., Browne, Salomon, & Bassuk, 1999). The absence of cultural/socioeconomic difference is also therefore suspicious if the *footbridge* intuition relies on reinforcement learning.

The lack of gender or cultural difference provides worry to think the predictions of the reinforcement account will fail to be realized. An additional concern is that many of the predicted results are counterintuitive. For instance, if the *footbridge* judgment depends on having used or witnessed personal force to bad effect, then all else equal, then people without the requisite reinforcement history—e.g., children raised in a particularly sheltered environment who were not allowed to watch television, attended violence-free schools, etc.—would be expected to fail to condemn *footbridge*. But it does not seem that people or children who have had minimal exposure to violence would simply find the *footbridge* action to be permissible, although this has not been examined.

One possible "catch-all" response to the concerns I have raised is to specify that the threshold of conditioning required to obtain the necessary model-free value (for, e.g., personal force) is quite low, such that the vast majority of children undergo it through minimal and unavoidable exposure to television or other cultural influences. This could explain why *footbridge* and other similar dilemmas are so widely condemned. Evidence for a "minimal threshold" of reinforcement learning required to influence moral judgment would be highly beneficial to the reinforcement learning account.

I will now switch gears to briefly address another prediction of Cushman's account involving reinforcement history: the effect of "typicality." Cushman's account predicts systematic differences in the extent to which actions are condemned according to the degree that they are *typically* harmful, since typically harmful actions would be more strongly linked to negative outcomes. This point is acknowledged by Cushman:

> . . .a key prediction of the current proposal is that typically harmful acts (e.g., pushing a person with your hands) will be considered morally worse than atypically harmful acts (e.g., pushing a person with your buttocks), even when the degree of physical contact and direct transfer of bodily force are equated.
>
> [Cushman, 2013, p. 282]

It is plausible that pushing the man off the bridge using one's buttocks would be viewed as more permissible than standard pushing. However, this example is confounded by the fact that the scenario invoked is humorous. A small piece of counterevidence to Cushman's claim is that Greene et al. (2009) found that people are just as likely to think that pushing the person off the bridge is impermissible if it is performed using a long pole (Greene et al., 2009). Pushing using a long pole is clearly atypical.

Following the line of thought that typicality of harmful consequence matters, a natural prediction is that actions involving a *high* degree of force will be more condemned than actions that involve only a small degree of personal force. Light force (like the kind required to playfully nudge a friend) typically results in less injury or distress signals (if any) than violent force, such as that utilized by an abusive partner; therefore, model-free values would be more severe for highly forceful actions than less forceful ones. Congruent with the claim, given that harmless actions like pointing a fake gun or smacking a realistic doll against the table are aversive at all (Cushman et al., 2012), the aversion would likely increase as the amount of force employed increased (e.g., from lightly hitting the doll to smashing it against the table). However, there is some reason to doubt this prediction about the degree of force. Imagine that the heavy man happens to be standing on the edge of the bridge making his position quite precarious (but nevertheless stable in the absence of intervention), requiring only the lightest of nudges to topple him over. Does lightly pushing him off, to save the five, seem any less morally bad? It seems not, although of course this cannot be answered definitively without a more extensive empirical investigation.

To summarize this section, a challenge for the reinforcement learning theory is to account for convergence in moral intuitions such as *footbridge*, especially among populations that plausibly differ in reinforcement history (e.g., women and men). Furthermore, the reinforcement account commits itself to predictions that there is prima facie reason to think will not be borne out: e.g., degree of force effects. These considerations are not conclusive, but are worth bringing to the table.[8]

4. Harming as a means vs. a side-effect

My third criticism concerns Cushman's account of the influence of intention on moral judgment. I argue that Cushman offers an incomplete explanation for the effect of intention on moral judgment: that the goal [harm a person] is more likely to lead to aversive outcomes like victim distress and reprimand than the goal [divert a train]. This is an incomplete explanation for the means/side-effect distinction, which specifies a moral difference between intended harm and foreseen-unintended harm. I will consider two ways that Cushman could "complete" the explanation, both of which are subject to concern.

A few words about intention and moral intuitions. In *footbridge* (and other cases that are usually condemned, like *transplant*) one intentionally harms a person; the harming is intentional because the victim is a *means* to saving the others (e.g., by stopping the trolley). In *switch*, the harming is merely a side-effect of diverting the train (Foot, 1967). It is important to note that intention is crucial to explaining trolley intuitions. The effect of personal force on trolley judgments does not emerge unless the harming is intentional (Greene et al., 2009). Furthermore, intention affects moral judgment independently of lower-level action properties—that is, the effect of intention does not emerge simply in virtue of intentional harm being more likely to be associated with aversive lower-level actions like stabbing, shooting, or pushing (even though in the *switch*/*footbridge* cases, it happens to be). For example, bombing civilians as a means is seen as morally worse than bombing civilians as a side-effect of bombing an enemy, even though the lower-level action ("bombing") is the same in both cases. A theory that purports to explain the contrasting moral intuitions in these dilemmas must explain why intention *alone* can make a difference.

Cushman situates his discussion of intention in the context of hierarchical reinforcement learning. In hierarchical reinforcement learning, the model-free system learns to select subgoals that best serve superordinate goals. With respect to intention, Cushman notes that only the subgoal in *footbridge* ("harm a person") would be associated with a negative action value.

> Generally, executing "select subgoal: harm a person" leads to aversive outcomes such as victim distress, reprimand, and so forth. Thus, a model-free system will tend to associate negative value with executing subgoals of the form "harm a person." By contrast, it will tend not to associate negative value with executing subgoals of the form "divert a train" or, more abstractly, "divert a threat," "save several lives," and so on, because these subgoals are not typically associated with aversive outcomes.. [Cushman, 2013, p. 283]

This explanation skips over an important structural feature of the trolley cases. It is true that the subgoal [harm a person] is more likely to lead to negative outcomes like victim distress and reprimand than the goal, [divert a train]. But why should we think that model-free reinforcement learning (or even hierarchical reinforcement learning) operates only over goals (and subgoals)? Even though the harming is not a goal in the switch case, one still ends up causing harm to a person in virtue of flipping the switch and diverting the train. It is true that this harm is a side-effect in the switch case, rather than a means. But it is harm nonetheless, and people clearly represent this unfortunate consequence of diverting the train. (I.e., it is clear to all readers of the switch case that if the switch is flipped, a person will be killed). The act [harm a person] surely leads to negative outcomes, even if the harming is not the goal. We should be skeptical, then, that the decidedly worse reinforcement history of the goal [harm a person] compared to the goal [divert a train] definitely settles the issue. The comparison of interest is whether harming a person as a goal has a worse history than performing actions one *foresees* will bring about harm to a person.

To put the objection more succinctly: call the act in footbridge [harm-intend]. The act is labeled such because it involves intentionally bringing about the consequence of harming a person. Call the act in switch [harm-foreseen], since it involves performing an act that one foresees will harm a person. Cushman reasonably claims that [harm-intend] is worse (i.e., has a worse reinforcement history) than [divert a train-intend], the two actions depicted in the narrow intention box in Fig. 1. However, the comparison of these actions is insufficient to account for the means/side-effect distinction. To account for the means/side-effect distinction in terms of model-free learning, it must be the case that [harm a person-intend] has a worse reinforcement history than [harm a person-foreseen]. It is not nearly as obvious that [harm a person-intend] results in worse outcomes than [harm a person-foreseen]—after all, by definition, both result in harm of a person!

One strategy to account for why [harm-intend] is judged more harshly than [harm-foreseen] is to specify that model-free reinforcement learning—or at least, one mechanism of model-free reinforcement learning (e.g., hierarchical reinforcement learning) only

[8] Another possible issue related the congruency of Cushman's account and the empirical data is the apparent context-dependency of moral intuitions (e.g., Bartels, 2008; Rai & Holyoak, 2010; Shallow, Iliev, & Medin, 2011). These factors represent major influences on moral judgment, but are "surface level" manipulations—i.e., they do not change the learned associations with actions. It is therefore unclear how the model-free/model-based account can accommodate them. I'd like to thank an anonymous reviewer for this suggestion.

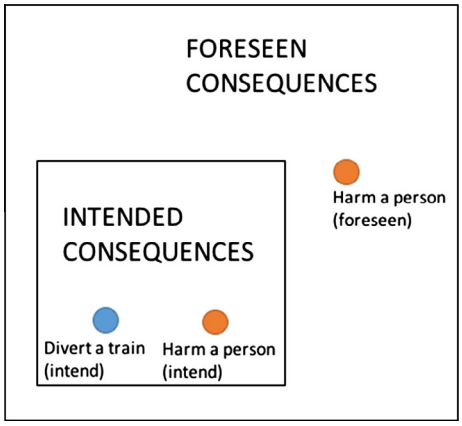

Fig. 1. The nested structure of consequence types. Intended consequences are nested within foreseen consequences, since the vast majority of intended consequences are foreseen, but not all foreseen consequences are intended. The two orange dots both represent the action "harm a person," but the consequence is intended for the inner-box action but merely foreseen for the outer-box action. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

operates over intended (narrow box) actions. Cushman seems to align himself with this proposal in the following quote: "A system that represents action plans in terms of their hierarchical structure must necessarily represent switch-flipping in terms of "harming a person" in the means case, whereas it can merely represent switch-flipping in terms of "diverting a train" in the side-effect case" (Cushman, 2013, p. 283).

Let us consider this proposal: that model-free system defines actions—at least in some contexts—only over their *intended* consequences. It is worth noting that Cushman and colleagues explicitly reject the (stronger) proposal that model-free learning *cannot* operate over foreseen-consequence defined actions. Miller et al. (2014) argue that unintended but foreseen harming specifically constitutes an action that has a model-free representation attached to it[9]:

> Through [a] process of conditioning, negative affect can become associated with actions that are essentially defined in terms of their goals *or foreseen consequences* [emphasis mine] (e.g., "murder," or "doing harm") and not simply with specific physical movements or motor plans (e.g., shooting a gun, thrusting a knife).
>
> [Miller et al., 2014, p. 580]

Indeed, this contention—i.e., that habit learning can operate over actions at the foreseen scope—accords with the authors' empirical results (e.g., Miller et al., 2014).[10] It does not appear, then, that proponents of the reinforcement learning account find it plausible that model-free values are not assigned to actions defined over their foreseen-unintended consequences tout court, consistent with Miller et al.'s findings. Another option is still available: that *some* mechanisms of model-free learning define actions over goals but not foreseen side-effects. Hierarchical reinforcement learning—the mechanism which the authors specify to assign values to sub-goals—may be just such a mechanism. If hierarchical reinforcement distinguishes between means and side-effects, and people can be shown to possess this discerning mechanism of model-free value assignment, this would support the author's grounding of the means/side-effect distinction in model-free learning.

A recent publication—"Habitual Control of Goal Selection in Humans" Cushman and Morris (2015) may provide evidence that bears on this issue. The title of the publication suggests that the authors may have uncovered a mechanism of model-free value assignment operative in cognition that distinguishes between goals and foreseen side-effects. Although this publication does not explicitly connect habitual goal selection to Cushman's account of deontological moral judgment, the experiments may provide reason to think that in these contexts model-free reinforcement learning operates specifically over goal-defined actions. It is therefore worth considering these experiments in detail.

In the experiments described, participants performed multistep choice paradigms in which certain actions at one stage resulted in a common outcome at the subsequent stage—e.g., selecting either the numbers "1" or "3" at stage 1 resulted (with .8 probability) in "blue" at stage 2 of the paradigm (experiment 1). Thus, the actions were linked by a shared expected outcome at stage 2. When the selection of a stage-1 choice like the number "1" resulted in an unexpected low-probability (.2) transition to a high reward "green" state at stage 2 rather than the expected state, participants were more likely to choose the *alternative* stage 1 action (i.e., the selection of "3") with the same *expected* outcome ("blue") on the next trial. The authors take this to indicate that model-free values can be linked to actions defined by a shared goal (i.e., "get to blue"), as opposed to more concrete actions (i.e., the selection of a particular integer). The results of these experiments are important because they militate against the traditional view that model-free learning operates only over concrete or immediate actions like "push red button" or "select number 3."

Despite the significant implications of the experiments, I submit that that this work does not provide convincing evidence of an operative mechanism that assigns model-free values to goals but not foreseen side-effects. The experimental results can be explained model-free value assignment to actions defined over their *foreseen side-effects*, rather than goals. Transition to the blue/red states in experiment 1, e.g., were both goals of the relevant stage 1 actions *and* foreseen side-effects of those actions. That the actions were defined over these outcomes (which the experiments clearly show) does not distinguish between the possibilities that this definition occurred in virtue of the outcomes being the goals of the actions, or in virtue of the outcomes being foreseen side-effects of the actions. As far as can tell, in none of the experiments was there a condition in which the actions were foreseen but *un*intended.

Importantly, the authors do rule out an alternative explanation for the results: that model-free values were simply defined over stage 2 resultant *states* (e.g., "blue") rather than outcome defined actions (e.g., "get to blue"). However, this does not preclude the possibility that values were attached to actions defined by their foreseen consequences. A foreseen-consequence defined *action* is conceptually distinct from an outcome (i.e., the resultant state of the action)—a distinction endorsed by Miller et al., 2014, p. 580). Because such a distinction is available, ruling out that values were attached to outcomes does not rule out that values were attached to actions defined over those foreseen outcomes.

Thus I submit that the first strategy—that the model-free system only defines actions over their intended consequences—is currently of minimal evidential support, although this evidential

[9] Miller et al. offer this as an explanation for why scores on their action aversion scale also predict moral judgment in switch-style cases involving impersonal harm as well as footbridge-style cases. The authors say, "Why should responses to 'impersonal' actions, which lack aversive surface properties, correlate with action subscale items that possess those surface properties but do not involve doing actual harm? The simplest answer is that individuals might differ in their general sensitivity to action-

[10] The authors found that their action aversion scale, which assessed aversion to harmless (but usually harmful) actions like pretending to stab someone as part of a stage play, marginally predicted responses in switch-style cases involving harming as a side-effect (Miller et al., 2014). The action aversion scale measures personal aversion to action divorced from any harmful outcome. People who scored high on this scale were less likely to think it is permissible to flip the switch (or to perform other actions that result in harm as a side-effect) in *switch*-style cases. This result would be highly unlikely if harming in switch was not associated with a model-free value at all.

status may change as research progresses. What's needed for Cushman's project is an experiment that demonstrates a context in which model-free values are assigned to actions defined by a common outcome when that outcome is a *goal* (or a subgoal of a larger goal) but not a salient foreseen side-effect.

However, there is also an alternative strategy. The alternative strategy is to contend that [harm-intend] is typically associated with worse consequences than [harm-foreseen], and thereby is more likely to be associated with negative value than [harm-foreseen]. This is the logic Cushman uses to explain the greater aversion to [harm a person-intend] than [divert a train-intend] (the two narrow box actions in Fig. 1).

One might worry from the outset whether it's plausible that the model-free system would treat [harm-intend] and [harm-foreseen] as separate action types with distinct model-free values rather than simply subsuming them under the single action of "harming." It would be detrimental and bizarre for the action-types to be updated independently, rather than allowing for "transfer" of values between the two. (E.g., it would be maladaptive for one's negative model-free value for [touch fire-intend] to fail to demotivate [touch fire-foreseen], e.g., when one reaches into a fire to retrieve a piece of food that fell in).

But let us set aside this worry for the moment. Assuming the model-free system attaches differential action values to intentional and unintended-foreseen harming, is it plausible that the reinforcement histories of these two action types differ for the majority of people, such that intentionally harming would be coded with a lower action value? When we decide to harm someone as our *goal*, does this generally result in worse outcomes than when we decide to harm someone as a *side-effect*? There is reasons to think this is implausible.

Cushman cites outcomes like victim harm and distress as being aversive. The proposal that victim harm and distress is worse for intended vs. foreseen-unintended harm is arguably a nonstarter with respect to actions defined by their harmful consequences. [Kill a person-intend] and [kill a person-foreseen]—the actions in footbridge and switch, respectively—by definition have the same consequence (a person is killed in each case). After all, one puzzle highlighted by the trolley dilemma is that intentional harm is seen as worse than foreseen-unintended harm *even when their outcome is exactly the same*. This applies mutatis mutandis to all ([___-intended], [___-foreseen]) action pairs (e.g., [suffocate-intend] and [suffocate-foreseen]; [starve-intend] and [starve-foreseen], etc.). Suppose we identified all possible ([___-intended], [___-foreseen]) pairs. Random sampling of intended and foreseen-unintended actions would produce no systematic differences in the victim harm implied, since the harms (specified by the "___" of each ([___-intended], [___-foreseen]) pair are precisely equivalent across the two boxes.

Perhaps this analysis is too quick. Within an abstractly-construed action type (like "harm"), there are "sub-consequences"—e.g., whether the specified harm produces a great deal of pain or only a small degree of pain. Gray and Wegner (2008) found that intentional harm is in fact reported to be more painful than unintentional harm—an intriguing and surprising result. However, it is important to note that the study distinguished intentional from accidental harm, not foreseen-unintended harm, mitigating its relevance to the means/side-effect distinction. While it may turn out that there is a difference in the amount that intentional and unintended harm, e.g., hurts, to rest an account of the means/side-effect distinction on this controversial assumption is at best perilous.

But victim harm and distress are not the only aversive consequences of harming others. Cushman also mentions punishment by authority as an aversive outcome of harming. Might punishment be significantly more severe for narrow box actions like [harm-intend] than for wide box actions like [harm-foreseen]? Clearly, whether a harm is intentional makes a moral difference at least sometimes (such as in trolley cases), and on those grounds we might expect [harm-intend] and [harm-foreseen] to be accompanied by different amounts of reprimand and punishment. It is intuitive and plausible that, say, a government that bombs civilians intentionally would be subject to more backlash than a government that bombs civilians as a necessary side-effect of bombing an enemy.

This proposal—that people are more averse to intentional harm than unintentional harm because intentional harm is more likely to be punished—appears promising. But unfortunately, I think it is subject to a significant worry. To put the concern succinctly: Conditions in which a foreseen harm that would be wrong to intend is excused are *rare*.

A standard characterization of the Doctrine of Double Effect specifies that an action that has a foreseen effect that would be prohibited if intended is permissible only if:

1. the intended action is permissible
2. the foreseen bad outcome is not intended
3. there is no way to produce the good outcome without also producing the bad outcome
4. the bad outcome is not disproportionate to the good outcome (see, e.g. Uniacke, 1998, p. 120).

Conditions 1–4 are met in the *switch* trolley case, but they rarely characterize real-life instances of unintended-foreseen harm. It is true, of course, that if an action with a bad outcome did meet the conditions specified, it would be punished less than analogous instances of intending the bad outcome. But because conditions 1–4 are rare, this hardly justifies the claim that foreseen harm is *in general* punished less than intended harm. The vast majority of foreseen-unintended harms that would be wrong to intend are *not* excused by the principle (e.g.; grabbing a buried item from a shelf knowing that doing so will topple and break the other items on the shelf). Such actions would be met with exclamations of disapproval and, in many cases, punishment. It is highly unlikely that the child's reinforcement history would provide any basis for a greater aversion to, avoidance of, or disapproval of intended compared to foreseen-unintended harm.

A short review of this section is in order. Cushman offers an incomplete explanation for the effect of intention on moral judgment: that the goal [harm a person] is more likely to lead to aversive outcomes like victim distress and reprimand than the goal [divert a train]. In order for this to be a satisfactory explanation for the influence of intention, we need an additional premise: either that model-free reinforcement learning (at least in some contexts) does not operate over actions defined by their foreseen consequences, or that [harm-intend] is associated with worse consequences (and therefore a more negative model-free value) than [harm-foreseen]. Both claims are subject to concerns.

With respect to the first claim, the current evidence of the operation of such a mechanism (e.g., hierarchical learning distinguishing between goals and side-effects) is minimal. The second claim is undermined by the fact that victim harm is (by stipulation) equivalent between pairs of ([__-intended], [__-foreseen]) actions in which the blank specifies a particular (harmful) consequence, and to specify that the difference lies in the "sub-consequences" of the harm (e.g., the extent to which intended and unintended pain "hurts") is a perilous assumption on which to rest the account. Situating the difference in the typical punishment is also hazardous—conditions in which harm is punished more if it is intended than if it is merely foreseen are rare. The occasional exemption of foreseen-unintended harm from moral censure in accordance with the Doctrine of Double Effect is a weak basis on which to rest a case for differential reinforcement for intended and unintended harm.

All-in-all, one should be skeptical—given the current evidence—that the effect of intention on moral judgment can be explained by model-free learning. This is not to say that the possibility has been definitively out-ruled, but that we need a much stronger case before acceptance of this view is warranted.

5. Conclusion

In this paper I identified three concerns about Cushman's use of model-free/model-based reinforcement learning to explain the contrasting moral intuitions in *switch* and *footbridge*. My first concern is that action devaluation by the model-free system—experienced as aversion to performing the action—is insufficient to prompt moral judgment. Action aversion prompts judgments of the unpleasantness, dislike, or badness of an action, but not judgments of *wrongness*. Something else—like rule representations—is required.

My second concern involves the potential inconsistency of the reinforcement account with certain empirical results—e.g., the consensus that certain acts (like *footbridge*) are wrong, the short developmental period required for moral competence, the lack of population differences in judgment (e.g., between men and women) for whom reinforcement history plausibly varies, and an apparent lack of clear intuition that typical harms are morally worse than atypical harms.

Finally, I argued that any account of the influence of intention on moral judgment situated in model-free learning will be relatively assumption laden and is not warranted be current evidence. This element of the account is in need of clarification and defense.

These criticisms are not knock-down criticisms of the view, but they challenge the current formulation of Cushman's proposal. However, the model-free/model-based proposal is still in its infancy. Future work on the framework may successfully address the concerns I have raised. The three questions I consider of primary importance are:

- What is needed in addition to model-free values to prompt judgments of moral wrongness?
- If the *footbridge* judgment (and other moral judgments) is partially grounded in reinforcement history, what explains the convergence in the *footbridge* judgment and other deontological intuitions? Is there evidence that people with different reinforcement histories make different moral judgments?
- What aspect of model-free learning, precisely, explains the role of intention in moral judgment? Does intended and unintended harm differ in their typical consequences? Are model-free values assigned only to intended harm in some contexts? What, precisely, is the evidence for this, and how is the possibility that model-free values were assigned to foreseen-unintended consequence defined actions excluded?

The model-free/model-based framework is no doubt an innovative lens through which our moral responses can be viewed—an account with much potential promise in accounting for features of moral judgment. However, the proposals themselves, the assumptions on which the account rests, and its relation to other accounts of moral judgment (e.g., the rule-based account) are in need of significant exposition and defense before acceptance of the fundamental features of the proposal is warranted.

Acknowledgements

I would like to thank Fiery Cushman and an anonymous reviewer for their insightful and thorough comments, which resulted in a much-revised and much-improved paper. I would also like to especially thank Shaun Nichols for his extensive feedback on the ideas contained in this paper. Research for this paper was supported by Office of Naval Research Grant #11492159 to Nichols.

References

Banerjee, K., Huebner, B., & Hauser, M. (2010). Intuitive moral judgments are robust across variation in gender, education, politics and religion: A large-scale web-based study. *Journal of Cognition and Culture, 10*(3), 253–281.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108*(2), 381–417.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161.

Bellebaum, C., Jokisch, D., Gizewski, E., Forsting, M., & Daum, I. (2012). The neural coding of expected and unexpected monetary performance outcomes: Dissociations between active and observational learning. *Behavioural Brain Research, 227*(1), 241–251.

Browne, A., Salomon, A., & Bassuk, S. S. (1999). The impact of recent partner violence on poor women's capacity to maintain work. *Violence Against Women, 5*(4), 393–426.

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences, 107*(32), 14431–14436.

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PloS One, 8*(9), e74434.

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience, 24*(1), 106–118.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology Inc, 17*(3), 273–292. http://dx.doi.org/10.1177/1088868313495594.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*(1), 2.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences, 112*(45), 13817–13822.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

Delgado, M. R., Jou, R. L., & Phelps, E. A. (2011). Neural systems underlying aversive conditioning in humans with primary and secondary reinforcers. *Frontiers in Neuroscience, 5*, 71. http://dx.doi.org/10.3389/fnins.2011.00071.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior, 23*(2), 197–206.

Dwyer, S. (1999). Moral competence. In K. Murasugi & R. Stainton (Eds.), *Philosophy and linguistics*. Westview Press.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. In *Virtues and vices and other essays*. Berkeley, CA: University of California Press.

Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of Adolescence, 27*(1), 5–22.

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PloS one, 8*(4), e60418.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Update, 11*(8), 322–323.

Greene, J. (2014). *Moral tribes: Emotion, reason and the gap between us and them*. Atlantic Books Ltd.

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*(3), 364–371.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 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*(2), 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 (New York, N.Y.), 293*(5537), 2105–2108. http://dx.doi.org/10.1126/science.1062872.

Gray, K., & Wegner, D. M. (2008). The sting of intentional pain. *Psychological Science, 19*(12), 1260–1262.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116.

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*(4), 613.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21.

Hauser, M., Young, L., & Cushman, F. (2008). Reviving Rawls' linguistic analogy. *Moral Psychology, 2*, 107–143.

Hood, B. M., Donnelly, K., Leonards, U., & Bloom, P. (2010). Implicit voodoo: Electrodermal activity reveals a susceptibility to sympathetic magic. *Journal of Cognition and Culture, 10*(3), 391–399.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of information processing in the basal ganglia* (pp. 249–270).

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

King, L. A., Burton, C. M., Hicks, J. A., & Drigotas, S. M. (2007). Ghosts, UFOs, and magic: Positive affect and the experiential system. *Journal of Personality and Social Psychology, 92*(5), 905.

Lansford, J. E., Skinner, A. T., Sorbring, E., Giunta, L. D., Deater-Deckard, K., Dodge, K. A., ... Tapanya, S. (2012). Boys' and girls' relational and physical aggression in nine countries. *Aggressive Behavior, 38*(4), 298–308.

Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics, 83*(4), 789–804.

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*(1), 26–34.

Mikhail, J. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in a theory of justice*. PhD dissertation. Cornell University.

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass, 7*(10), 707–718.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion, 14*(3), 573.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 16*(5), 1936–1947.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition, 84*(2), 221–236.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*(3), 530–542.

Nichols, S., Kumar S., Lopez, T., Ayars, A., & Chan., H. (2015). Rational learners and moral rules. *Mind and Language* (in press).

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, N.Y.), 304*(5669), 452–454. http://dx.doi.org/10.1126/science.1094285.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*(3), 467.

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental science, 13*(2), 265–270.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science, 34*(2), 311–321.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*.

Ross, B. H., Bartels, D., Bauman, C., Skitka, L., & Medin, D. L. (2009). *Psychology of learning and motivation: Moral judgment and decision making*. Academic Press.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research, 15*(2), 165–184.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.), 275*(5306), 1593–1599.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23*(1), 473–500.

Shallow, C., Iliev, R., & Medin, D. (2011). Trolley problems in context. *Judgment and Decision Making, 6*(7), 593.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 34*(13), 4741–4749. http://dx.doi.org/10.1523/JNEUROSCI.3390-13.2014.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of personality and social psychology, 88*(6), 895.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology, 66*, 223–247.

Thomson, J. J. (1985). Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Yale Law Journal, 94*(6), 1395–1415.

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron, 41*(2), 281–292.

Uniacke, S. (1998). The principle of double effect. *Routledge Encyclopedia of Philosophy, 3*, 120–122.

Wright, J. C., & Bartsch, K. (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly, 54*(1), 56–85.