Crim Law and Philos
DOI 10.1007/s11572-014-9334-1

ORIGINAL PAPER

The Psychological Origins of the Doctrine of Double Effect

Fiery Cushman

Abstract The doctrine of double effect (DDE) is a moral principle that distinguishes between harm we cause as a means to an end and harm that we cause as a side-effect. As a purely descriptive matter, the DDE is well established that it describes a consistent feature of human moral judgment. There are, however, several rival theories of its psychological cause. I review these theories and consider their advantages and disadvantages. Critically, most extant psychological theories of the DDE regard it as an accidental byproduct of cognitive architecture. This may provide philosophers with some reason to question its normative significance.

Keywords Moral psychology · Doctrine of double effect · Cognition · Reinforcement learning

The doctrine of double effect (DDE) is an unlikely target for psychological research. First articulated by St. Thomas Aquinas (1988, p. 70), it was an orphan of Catholic doctrine for centuries before being adopted into the philosophical mainstream by Philippa Foot (1967). Even today it remains hard for novices to understand, hard for philosophers to justify, and hard for anybody to apply beyond decidedly contrived dilemmas. Moreover, while it does capture a detectable pattern of influence on ordinary people's moral judgments, the effect is Lilliputian even by the forgiving standards of social psychology. So it is remarkable that, despite these inborn disadvantages, the DDE has inspired a considerable body of research (Cushman and Young 2011; Cushman et al. 2006; DeScioli et al. 2012; Greene et al. 2009; Hauser et al. 2007; Lotto et al. 2013; Mikhail 2000, 2007, 2011; Royzman and Baron 2002; Schwitzgebel and Cushman 2012; Sinnott-Armstrong et al. 2008; Waldmann and Dieterich 2007).

F. Cushman (✉)
Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, MA 02138, USA
e-mail: cushman@fas.harvard.edu

Published online: 16 September 2014

Yet, some of these very disadvantages may be the secret to the DDE's appeal. All psychological theories of moral judgment predict that people will consider unjustified harm reprehensible, that they will excuse accidental behaviors, and that they will scorn others' hypocrisy while rationalizing their own ethical lapses. These clues are about as useful in identifying the psychological causes of our moral attitudes as a five-fingered glove print on a gun is useful in identifying the shooter. But the DDE is so unexpected, bizarre and subtle an effect that it holds the promise of a three-and-half-fingered grip: a theory that explains it cannot be far from the culprit's hand.

The DDE as a Psychological Effect

The DDE operates against the background assumption that it can be permissible to cause harm when it is an unavoidable consequence of attaining a greater good. Thus, for instance, it may be permissible to bomb an enemy munitions factory, even though some civilians will be killed as a side-effect, because doing so will prevent the enemy from using the completed munitions to kill many *more* civilians in the future. The distinctive feature of the DDE is a limitation on the scope of such permissible harms. It states that it is impermissible to intend harm to another person as a means to achieve a greater good, even while it is permissible to foresee harm to a person as a side-effect of achieving a greater good. Thus, while civilians may be harmed as a side-effect, it is impermissible to intentionally target civilians in order to "break the will" of the enemy and thereby prevent a greater number of civilian casualties. In the latter case, harm to the civilians is intended as a means of accomplishing the goal, rather than occurring as a foreseen side-effect of accomplishing the goal.

Foot's "trolley problem" offers perhaps the clearest expression of this moral distinction between means and side-effects. According to the DDE, it is permissible to redirect a runaway trolley away from five people on the main track and onto a side track where it will hit one person. This is because the foreseen side-effect of killing one person on the side track is outweighed by the benefit of saving five lives. But, it is impermissible to throw a person in front of the train in order to slow it down and thereby prevent it from hitting five people ahead on the main track. This is because throwing the person in front of the train requires intending her death as a means to saving the five people. A handy test of the means versus side-effect distinction asks: could the victim be removed from the situation without interfering with the agent's plan? In the "switch" version of the trolley problem the answer is yes. If the one person were not on the side track, the agent's plan to redirect the train would still succeed; so much the better for everyone involved. But in the "push" version of the trolley problem the answer is no. If the one person were not available to be hurdled in front of the train, the agent's plan would be spoiled. Either she would have to come up with some other way of saving the five, or else they could not be saved at all. This is the sense in which the death of the one person comprises a necessary part of her plan.

It is quite clear that ordinary people (i.e., people other than philosophers) do not apply the DDE by conscious reasoning from an explicit principle, the way that Aquinas first articulated it, or that Chief Justice William Rehnquist applied it in his majority opinion in *Vacco v. Quill*. Very few people outside of specialized academic disciplines have heard of it, and it is often very hard to explain to a novice. When ordinary people produce a pattern of moral judgments consistent with the DDE and are asked to explain *why* their judgements exhibit that pattern, only a small proportion are able to articulate any moral principle that resembles the DDE (Cushman et al. 2006). So if the DDE characterizes some fundamental

property of human moral judgment, it is one that operates automatically and outside of conscious awareness.

But there is a deeper sense in which the ordinary psychological manifestation of the DDE is quite unlike a philosopher's precise formula. Suppose that you give somebody a 1–7 scale of moral wrongness: 1 means completely acceptable, and 7 means completely wrong. If you ask people to evaluate a day spent picking daisies, the average response will be very close to 1; if you ask them to evaluate the murder of an innocent person, the average response will be very close to 7. But if you ask them to evaluate the switch version of the trolley problem, the average response will be close to 4, and if you ask them to evaluate the push version, the average response will be close to 5 (Cushman et al. 2006). Moreover, even this relatively modest effect is augmented by a confounding variable: people tend to feel worse about harms committed with direct bodily force, such as a push (Cushman et al. 2006; Greene et al. 2009). When isolated, the DDE appears to generate differences averaging about 0.3 points on a 1–7 scale. This stands in stark contrast to the philosophical version of the DDE, according to which the behavior performed in the switch case is as permissible as picking daisies in the park, and the behavior performed in the push case is as forbidden as murdering an innocent person. Thus, whereas the DDE specifies a categorical decision boundary for philosophers, it operates instead as a single source of influence on the judgments of ordinary people. All else being equal it is worse to harm as a means, and better to harm as a side-effect, but it is not categorically wrong or right in either case.

These two facts—that the DDE operates outside of conscious awareness, and that it exerts a modest influence on behavior rather than a categorical decision boundary—provide circumstantial evidence that the DDE is not explicitly represented in the brain at all. More likely, it is implicit in the working of cognitive mechanisms designed around rather different operating principles. And, while it may be tempting to therefore write off psychological studies of the DDE as irrelevant to the philosophical project of establishing its normative status, in fact precisely the opposite conclusion follows. Suppose that ordinary people's moral judgments conformed to the DDE because they all walked around with it clearly held in mind, like Aquinas and Foot and Rehnquist. This would provide philosophers will little basis for deciding its normative status. Many people walked around for many years with the explicit belief that the sun revolved around the earth, which only goes to show that people are often wrong. But now suppose, as seems to be the case, that the DDE exerts a subtle and unseen influence on moral intuitions, and that it derives implicitly from some structural feature of the mind that many not even be specific to the moral domain. It a reasonable guess that these same mechanisms would influence philosophers' intuitions about the DDE—intuitions which have often played a critical role in debates over the normative status of the DDE. In that case, if we were to learn something unknown about the psychological origins of the DDE, and thus the origins of the philosophers' intuitions, it might well influence our normative evaluation of the DDE. In other words, by understanding the psychological origins of philosophical intuitions, we would be in a better position to evaluate whether they depend on features that we are prepared to endorse upon reflection.

Causal Attribution

According to one family of explanation the DDE is simply a mistaken characterization of ordinary moral judgment, and its apparent influence is due to confounded factors that

typically co-occur with the core distinction between means and side-effect, but which can be dissociated from it (Royzman and Baron 2002; Sinnott-Armstrong et al. 2008; Waldmann and Dieterich 2007). The two best-developed accounts of this type both focus on processes of causal attribution.

Royzman and Baron (2002) account for the DDE in terms of "direct" versus "indirect" causal influences. Thus, for instance, pushing a man in front of a train is taken to be a more causally direct manner of killing than flipping a switch that redirects the train away from five and towards one. Although Royzman and Baron do not offer a full, precise account of the criteria for causal directness, they do provide examples of some relevant features. For instance, causal directness increases with the number of steps between an action and its harmful effect. This provides one potential analysis of the trolley problem: in the push case, the action on the person's body leads directly to an injury, whereas in the switch case, the action on the train only harms the person's body by way of a secondary effect on the path of the train. (Of course, this analysis depends critically on how one chooses to individuate causal steps.)

Waldmann and Dieterich (2007) offers a related account of the DDE that also focuses on processes of causal attribution. This analysis turns on the locus of an agent's intervention. Specifically, in the push version of the trolley problem, the agent intervenes directly on the victim by forcing him from the footbridge onto the tracks. By contrast, in the switch version, the agent does not directly intervene on the victim, but instead on characteristics of the environment that later turn out to harm the victim. This factor, the locus of causal intervention, has a significant effect on moral judgment even when it is manipulated independently of the means/side-effect distinction.

These critiques of the DDE as a psychological theory hinge on the very accurate observation that the traditional contrast between the switch and push variants of the trolley problem is confounded along several dimensions. The confounds are not limited to causal directness and the locus of intervention, but also include the degree of direct physical contact between the agent and the victim. Fortunately, some attempts have been made to devise alternative versions of moral dilemmas that control for factors other than the distinction between means and side-effect that lies at the heart of the doctrine of double effect (Cushman et al. 2006; Greene et al. 2009). Typically it is found that the means/side-effect distinction plays a role in moral judgment even after the relevant confounds are eliminated.

To give just one example, consider the following pair of cases. In both cases, five people are drowning at some distance from a boat that could save them. The pilot of the boat sees them and must decide whether to rush over to save them. In the side-effect case, the boat must be accelerated so quickly by the pilot that his passenger would be jerked off the back and would drown. In the means case, the weight of the passenger is slowing the boat, but the pilot can accelerate so quickly that he would fall off the back and drown. So, in either case, the five drowning swimmers can be saved at the expense of the one passenger. This pair of cases is controlled in terms of causal directness, locus of intervention, as well as the degree of physical contact between agent and victim. Although the moral distinction drawn between the two cases is not large, it is reliable. Of course this does not mean that directness, intervention and contact do not influence moral judgment—they do!—but rather that there is still something to be explained about the distinction between harming as a means and harming as a side-effect.

There is an additional strike against the hypothesis that processes of causal attribution are responsible for the apparent difference between means and side-effect cases, and one that relies on quite a different method. Cushman and Young (2011) constructed several cases that are structurally similar to moral dilemmas, and which varied along

the means/side-effect dimension, but in which no lives or other morally-relevant values were at stake. For instance, one pair of cases was similar to the drowning swimmers case described above, except that the pilot had to choose between boating over to photograph playful seals at the "expense" of having some seaweed fall off the back of his boat. The participants who viewed these cases were asked, for instance, to what extent the agent caused the seaweed to fall off the back of the boat—in other words, to make a causal attribution analogous to the causal relation between agent and victim in a typical moral dilemma. This experiment revealed no effect of the means/side-effect distinction on causal attribution whatsoever. (By contrast, the same method showed robust differences in causal attribution generated by cases contrasting harm by action versus omission, consistent with prior evidence implicating causal attribution in this moral distinction.)

Intentional Attribution

An alternative psychological account of the DDE focuses not on causal attribution, but rather on the attribution of intent (Cushman and Young 2011). This approach is intimately related to some of the dominant philosophical approaches to the DDE. For instance, Foot writes: "The doctrine of the double effect is based on a distinction between what a man foresees as a result of his voluntary action and what, in the strict sense, he intends. He intends in the strictest sense both those things that he aims at as ends and those that he aims at as means to his ends" (Foot 1967, pp. 5, 6). There is an extensive treatment of these ideas in the philosophical literature, but the associated psychological model is easiest to understand if we approach it on its own terms.

The theory begins with the banal observation that we judge intentional harm as worse than accidental harm. Imagine a nephew who cares for his ailing uncle: it is morally wrong for him to intentionally kill his uncle (for instance, to get an inheritance), but it is not morally wrong for him to accidentally kill his uncle (for instance, by giving him drugs from a mislabeled bottle).

Now consider the distinction between harming as a means and harming as a side-effect, as in the trolley problem. It is not obvious whether we really ought to say that the harm is more "intentional" in one case than the other, and we will consider that question in more detail in a moment. But, as a matter of fact, people do say that the harm caused as a means is more intentional. In fact, they consider all effects brought about as a means to be more intentional than those brought about as a side-effect, even ones that have nothing to do with harm or morality. In the study by Cushman and Young (2011) mentioned above, people were asked not only whether (for instance) the boat driver *caused* the seaweed to fall of the boat more in one case than the other, but also whether the driver *intended* for the seaweed to fall off. When it came to intentional attributions, the distinction between means and side-effect made a reliable difference.

This suggests a potential psychological account of the DDE that proceeds in two steps. Step 1: Whether an act was performed intentionally matters a lot to moral judgments. Step 2: Whether an outcome was brought about as a means or instead as a side-effect influences people's attributions of intentionality. Thus, the means/side-effect distinction influences moral judgments, but not because the DDE is built into the psychological machinery of moral judgment. Rather, the moral influence of the DDE may derive from an ancillary feature of how we judge intentionality, having little to do directly with moral judgment at all.

Further evidence for this model comes from another experiment conducted by Cushman and Young (2011). Participants read several cases of harm as a means or a side-effect, but those cases were embedded in attempted versus accidental settings. Thus, some characters *attempted* to cause harm as a means (but failed), or instead attempted an action that would cause harm as a side-effect (again failing). Others *accidentally* caused harm as a means, in the sense that the harm that they caused accidentally turned out to be a necessary means of bringing about a goal they happened to have, and *mutatis mutandis* for cases of side-effects. The results showed that the means/side-effect distinction is preserved in cases of attempted harm, but eliminated in cases of accidental harm. This is consistent with the model that the means/side-effect distinction depends more generally upon the attribution of intentionality, which is preserved in cases of attempted harm but eliminated in cases of accidental harm.

Does this model imply that the DDE is an appropriate influence on moral judgment? It is widely agreed by philosophers and folk alike that it is appropriate to draw a moral distinction between intentional and unintentional actions. But what does it mean for an action to be intentional, in a sense relevant to moral judgment? Consider again the nephew who kills his uncle intentionally or accidentally. Among the features that differ between these cases, three are especially important: in the first case, the nephew *desires* for his uncle to die, forms a *plan* to kill his uncle, and *believes* that the actions he performs will succeed in killing his uncle. In the second case, the nephew who kills accidentally does not share any of these mental states, and he receives little or no moral condemnation. So, which of these mental states—desire, planning or knowledge—is relevant to moral judgment? The influences of belief and desire information on moral judgment are very large and consistent (Cushman 2008; Young et al. 2007). These influences also have a very apparent function: a person who desires to harm another, or who willingly acts in ways that he believes will harm another, often presents an ongoing risk. This is true even if the harm does not constitute a part of his plan. Say, for instance, your plan is to lock a meat freezer for safe-keeping over the weekend, and you just happen to know that your uncle is inside. This is definitely immoral and it indicates an alarming disregard for the welfare of others on your part, even though the harm does not constitute a part of your means-end planning.

In contrast, the effect of planning—that is, the distinction between harm caused as a means and harm caused as a side-effect—exerts a much smaller influence on moral judgment. Moreover, when isolated from belief and desire information, there is little apparent function in distinguishing between the moral status of harm-as-means and harm-as-side-effect. A person who believes that he will cause harm, but does not desire to cause harm, would seem to present an equal risk for future harmdoing whether he happens to cause harm as a means or as a side-effect of his behavior.

Thus, the influence of the means/side-effect distinction on moral judgment may be due to a mere statistical co-occurence. The harms that we foresee and desire are often planned; the harms that we do not foresee or desire are, necessarily, unplanned. We commonly cluster these distinct types of mental states under the broad descriptions of "intentional" and "unintentional" action, and we take this broad distinction to be morally relevant because of the important role that beliefs and desires play in assessing the likelihood of future harmdoing. In circumstances where we perceive a harm as a means, this may activate our prototype concept of an intentional harm and slightly enhance moral condemnation. This enhancement may not reflect any functional importance of the means/side-effect distinction in the moral domain, but rather the tendency for that distinction to co-occur with mental states of genuine functional importance.

Universal Moral Grammar

There is an alternative view that posits a role for the DDE as a basic design feature of the moral domain: the theory of Universal Moral Grammar (UMG) (Mikhail 2000, 2007, 2011). UMG consists of two core proposals. First, at an abstract level, it proposes that the human competency in moral judgment is the product of a single, relatively discrete psychological system. Borrowing from the principles and parameters approach to generative linguistics (Chomsky 1957, 1965), this system is proposed to operate according to a set of production rules. These production rules are largely innately specified but, as with language, there may be some latitude for developmental processes to select among a larger set of candidate rules.

Second, at a much more concrete level, UMG proposes a specific set of production rules that account for characteristic patterns of human moral judgment. It is useful to divide the production rules into two sets. The first set builds on work by Goldman (1971) that aims to explain how rich causal models of human action (including unobservable mental states) are inferred from perceptual data (including observable actions). For instance, if you observe a person reach for his back pocket while standing in front of a vending machine that sells soda, you might reasonably infer that he intends to retrieve money in order to buy a soda, and moreover that he is probably thirsty. Such internal phenomenal states, plans and goals cannot be directly perceived, but are readily inferred. This first of production rules described by Mikhail constitute a crucial input to the UMG system but are not specific to the moral domain.

The second set of production rules converts structural descriptions of actions, events and circumstances into moral judgments. Several of these rules specify the relative moral value of different outcomes: living is good, physical harm is bad, and death is even worse than physical harm. Other rules specify moral constraints on actions that bring about those outcomes: it is categorically prohibited to intentionally kill, for instance, but it is also prohibited to allow somebody to die when saving him would come at a trivial cost to oneself.

According to UMG, a combination of moral and non-moral production rules account for the DDE. Like virtually all psychological accounts of mental state inference, the non-moral system draws a distinction between beliefs and goals. It also treats the "means" to a goal as if it were itself a goal; we might call this a "sub-goal". Thus, for instance, if you have the goal to make a sandwich and the sub-goal to get some bread from the cabinet, the framework represents them equivalently as goals. (This is a sensible feature, since of course the goal to make a sandwich may be a sub-goal of sating hunger, and so on.) The relevant production rules that infer knowledge and goals operate outside the moral domain.

Within the moral domain, UMG distinguishes between two prohibitions on homicide: one that prohibits killing that is merely foreseen *(*i.e., harm as a side-effect, not as goal), and a second that prohibits killing that is a goal. It specifies that foreseen killing can be justified when it is necessary to save a greater number of lives (a version of the "Rescue Principle"), but that killing as a goal cannot be justified in this manner. Because UMG represents killing as a means identically to killing as a goal (including outside the moral domain), the result is a categorical prohibition on killing as a means to an end.

Under UMG, is the DDE better described as an error, or instead as a design feature? On the one hand, the theory proposes that outside the moral domain no distinction is made between intrinsically desired goals and instrumental sub-goals. This conflation is critical to the ultimate impact of the means/side-effect distinction on moral judgments, and it parallels the observation of Cushman and Young (2011) that events brought about as a means

are judged to have been brought about more intentionally than events brought about as a side-effect. From this perspective, the DDE appears to result from a computational property of mental state attribution operating outside the moral domain, rather than a design feature of the moral domain proper. On the other hand, UMG proposes a set of production rules specific to the moral domain that essentially prohibit cost/benefit analysis in the face of goal-directed harm, but permit it in the face of merely foreseen harm. This is a design feature.

Yet, there are several reasons to doubt whether the characterization of the DDE as a design feature (even partially) is accurate. First, as noted above, it is rather difficult design feature to explain from a functional or adaptive perspective. Why would natural selection favor moral machinery that categorically includes or excludes cost/benefit analysis dependent upon the status of a harm as a goal or sub-goal? The individuals sacrificed are equally dead in either case; the individuals saved are equally living. Surely to *them* the structure of the agent's specific plan of action is irrelevant! A system of moral evaluation based exclusively on foresight is much more easily explained. From the perspective of a potential victim, it mandates: "if you foresee your action causing harm to me, then don't do it."

Second, the sharp boundaries of permissible harm predicted by UMG are not reflected in the actual moral judgments of ordinary people. Recall that, on a 1–7 point scale of moral permissibility, the distinction between harming as a means and harming as a side-effect is typically a half-point or less. This suggests that the DDE operates not as a critical tipping point between wholesale permissibility and prohibition, but rather as a subtle nudge in one direction or the other.

Action Planning

Greene and colleagues (2009) found an unexpected relationship between the DDE and a second influence on moral judgment that they termed "personal force". The attempt to explain this statistical interaction lead to a new theoretical perspective on the psychological basis of the DDE itself.

As noted above, the traditional contrast of the switch and push versions of the trolley problem confounds several independent effects. One of the most salient differences between these cases is that the agent acts upon the victim with direct physical force in the push case, but acts indirectly in the switch case. Early research found that mere physical contact between agent and victim was sufficient to produce harsher moral judgments (Cushman et al. 2006). But, additional work by Greene and colleagues demonstrated that the key factor is not merely physical contact, but rather the direct application of muscular force—a transfer of momentum from agent to victim. This factor is called "personal force". Thus, for instance, pushing a person off a footbridge with a pole is judged to be just as wrong as shoving a person with your hands, whereas dropping a person through a trapdoor on a footbridge by flipping a switch is judged less wrong. The first two actions involve personal force, while the third does not.

The key finding of Greene et al. (2009) was a statistical interaction between the DDE and personal force in their effect on moral judgment. Either factor alone exerted a small effect, but the two of them together delivered a very large effect—a whole greater than the sum of its parts. (In fact, current evidence suggests that personal force exerts no influence whatsoever outside of the context of harm caused as a means to an end.) This statistical interaction with personal force is not explained in any obvious way by accounts of the DDE that rest on causal or intentional attribution, such as those discussed above.

So, where in the brain would we expect a representation with roughly the content, "Harm caused by direct action as a means to a goal"? This is the kind of representation is necessary in order to explain the observed statistical interaction: one that depends upon the joint presence of information about goals and information about direct motor action. One likely candidate is a system of motor planning and production. When humans plan their own actions, they must engage in means-end reasoning and select appropriate motor actions to be performed on their immediate environment. Moreover, such motor action plans would be a logical target for a system of moral self-regulation. That is, a mechanism designed to prevent people from doing harm to others might plausibly "inspect" motor action plans and reject ones that involve harm. A mechanism thus situated would fail to detect harms that are caused as side-effects (which would not be represented in the means-end processing stream of motor planning), or harms that are not directly produced by the motor action in question.

Of course, this proposed "action plan inspector" cannot be the only mechanism of moral self-regulation that operates in the brain. It is obvious that we are able to detect and avoid plans that cause harm as side-effects: recall the poor uncle in the meat locker. Thus, Greene and colleagues situated their proposal within the dual process theory of moral judgment (Greene 2008; Greene et al. 2001), according to which a separate set of mechanisms are responsible for the cost/benefit analyses that generate utilitarian influences on moral judgment. This second system is proposed to have full access to knowledge of side-effects, and to be indifferent to the dimension of personal force. Because it operates using a utilitarian cost-benefit analysis, it endorses harm in both the push and switch versions of the trolley problem. By contrast, the action plan inspector rejects the push plan (which involves personal force and harm as a means), and is silent about the switch plan (which involves neither). In the switch case, then, there is no conflict between the systems and the utilitarian solution is endorsed. But in the push case there is conflict between the systems, and consequently a lower rate of endorsement for the utilitarian solution.

Notably, this proposal locates the origins of the DDE in a mechanism designed to regulate one's *own* behavior. It is reasonable to ask, then, why studies consistently show that the DDE also influences moral judgments of third party behavior (e.g., Cushman et al. 2006). That is, why do we regard it as wrong for some *other* person to push a man in front of a train in order to save five, but right to flip a switch with equivalent consequences? One potential solution is to suggest that we evaluate other peoples' behaviors by simulating performing those behaviors ourselves, and judge the behaviors to be wrong in the event that they engage our own systems of moral self-regulation. We have described as the "evaluative simulation" hypothesis (Miller and Cushman 2013). In this respect, theories of the DDE that implicate systems of action planning stand in fundamental contrast to those that implicate processes of intentional attribution. The action planning theory locates the origins of the effect in the psychology of choosing one's own action and regulating one's own behavior, whereas the intentional attribution theories locate it in psychology of interpreting the actions of a third party and delivering a judgment of the third party. This distinction between the theories may play a critical role in future research that adjudicates between them.

Hierarchical Reinforcement Learning

One shortcoming of early versions of the action planning hypothesis is that it offered little direct connection to current neurobiological models. Yet, our understanding of the

mechanisms that regulate the production of human motor action is very well developed. Over the last 20 years it has been revolutionized by the field of reinforcement learning, a family of computational models that are used to specify how a network of neural systems in the midbrain, basal ganglia and frontal cortex learns the value of different actions and then chooses the appropriate action for a given context.

Reinforcement learning algorithms can be broadly classified into two types, and the distinction aligns elegantly with Greene's dual process theory of moral judgment (Cushman 2013; Crockett 2013). Both algorithms choose which actions to perform based on a representation of the value of those actions in a particular context; they differ in terms of how value is calculated. The "model-based" family of algorithms calculates value by guessing at the likely outcome of different actions, given a causal model of the world (representing, for instance: "if I flip this switch, it will send a train down the side track, and a person is standing on the side track, so it would kill him"). In contrast, the "model-free" family of algorithms derives an estimate of value according to past history of reward or punishment (representing, for instance: "pushing is bad", a representation established via direct experience of the bad consequences following past pushing).

A creative series of experiments illustrates the core difference between model-based and model-free mechanisms (e.g., Dickinson et al. 1995; Dickinson and Shanks 1995). A hungry rat is placed in a box where it learns to push a lever for a food reward. The rat is then taken out of the box and fed until it shows no further interest in food of any type. This procedure is referred to as "devaluation" because food is no longer rewarding to the rat. When it is put back in the box, the operation of model-free mechanism makes it tend to return to the lever and press it repeatedly, even though it has no interest in the food it obtains and lets it sit on the floor. These mechanisms blindly represent something like, "pressing this lever is good", without any knowledge that the particular happy ending—food—is no longer desired. In contrast, the operation of model-based mechanisms inhibits lever pressing because these mechanisms use a causal model to represent the specific outcome of pressing the lever, and that outcome is undesired. In different circumstances rats can be made to follow one or the other set of mechanisms; like humans, they appear to possess psychological systems of both types.

A model-based mechanism would view the trolley problem as a choice between a pair of anticipated outcomes: death to one person, or death to five. If we assume that all else is equal and lives matter most, a model-based mechanism would therefore endorse the utilitarian response to both the switch case and the push case. But a model-free mechanism takes a much more restricted view of the situation. It would estimate the value of "pushing a person" (based on its past reinforcement history) in the push case, and estimate the value of "flipping a switch" in the switch case. It is easy to see why a model-free mechanisms would assign very different values in these two cases: one typically leads to bad outcomes, while the other does not.

In order to make reinforcement learning algorithms work in practice it is necessary to break down actions into sets of recurring features, so that a lesson learned about one specific action can be generalized productively to others that are similar. This may explain why a factor like personal force exerts influence on moral judgments. Rather than separately encoding the negative value of hitting, kicking, slapping, etc., an efficient compression of reinforcement history might take a more abstract form that generalizes across all of these actions as personal force, thus extending to other actions not yet performed (head-butting, elbowing, etc.)

Current reinforcement learning theories provide a potential explanation for the DDE, but in order to understand how, it is necessary to consider the way that model-based and model-free systems appear to interact to produce hierarchically organized behavior. It has long been recognized that humans plan actions in a nested hierarchy of goals and sub-goals (Lashley 1951; Miller 1956): if you want to make a sandwich, get bread; if you want get bread, open the refrigerator; and so on. At one level, the execution of goal-oriented behavior is a hallmark of model-based reasoning. After all, without relying on a causal model, an agent would have no way of assessing whether a particular course of action would lead to the *specific* goal that he has in mind. (Recall the rats subjected to the devaluation procedure, whose behaviors were mismatched to their goals under a model-free regime.)

But, at a lower level, it is possible to use model-free mechanisms to carry out some of the subsidiary computations that enable means-end reasoning (Frank and Badre 2012; Ribas-Fernandes et al. 2011). Suppose that you have in mind the goal to make a sandwich. How is it that you identify "getting bread", in particular, as the next relevant step? It would be impractical to exhaustively search the full space of all possible sub-goals to see which one advances your overall sandwich mission. (Imagine starting at the beginning of the alphabet: "Abscond with an aardvark? No. Abscond with an anteater? No. Abscond with a …".) Clearly, it is necessary to restrict the selection of sub-goals to a very small set of viable candidates. One attractive means of accomplishing this goal is to rely on a model-free mechanism that associates the cognitive state "Goal: Make a sandwich" with the cognitive action "Load sub-goal: Get bread", based on the previous reward history: every time such a sub-goal was selected in past instances of the superordinate goal, things worked out well. Thus the appropriate sub-goal can be efficiently selected, or a set of possible appropriate sub-goals identified for further evaluation by the model-based system.

This mechanism depends on a simple but powerful trick of the mind that has been intensively studied over the past 10 years (Dayan 2012; O'Reilly and Frank 2006). Model-free mechanisms typically operate over perceptual states ("I'm in a Skinner box …") and motor actions ("… so I'll push the lever"). But in order to facilitate sub-goal selection and other constituent elements of higher-level cognition, they may also operate over conceptual states ("My goal is to make a sandwich …") and cognitive actions ("… so I'll load a new sub-goal: get bread").

This very trick of the mind may give rise to the DDE (Cushman 2013). If a model-free mechanism assigns value representations to sub-goals, it seems likely that the sub-goal (harm a person) would be historically associated with very bad outcomes. Consequently, any plan that implicates harm to a person as a sub-goal would be strongly inhibited by a model-free mechanism. Harm brought about as a side-effect would be invisible to this mechanism, however—side-effects would only be evaluated by a model-based mechanism, evaluating all the likely outcomes of an action based on a causal model.

If the DDE is a byproduct of the architecture of hierarchical action selection, then it should not be limited to the moral domain: people should avoid using scary, painful or disgusting actions as a means to accomplishing goals, just as much as they avoid using harm to others. Consistent with this prediction, preliminary research indicates that the means/side-effect distinction influences purely self-interested decisions, such as whether to smear animal feces on one's body in order to deter a hungry wolf from attacking (Cushman, unpublished data). This would not be predicted by models of the DDE that depend on intentional attribution (Cushman and Young 2011) or domain-specific deontic principles (Mikhail 2011), but further research of this type is necessary.

Philosophical Implications

It is not yet clear which psychological account of the DDE is correct, or whether it is in fact a perfect storm of several influences. But even when the cause of a storm is uncertain, you can tell which way the wind is blowing. The psychological accounts of the DDE described here mostly point towards the conclusion that it derives from a set of factors that philosophers are unlikely to endorse as normatively relevant.

To begin with, it is clear that several of the specific hypothetical cases used to describe and explore the DDE in the philosophical literature, such as the trolley problem, are confounded along several other dimensions that independently influence moral judgment. This would be of little concern to philosophers if they were able to insulate their case-based intuitions against unwanted sources of influence, or if they could directly perceive which influences were responsible for which judgments. Unfortunately, research suggests just the opposite. Schwitzgebel and Cushman (2012) conducted a comparative study of 300 philosophers and 900 academics in other disciplines, all affiliated with well-ranked universities and colleges. Participants judged a series of dilemmas such as the trolley problem, and subsequently were asked whether or not they agreed with several statements summarizing prominent moral principles, including the DDE. They found that all participants' judgments of individual cases were influenced by the order in which they judged the cases (for instance, switch before push or push before switch). These order effects were comparable among groups—if anything, stronger among philosophers and strongest of all among those holding PhDs with a specialization in ethics. They then assessed the impact of the order of *judging specific cases* on subsequent *endorsement of general principles*. Here, too, they found an effect—and the strongest of all was the effect on philosophers' endorsement of the DDE. In other words, philosophers were significantly more likely to endorse the general principle when they had previously judged specific cases in one order, compared with another. This indicates that unwanted influences operate on philosophers' judgments of specific cases—and, worse, because philosophers are unaware of those influences, they can impact the subsequent endorsement of general principles. Although "order of presentation" may not apply to the typical philosophical engagement with the trolley problem, which occurs over months or years, factors such as "personal force", "directness", "locus of causal intervention", etc. certainly do. This suggests that philosophers may endorse the DDE at least in part because they misattribute strong intuitions to the DDE that are in part driven by other factors.

Apart from the influences of confounding factors, however, current models of the psychological basis of the DDE raise even more fundamental concerns about its normative status. Although accounts in terms of intentional attribution and action planning diverge completely in their mechanistic details, they agree on one point: the effect of the means/side-effect distinction on moral judgment is best explained not as an adaptive design feature specific to the moral domain, but rather as an incidental effect of the operation of mechanisms that are not specific to the moral domain at all. In the simplest terms possible, the DDE looks like a psychological mistake.

Of course just because something is an adaptive mistake doesn't mean that it is bad. (Neither is a trait with an adaptive function necessarily good.) It may be that human psychology inadvertently stumbled upon an operating principle that happens to align with a deep metaphysical moral truth, or that happens to push us pragmatically towards a high-point in the landscape of human welfare. These are important questions to answer, and they will not be resolved by psychological research alone.

But, as a matter of fact, philosophical defenses of the DDE have tended to rely heavily on intuitions derived from hypothetical examples. Our best psychological theories of those intuitions explain them without reference to metaphysics or welfare, but rather in terms of the ragged edges of imperfect minds. This may not be enough to win the match against philosophers who endorse the DDE, but it ought to be enough to return the ball to their court and await some further play.

Acknowledgments This research was supported by Grant N00014-13-1-0281 from the Office of Naval Research.

References

Aquinas, T. (1988/1274). *Summa Theologiae* IIa IIae, Question 64, Article 7. In P. Sigmund (Ed.), *St. Thomas Aquinas on Politics and Ethics*. Norton, New York.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge: MIT Press.

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

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

Cushman, F. A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108 (2): 353–380.

Cushman, F. A., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science,* 35 (6): 1052–1075.

Cushman, F. A. (2013). Action, outcome and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17 (3): 273–292.

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, 22 (6): 1068–1074.

DeScioli, P., Asao, K., & Kurzban, R. (2012). Omissions and byproducts across moral domains. *PLOS One*, 7 (10): e46963.

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

Dickinson, A., & Shanks, D. (1995). Instrumental action and causal representation. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal Cognition* (pp. 5–25). Oxford: Oxford University Press.

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

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in cortiostriatal circuits 1: computational analysis. *Cerebral Cortex*, 22 (3): 509–526.

Goldman, A. (1971). The individuation of action. *Journal of Philosophy*, 68: 761–774.

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.

Greene, J. D. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 3). Cambridge, MA: 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 (3): 364–371.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. (2007). A dissociation between moral judgment and justification. *Mind and Language*, 22 (1): 1–21.

Lashley, K. S. (1951). *The problem of serial order in behavior*. New York: Wiley.

Lotto, L., Manfrinati, A., & Sarlo, M. (2013). A New Set of Moral Dilemmas: Norms for Moral Acceptability, Decision Times, and Emotional Salience. *Journal of Behavioral Decision Making*, 27 (1): 57–65.

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'*. (Doctoral Dissertation), Cornell University, Ithaca.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Science*, 11 (4): 143–152.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63 (2): 81.

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.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18 (2): 283–328.

Ribas-Fernandes, J. Ú. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, 71 (2): 370–379.

Royzman, E., & Baron, J. (2002). The Preference for Indirect Harm. *Social Justice Research*, 15 (2): 165–184.

Schwitzgebel, E., & Cushman, F. A. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language*, 27 (2): 135–153.

Sinnott-Armstrong, W., Mallon, R., McCoy, T., & Hull, J. G. (2008). Intention, temporal order, and moral judgments. *Mind & Language*, 23 (1): 90–106.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychological Science*, 18 (3): 247–253.

Young, L., Cushman, F. A., Hauser, M. D., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, 104 (20): 8235–8240.