Cognition xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Cognition

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

Original Articles

# Moral vindications

Victor Kumar

*Philosophy Department, Boston University, 745 Commonwealth Ave, Room 516, Boston, MA 02215, United States*

ARTICLE INFO

*Article history:*
Received 13 April 2016
Revised 17 April 2017
Accepted 1 May 2017
Available online xxxx

*Keywords:*
Moral learning
Intuition
Rationalization
Vindication
Debunking
Moral luck
Honor
Disgust

ABSTRACT

Psychologists and neuroscientists have recently been unearthing the unconscious processes that give rise to moral intuitions and emotions. According to skeptics like Joshua Greene, what has been found casts doubt on many of our moral beliefs. However, a new approach in moral psychology develops a learning-theoretic framework that has been successfully applied in a number of other domains. This framework suggests that model-based learning shapes intuitions and emotions. Model-based learning explains how moral thought and feeling are attuned to local material and social conditions. Philosophers can draw on these explanations, in some cases, in order to vindicate episodes of moral change. Explanations can support justifications by showing that they are not mere rationalizations. In addition, philosophical justifications are a fertile source for empirical hypotheses about the rational learning mechanisms that shape moral intuitions and emotions.



## 1. Introduction

Human beings need explanations. We're especially motivated to seek explanations for the behavior of other human beings around us, a trait that emerges early in development. You don't need to be a parent to know that children as young as two incessantly pose the question "why?" Children are even known to seek explanations for this very tendency. One three-year-old is recorded to have said to her mother: "Mommy, I always ask why. Why do I always ask why?" (Callaman & Oakes, 1992: 222; cited in Nichols, 2015: 17).

A few curious children grow up to become more serious students of human behavior. Both psychologists and philosophers wonder why we make the moral judgments and decisions that we do. In a moral context, however, "why questions" are ambiguous—between questions of explanation and questions of justification. When we ask *why* someone donated her hard-earned money to charity, we might be wondering what reasons *caused* her to do so, or we might be wondering what reasons *justify* her choice. The "reason" for a moral judgment or decision—the answer to a "why question"—may be a causal explanation or it may be a normative justification.

Explanation and justification often diverge. Imagine someone who donated to charity because she was moved to imitate a celebrity. This explains her choice but does not justify it. However, when we ask the same person why she donated, she cites a duty to help those in desperate need. This is a perfectly good rationale but does not reveal the actual cause of her choice. We have a justification, but we lack an explanation. The distinction at hand suggests a natural division of labor between psychologists and philosophers who study morality. Moral psychologists seek explanations for moral judgments and decisions; moral philosophers seek justifications.

Some interdisciplinary work on moral thought, however, blends science and philosophy, weaving empirical and normative threads into the same cloth. Sometimes an answer to a "why question" yields both an explanation and a justification—both a cause and a rationale. For example, if research suggests that heterosexuals become more accepting of gay people by empathically appreciating the harms that certain friends and family members suffer, and then generalizing their empathic insight to other gay people whom they don't know personally, the cause of this attitude change also justifies it (Kumar, 2017a). When one and the same account of moral change combines explanation with justification, it is what we shall call a "vindication" (Kumar & Campbell, 2017).

Vindications of moral change can be found in the field of moral learning. In general, learning theory in moral psychology studies the implicit and explicit learning mechanisms that shape moral judgments and decisions (along with other attitudes). Learning mechanisms typically effect stable moral change by reorganizing underlying intuitive and affective structures (Campbell & Kumar, 2012). Moreover, learning can attune moral attitudes in ways that are sensitive to morally relevant aspects of the material and social environment. Under these conditions, moral change is progressive.

*E-mail address:* victor.c.kumar@gmail.com

http://dx.doi.org/10.1016/j.cognition.2017.05.005
0010-0277/

Traditionally, moral philosophers seek ethical theories—like utilitarianism—that explain why any action is right or wrong, in any and all conceivable circumstances. However, one might reasonably worry whether such "ideal theories" are genuinely knowable. A similar epistemological worry has led some political philosophers to be skeptical about theories articulating the structure of an ideally just state (Anderson, 2011; Sen, 2009). Relinquishing an ideal ethical theory, philosophers might instead focus on something more accessible: moral progress (see Buchanan & Powell, 2016). We stand a better chance knowing which remote and recent moral changes have been morally progressive (or regressive) than we do of knowing any complete and definitive moral code. Empirical work has the potential to inform non-ideal theory if it can support generalizations about how moral progress occurs, and thus help to formulate methods for achieving further moral progress (Kumar, in press-a). Furthermore, as we'll see, empirical work offers valuable constraints for non-ideal theorists pursuing philosophical justifications of moral change. Explanations can support justifications by showing that they are not mere rationalizations.

Scientists and philosophers who study moral learning can profit from one another. On the one hand, investigating why moral attitudes change can shed light on their rationale. On the other hand, exploring why moral change is rational can offer clues about the psychological mechanisms that lie behind it. The aim of this essay is to show how explanation and justification of moral change are mutually informative.

## 2. Debunking

My principal topic in this essay is learning mechanisms that vindicate moral change. I will begin, however, by surveying adjacent and more familiar terrain at the intersection of cognitive science and moral philosophy: arguments that attempt to debunk (rather than vindicate) moral attitudes. Later on, I will describe a new approach in moral psychology and the vindications of moral change that it promises.

According to some researchers, moral intuition is a psychological module designed by natural selection: a fast, automatic, unconscious system, relatively isolated from the rest of the mind (cf. Fodor, 1983; Sperber, 1994). Moral intuition, on a nativist view, is relatively *inflexible*: much of its contents are fixed either innately or in a critical developmental window. For example, Mikhail (2011) and Dwyer (2006) argue for the existence of an innate and universal moral grammar, triggered by social cues in early development and capable of producing only a limited variety of local moral languages.

Greene (2008, 2013, 2014) has developed a theory of moral intuition that blends nativism and empiricism. According to Greene, morality is fundamentally a biological adaptation for living in small hunter-gatherer groups that were in severe competition with one another. Moral intuition is a set of simple heuristics that are useful for coexistence within tribal groups. Intuition, Greene says, is like the "point-and-shoot automatic settings" on a digital camera: "highly efficient, but not very flexible" (Greene, 2014: 696). However, Greene also says these "automatic settings. . .can be acquired or modified through cultural learning. . .[and] individual experiences" (698).

Greene thinks that this view of moral intuition has important philosophical implications. He argues that moral heuristics often lead us astray in contemporary, large-scale, technological societies. For example, Greene's own experimental work suggests that intuition produces an emotional aversion to harm inflicted by "personal force," roughly, the use of direct, muscular force to inflict violence upon people. In the Pleistocene environment of evolutionary adaptedness, this was virtually the only way to harm someone, and so a "personal force heuristic" made good sense. But, as Greene says, it doesn't matter, ethically speaking, if you harm someone remotely rather than directly. Thus, while an emotional aversion to personal force is often useful, it is error-prone (Greene, 2014: 713). This is especially true in complex technological societies in which harm is often systemic rather than dyadic, in which death and destruction are increasingly possible through the press of a button. In sum, according to Greene, moral intuition is designed for a lost past and is therefore untrustworthy in new and relatively unfamiliar conditions (714–7). Intuition is thus similar to the psychological drives that produce hunger (697). Adapted for an old environment in which foods rich in calories were rare, these drives now lead us astray in a new environment saturated with candy and fast food.

Greene uses his theory of moral intuition to argue against *deontology*, a class of theories in moral philosophy according to which the rightness or wrongness of actions is partly independent of their consequences. Greene argues that evolutionary relics like the personal force heuristic underlie "characteristically deontological" moral intuitions. The moral theorizing of deontologist philosophers is, at base, an exercise in rationalization—an attempt to offer sophisticated post hoc justifications for intuitions that are actually based on simple heuristics. Since deontological intuitions are untrustworthy, Greene argues, so too are rationalizations of these intuitions within deontology.

At the beginning of the essay, I introduced the idea of an explanation that justifies. Greene offers the inverse: an explanation that "unjustifies." He argues that once we understand the source of certain moral intuitions, we find that the intuitions are dubious, along with any moral beliefs or theories based on them. This is not a vindication, but, rather, a "debunking" (see Nichols, 2014). Greene's explanation suggests that some of our moral beliefs are founded upon error-prone psychological processes. Consequently, he argues, we should give up these insecure moral beliefs and instead adopt those that are shielded from intuitive error. Utilitarianism, Greene thinks, is relatively free of influence from error-prone heuristics.

Greene's debunking argument against deontology has been met with very little sympathy among philosophers. Many critics deny that scientific research on moral beliefs has any potential to inform moral philosophy (see, e.g., Berker, 2009). As a whole, philosophers tend to heed the "is-ought gap." According to this Humean doctrine, there is a logical gap between empirical claims about what "is" the case and philosophical claims about what "ought" to be. We can gain all the knowledge to which we aspire about how a moral judgment *is* in fact made, but that won't tell us about how we *ought* to make it.

On the contrary, however, properly formulated debunking arguments are perfectly consistent with the is-ought gap—so long as empirical premises are complemented by normative premises (Kumar, in press-a). Greene, indeed, disavows any attempt to infer normative conclusions from only empirical premises (Greene, 2014: 711). He explicitly relies on normative premises too. Thus, Greene claims that it does not matter whether or not harm is inflicted through personal force—that this is a *morally irrelevant factor* on which to base moral judgments—and his philosophical critics, on pain of acute implausibility, cannot but accept this normative claim (Kumar & Campbell, 2012).

However, Greene's debunking argument against deontology proves to be unconvincing in light of a more complete description of the underlying psychological processes. Deontological intuitions are sensitive to a range of factors aside from whether harm is inflicted through personal force. For example, intuitions track the degree of harm inflicted, whether it was caused intentionally or only accidentally, whether it was intended as a means to an end

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

or merely as a foreseen side effect, whether it was a deserved response to aggression or unprovoked, and so on. Greene argues that deontology lacks rational credibility because it is based on untrustworthy heuristics that underlie deontological intuitions. However, if this argument is to succeed in lowering the credibility of deontology below that needed for rational credence, Greene must employ a normative premise to the effect that these *other factors* influencing intuition also do not lend rational credibility to the deontological beliefs that rest upon them. The problem for Greene is that this normative premise is not at all plausible (see Kumar & May, in press). It is highly controversial that degree of harm, whether or not harm is intentional, and whether it is deserved are *morally irrelevant* bases for belief. Greene might argue that deontological beliefs tend to be affected by morally irrelevant factors more so than utilitarian beliefs, but the empirical evidence for this comparative claim has not been supplied, and it is not supported by cherry-picking instances, such as trolley cases (see Kumar & Campbell, 2012 for further discussion).

Even if Greene's empirical premises were uncontroversial, his debunking argument against deontology would require an unsustainable normative premise. In the next section, however, I will raise doubts about Greene's empirical premises in light of recent work in moral learning theory. Ultimately, as we'll see, learning theory shows how moral intuition might be flexible rather than innate, often in seemingly rational ways.

## 3. Learning

Moral intuition tends to be inflexible in one sense—over the short run. In the moment, moral intuitions tend not to be sensitive to countervailing reasoning. For example, many people have the intuition that bigoted speech is morally wrong and should be censored; this intuition often persists in the face of explicit reasoning about the value of free speech, even when people endorse the reasoning.

However, intuition does seem to be flexible over the long run. Whereas explicit moral reasoning is "synchronically" flexible, moral intuition is "diachronically" flexible (Campbell & Kumar, 2012: 282–3; see also Gottlieb & Lombrozo, in press; Pizarro & Bloom, 2003). This flexibility is what makes stable, progressive moral change possible. People do not revise only their explicit beliefs about issues like the moral status of women and people of color. Over the course of a person's life, intuitions change too. Witness, for example, the revolutionary change in moral intuitions about gay people and, in earlier stages as yet, those who fall outside of traditional gender binaries (Broockman & Kalla, 2016; see also Kumar, 2017a; Kumar & Campbell, 2017).

The most striking work on moral learning to date has been conducted by Nichols, Kumar, Lopez., Ayars, and Chan (2016). A number of studies suggest that moral rules underlie intuition, and thus that straightforward cost-benefit analysis does not lie behind all moral judgments. Some of these rules are deontological: they draw moral distinctions between acts vs. omissions, intended vs. unintended actions, and means vs. side effects. According to Greene, it is not rules but crude, emotionally backed heuristics that produce the relevant pattern of judgments and decisions; rules come into play only as rationalizations of these judgments and decisions. Other researchers, by contrast, accept the existence of deontological rules, but, like Greene, offer an evolutionary explanation of their etiology. According to moral nativists who advance the "linguistic analogy," children possess implicit deontological rules without being exposed to sufficient information about their content in the children's early learning environment (Dwyer, 2006; Mikhail, 2011). Paralleling Chomsky's case for linguistic nativism, moral nativists argue that the rules, because they cannot be learned, must be innate.

Nichols, however, develops a statistical learning model that can explain how children acquire certain deontological rules. He begins with a simple analogy (further simplified here). Imagine that one of two dice is being rolled, though you are not told which. One is four-sided and the other is eight-sided. Your job is to notice which numbers turn up and then form a hypothesis about which of the two dice is producing them. If you record anything from 1 to 4, that makes it more likely that the die being rolled is four-sided rather than eight-sided. For the likelihood of that evidence is slightly higher on the former hypothesis than on the latter. Moreover, with each successive piece of data that falls between 1 and 4, you acquire more and more evidence that it's the four-sided die rather than the eight-sided die—so long as none of the data fall between 5 and 8. It would be a huge coincidence if many rolls of an eight-sided die yielded only half of its possible values.

The key feature of this example is that the two hypotheses are nested. The evidence consistent with one (1–4) is a proper subset of the evidence consistent with the other (1–8). As evidence accumulates that is consistent with the narrower hypothesis, that hypothesis increases in probability. As Nichols discerns, deontological rules *also* involve nested hypotheses. For example, intended actions are a proper subset of foreseen actions; every intended action is foreseen, but not all foreseen actions are intended. So, if children are repeatedly told that certain actions are wrong, and if all those actions are prohibited by both the narrow and the broad rule, a statistical learning model predicts that children will infer the validity of the narrow rule. Thus, if a child is told that her foreseen actions are wrong, but only when these foreseen actions are intended, then she has reason to infer that the rule covers only intended actions and not those that are unintended but merely foreseen.

By itself, this is a "how-possible" story for rationally learning deontological moral rules. But Nichols and colleagues have been turning it into a "how-actual" story, conducting experiments in which participants are more likely to infer that a narrower deontological rule is at play when they are given examples of rule violations that are predicted by the narrow rule. Given statistical learning algorithms, the relevant evidence is, rather than impoverished, sufficiently rich to generate a competence with certain deontological rules. In general, people follow moral rules without explicitly apprehending their content—few can articulate the distinction between actions that are intended and those that are unintended but merely foreseen, even though they employ the distinction in moral judgment (Cushman, Young, & Hauser, 2006). So, it seems, unconscious, rational learning mechanisms may explain how people achieve an intuitive grasp of deontological rules.

Nichols' research casts doubt on the empirical premises in Greene's debunking argument against deontology. Greene thinks that deontological intuitions are innate rather than learned, since they are elicited only by "up close and personal" harms that were the norm for our tribal ancestors. However, if deontological rules are learned, as Nichols' research suggests, deontological intuitions are not simply untrustworthy relics of a lost past. Even if we accept Greene's hypothesis that deontology is founded upon intuitions, we have reason to doubt that these intuitions issue from inflexible heuristics.

Nichols' research offers a clear demonstration of the potential fruitfulness of learning models in moral psychology. However, Nichols offers less than we might hope for as we search for moral vindications. Even if he is right about how children learn moral rules, and even if this learning follows statistical norms of inference, this does not vindicate the products of rule learning. The reason is that nothing in the learning story vindicates *the rules*

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

*themselves*. That children have the power to acquire rules from subtle information in their learning environment may lead to the adoption of unjustifiable or even reprehensible rules (see Huebner, 2016). What's needed, then, is an understanding of the conditions under which learning mechanisms engender progressive moral change.

## 4. Justification & rationalization

As we've seen, some explanations of our beliefs, if correct, would rob them of credibility. A significant proportion of our moral beliefs are based on unconscious psychological processes; introspection cannot tell us about their provenance. It is tempting to rationalize these beliefs—to find some justification for them after the fact. However, rationalization is untrustworthy because human beings—philosophers especially, perhaps—are so good at it (Haidt, 2001). Perhaps no matter what our starting points, we would be able to find some justification for them, even if our starting points rest on error-prone heuristics.

What if, however, certain unconscious psychological processes that influence moral attitudes are more rational than evolved heuristics? A causal explanation would not be embarrassing if it implicated sound principles for updating moral attitudes in light of empirical evidence about new material and social conditions. If we want to find a credible justification for our moral attitudes, then, we should eschew untrustworthy rationalization and instead seek vindications—accounts of moral change that offer both explanation and justification. To do so, however, we need a better understanding of vindicating arguments and their philosophical significance.

Debunking arguments rest on empirical and normative premises. The empirical premise specifies the causal process that underlies a set of beliefs. The normative premise claims that this premise is unreliable. Thus, Greene makes the empirical claim that deontological beliefs are generated by a set of moral heuristics. He also makes the normative claim that these moral heuristics are unreliable because they are sensitive to factors that are morally irrelevant in our current environment.

Vindicating arguments similarly rest on empirical and normative premises. Like debunking arguments, the empirical premise describes the casual history of a set of beliefs. In this essay, we are interested in causal processes that explain episodes of moral change. The normative premise claims that this process is reliable, sensitive to morally relevant factors. Statistical rule learning is a rational process, deriving rules in response to subtle patterns of feedback in the environment. As we've noticed, however, this feedback can be cued either to morally relevant factors or to morally irrelevant factors. So, for purposes of vindication, we need another learning mechanism and an understanding of the conditions under which it is reliable.

Several other types of learning mechanisms seem to shape moral intuition. One type is simple, "model-free" reinforcement learning. Immediate rewards and punishments influence moral intuition; brute conditioning influences other thought and behavior too. Another type of learning mechanism seems to be at work in domains of cognition outside of morality, including language, causation, and theory of mind: "model-based" learning (cf. Gopnik & Wellman, 2012). Cushman (2013a) and Crockett (2013) both develop computational theories of model-free and model-based learning in moral cognition. On their characterization of model-based learning, people implicitly construct complex value-laden models of their material and social environments. These models then generate expectations about the value of performing various actions. People tend to perform actions with high expected value and then receive feedback about their actual value. This feedback, and its discrepancy with expectations, is used to revise the model and consequent expectations of value.

Recently, Railton (2014) has argued that model-based learning explains why moral intuition is reliable—why moral intuitions can be trusted in everyday judgment and decision making, as well as in moral philosophy. However, model-based learning might also provide vindications of moral change: accounts of changes in moral intuition that not only explain but also justify. Railton attempts to vindicate intuition; my focus is more specifically on vindicating *changes* in intuition. One reason that model-based learning is equipped to do both empirical and normative work is that it is, by any standard, a rational psychological mechanism. It allows people to implicitly track statistical regularities that bear on the satisfaction of values and update their intuitions accordingly. Model-based learning is potentially vindicating not because it changes intrinsic moral values, but because it allows people to improve the means they take to achieving moral ends.

However, model-based learning is reliable only when the values that guide it are cued to morally relevant factors. Consider the value against harming others. If model-based learning allows people to implicitly track which sorts of activities are likely to bring about harm, it can lead them to disvalue these activities. This sort of instrumental learning is reliable because it allows people to gain implicit knowledge about the means to morally worthy ends. The normative premise here is that one ought to avoid harming people, all else being equal, and this premise is as plausible as any to which philosophers might wish to appeal. Thus, instrumental model-based learning may vindicate moral change, but only when learning is anchored in appropriate values.

It's clear enough how the causal history of moral change might justify it. If the explanation for my change in attitude toward Syrian refugees is attunement of my sympathy for the harms they suffer, the cause of my new belief supports it (given the normative premise that harm-avoidance justifies better treatment toward refugees). I will suggest, however, that model-based learning explanations can support not just new beliefs *but also philosophical justifications for them*.

This claim will trouble philosophers who think that moral truths are knowable only from the armchair (perhaps a priori). Philosophical critics might argue that empirical vindication of a moral belief must presuppose the armchair reasoning that justifies it. We can know that the causal history of a belief is vindicating only if it reflects good reasons for the belief that we already possess. Thus, it would seem, etiological analysis is otiose. If we know (from the armchair) the reasons for a true moral belief, then any knowledge (from the lab) cannot but support the belief. To address this challenge, we must grasp more clearly the threat that rationalization poses to moral philosophy.

That a reason is good *in general* is a normative claim, supported by philosophical considerations rather than empirical considerations. However, even if a reason is good in general, that does not guarantee that it genuinely *applies* to any given belief. Suppose that inequalities are justified if they arise from incentives that benefit everyone; equalization would destroy incentives and thereby leave everyone worse off in the long run. This is a good reason in general for inequality, let's assume, but it does not show that any particular inequality is justified—not unless we know that this inequality in question *actually does* benefit everyone. It is tempting to cite this reason for inequalities that are self-serving or merely deeply entrenched in society. The problem is that people are remarkably clever at offering good reasons for their beliefs no matter what their beliefs happen to be. This is rationalization, untrustworthy because it tends to be over-applied. However, if a good general reason is reflected in the etiology of a belief, this bolsters the idea that the reason genuinely applies to the belief. For then the belief has been formed in response to conditions in the environment to

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

which the general reason applies. Thus, explanations can support justifications by showing that they are not mere rationalizations.

Is it *necessary* that a moral belief be formed on the basis of good reasons in order to be credible? No. New moral beliefs may arise because they are self-serving, say, but they may nonetheless constitute moral improvements. To wit: we have the right view, but we acquire it for the wrong reasons. In that case, justification is not reflected in etiology. However, although accidental moral progress does happen, it is not typical. The ability to figure out what is right and wrong through reason alone is limited. Moral progress tends to occur when people carry out successful "experiments in living" (Mill, 1859; see Anderson, 2016). Thus, when a moral change is progressive, the justification for it is usually reflected in its etiology. When the justification is not so reflected—when the justification is only discovered post hoc—this is a sign, though not a guarantee, that the justification is merely a rationalization. This is why vindication offers prima facie support for the corresponding justification.

In the rest of the essay, I will pursue moral vindication by looking at three cases of model-based learning and the way in which they seem to not only explain moral change but also justify it. The vindications offered are instrumental, and therefore broadly consequentialist (neutral, however, regarding what precisely counts as a good consequence). I will propose that several remote innovations in the history of morality occurred because they led to better consequences than the alternatives or the status quo. Thus, the explanation for these moral changes supports a consequentialist justification for them, because it shows that the consequentialist justification is not a mere rationalization. The arguments suggest that certain episodes of incremental moral change are progressive. They are, however, defeasible, leaving open the possibility that alternative changes would produce even better consequences. Each case study below, as we'll see, involves a rich interplay between science and philosophy. Our discussion will revolve around the role in moral cognition of negative emotions like anger and disgust. But we'll begin pursuit of moral vindication with the curious case of moral luck.

## 5. Moral luck

Imagine that you are cruising blithely down an empty street, well over the speed limit, when suddenly a young child steps in front of your car and is tragically killed. You acted recklessly and are responsible for the child's death. The child's parents have cause to blame you for their tragedy; the law imposes a lengthy prison sentence for manslaughter. However, imagine now that I too have been driving over the speed limit, but I luckily get away with my recklessness. I do not suffer the same blame to which you are subjected; even if the police catch me, I am not sentenced to prison.

This pair of cases illustrates the problem of moral luck (Nagel, 1979). You and I were both reckless in precisely the same way. It is simply a matter of luck that you are convicted of manslaughter and I am given only a ticket for speeding. The puzzle for philosophers and legal theorists is that it seems unfair or unreasonable to blame and punish two people differently on the basis of factors that are utterly beyond their control—in this case, whether or not a child happened to step in front of our vehicles. So, what should we do? Should we change our moral and legal practices? Or should we accept that responsibility is partly a matter of luck?

The problem of moral luck arises out of a clash between a plausible, general moral principle—that people are responsible only for what is under their control—and strong moral intuitions about many particular cases. The problem has occupied philosophers at least since Smith (1759), and perhaps as far back as Plato and Aristotle. But moral luck is also of interest as a puzzle in empirical moral psychology. *Why* do we blame and punish others—and ourselves—on the basis of luck? What explains our intuitions? This is not an easy question, but understanding the psychology of moral luck is more tractable than solving the difficult and longstanding philosophical problem. As we'll see, illuminating the psychology of moral luck can also advance the philosophical debate.

Fiery Cushman and colleagues have recently carried out a series of important studies on moral luck (Cushman, 2008, 2011, 2013b, 2015; Cushman, Sheketoff, Wharton, & Carey, 2013). Cushman's findings reveal that two psychological processes guide judgments about blame and punishment. The "mental process" identifies an agent who intended harm or was negligent. The "causal process" records a harmful outcome and then searches for the agent who caused it. The total amount of blame and responsibility assigned to someone depends on the contribution of both processes. (The mental process seems to contribute more than the causal process, though research does not yield any precise determination of their relative proportions.) So, when we confront two people who have been negligent, the mental process assigns blame and punishment to both. But when only one person actually caused harm, the outcome process offers an extra slice of blame and punishment to her. This is, it seems, why we have the intuitions that arise in us about numerous cases of moral luck, including the hypothetical case of the two reckless drivers.

In philosophy, we find a division between realists and skeptics about moral luck. Realists accept that moral responsibility outstrips control (Kumar, in preparation-b; Moore, 1997, 2009; Walker, 1991). Skeptics, by contrast, insist that we should hold people responsible only for things that are within their control, and therefore that we should abandon or override the recalcitrant intuitions that seem to support the existence of moral luck (Richards, 1986; Thomson, 1993; Wolf, 2001). Some skeptics attempt to debunk our intuitions by arguing that they are distorted by cognitive biases (Domsky, 2004; Royzman & Kumar, 2004). One view says that when assigning blame and punishment, we care only about a person's mental states—their malign intentions or their recklessness—but through the bias of hindsight we infer from bad outcomes that the agent's intentions were bad too. However, if Cushman's model is correct, the intuitions at play are not the product of a general bias; they issue from a dedicated mechanism that assigns responsibility on the basis of outcomes. Our moral judgments are not generally influenced by hindsight bias, Cushman's work shows, since it's only judgments about blame and punishment, and not judgments about wrongness, impermissibility, or character that are influenced by luck (Cushman, 2008). Skeptics attempt to debunk intuitions by relying on an empirical premise about their source in general cognitive biases, but in the case of intuitions about moral luck, this premise is false.

Cushman isn't content to have only a psychological model of luck intuitions. He also develops a model of the evolutionary pressures that may have given rise to the two processes underlying blame and punishment (Cushman, 2013b). The evolutionary roots of punishment are relatively well understood, in terms of their contribution to the evolution of pro-sociality (Boyd & Richerson, 1992). A group of pro-social agents will fare better on average than a group of anti-social agents, but anti-social agents within the pro-social group will do better than others, and thus over time will increase their relative proportion in the population. Punishment evolved as a mechanism for introducing costs on anti-social agents, making pro-sociality fitness enhancing not just across groups but also within them, thus securing the group-level benefits of pro-sociality. Blame seems to fit within this general framework. Arguably, punishment functions to materially sanction others, while blame functions as a type of social sanction.

Moral intuition plays at least two main roles in the relevant social dynamics. On the one hand, it gives rise to pro-social feelings

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

of sympathy that engender altruism and cooperation. On the other hand, intuition gives rise to punitive feelings of anger that engender blame and punishment. (Blame and punishment arguably have a cognitive dimension too. See, e.g., Wallace, 1996). Cushman notes that pro-social feelings are relatively flexible. When kindness pays, we learn to be more kind; when kindness is costly, we learn to be less kind. Interestingly, however, punitive feelings are relatively *in*flexible. We're disposed to blame and punish others whether or not doing so incurs costs or benefits (Carlsmith, 2006, 2008).

Why is punishment inflexible and why is it based partly on the outcomes of a person's action? Cushman argues that this way of assigning punishment provides an excellent learning environment in which to nurture pro-social feelings. The outcomes of a person's action are easier to identify than the mental states that lie behind their action; best to rely on both as the bases for anger and punishment. So, punishment is more reliable if it is based partly on outcomes. That is, punishment based partly on outcomes is more likely to correctly apply its own standards (compared to punishment solely based on mental states, which is likely to be incorrectly applied). Furthermore, if punishment were flexible, this would provide anti-social agents with an opportunity to introduce disincentives on punishment that inhibit punitive feelings among others and allow them to get away with anti-social behavior. What this suggests, then, is that the intuitions that underlie moral luck evolved in order to facilitate effective punishment, which itself facilitates effective learning of pro-social feelings.

The reason that punishment of outcomes facilitates moral learning, according to Cushman, is that model-based learning underlies pro-social feelings. People construct implicit models of the connection between their feelings, the outcomes of the behavior produced by those feelings, and the costs and benefits of those outcomes. Through their behavior, people receive feedback that is used to update their models. Thus, model-based learning attunes pro-sociality in the face of punishment that is tied to outcomes.

Cushman finds support for this learning theory in experimental work (Cushman and Costa in prep). In Cushman's study, pairs of participants are recruited and assigned to one of two roles. "Shooter" is given the task of throwing darts at a board with two different kinds of targets. "Trainer" wins or loses money depending on which targets Shooter hits. Trainer's task is to reward and punish Shooter so as to increase her accuracy and thereby increase his own payoff. In one condition, Trainer rewards Shooter whenever the target she *announces* is the one that increases his payoff. In the other condition, Trainer rewards Shooter whenever she *hits* the right target. Cushman and colleagues find that Shooters improve in both conditions, but are nearly twice as good at improving their performance in the latter condition. Thus, it seems people are particularly equipped to track contingencies between the outcomes of their actions and reward/punishment.

Cushman's main aim is to develop a psychological and evolutionary explanation of moral luck. The influence of two psychological processes explains why we react differently to two people who perform the same action but bring about different outcomes. These psychological processes exist in order for punishment to effectively facilitate model-based learning of pro-sociality. As we'll see next, though, this empirical explanation of moral luck also lends support to a philosophical justification.

Two philosophical theories of punishment are attractive. Retributivism is the view, roughly, that punishment is justified if agents merit suffering in virtue of their bad intentions or state of mind. Consequentialism is the view that punishment is justified if it leads to positive consequences, in particular, if it deters immoral behavior. This debate about punishment parallels the two main schools of thought in the philosophical debate about moral luck. Skeptics believe that moral luck is illusory, that we should hold people responsible only in accord with what their malicious intentions or their negligence deserves. Realists, however, believe that moral luck is real, and that holding people responsible for the outcomes of their actions is justified because it improves moral behavior. Often, realists claim not that any particular act of blame or punishment is justified on grounds of deterrence, but that our practice as a whole is justified because it leads to better consequences than alternative practices.

Cushman's explanation offers new support for a partially consequentialist rationale for moral luck over a pure retributivist theory (Kumar, in preparation-b). Cushman's explanation shows that the reason why our moral judgments are sensitive to luck is that this brought about better consequences in the evolution of altruism and cooperation. Specifically, this sensitivity provided a learning environment that favored model-based learning of pro-social feelings. Consequentialists are disposed to argue that moral luck is justified because holding people responsible for the outcomes of their actions has better consequences. However, philosophers in this camp provide virtually none of the required evidence that blame and punishment on the basis of lucky outcomes actually has positive deterrence effects. So, their consequentialist justification threatens to be nothing more than a rationalization, given the possibility that one can cook up a consequentialist justification for our practices no matter what their shape. Cushman's psychological and evolutionary explanation comes to the aid of consequentialists, since it indicates that moral luck does have deterrence effects—and explains why. Thus, moral luck reflects the adoption of standards of punishment that are better than alternative standards.

Cushman's research offers a *vindication* of moral luck in terms of moral learning. The explanation for why we blame and punish others for things that are beyond their control reflects a rationale. We impose material and social sanctions for unlucky outcomes because this provides a better set of learning conditions for prosociality to flourish. A consequentialist justification for moral luck therefore becomes more plausible, since it has actually shaped our psychological dispositions and practices, and thus is no mere rationalization. We couldn't have known precisely why morality is partly a matter of luck without understanding *why* we have become the sorts of creatures who think, feel, and act as if this is the case.

Some moral luck skeptics think that the intuitions sensitive to luck can be debunked. As shown above, the empirical premises in this debunking argument are implausible. Furthermore, however, Cushman's explanation for our luck intuitions does not only rule out a debunking explanation. It shows that consequentialism is not merely post hoc, that the general consequentialist reason to deter immoral behavior applies in this particular case. This is not a decisive solution to the philosophical puzzle, but it does advance the philosophical debate. To undermine this piece of empirical support for realism about moral luck, critics might pursue the possibility that outcome-based punishment once nurtured pro-sociality, but no longer does. Absent any such argument, however, we have prim facie reason to believe that consequentialism applies to our practices of moral luck.

## 6. Honor

A scientific explanation of moral learning has the power to lend credibility to a philosophical justification. We can be confident that a philosophical justification is not untrustworthy rationalization, since the factors that seem to justify a change in moral attitudes are reflected in an explanation of its etiology. However, as we'll see next, cross-disciplinary fertilization occurs in the opposite direction too. A philosophical justification of moral attitude change can yield a hypothesis about its etiology. In this section and the next, we'll explore philosophically informed moral learning theory.

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

Our next case study targets the emotions and intuitions that underlie moral honor.

Several philosophers have recently turned their attention to norms and values associated with honor and how they fit with the rest of morality (see, e.g., Appiah, 2010; Demetriou, 2014; Kumar & Campbell, 2016). Honor is arguably present in all societies, but so-called honor cultures tend to emphasize it. What that means, essentially, is that members of honor cultures tend to care a great deal about respect based on social identity. Who a person *is* gives them the right to be honored and respected, and what they *do* can enhance or diminish their honor.

In many honor cultures, including the American South, people not only care about respect based on social identity, but also have a lower threshold for anger in response to perceived attacks, threats, and insults. To fail to respond is to show oneself to be unworthy of respect—that is, dishonorable. Nisbett and Cohen (1996) famously confirmed this pattern. They found that in a lab setting Southern men were more likely than Northern men to affirm honor-based values, more likely too to express anger when threatened or insulted. In one experiment, for example, these tendencies were revealed by levels of stress hormones. Southern men were more likely than their Northern counterparts to show elevated cortisol in response to aggressive behavior.

There is a Standard Story about honor cultures in which men are prone to anger, originally proposed by Nisbett and Cohen and now widely endorsed. These sorts of honor cultures tend to arise in communities where property is portable and there is no social institution that reliably enforces property rights. For example, Southerners traditionally migrated from herding areas of Scotland where their livestock were subject to theft and for which the chance of restitution was low. Northerners traditionally migrated from farming communities. It is, in short, much easier to steal livestock than it is to steal crops. The Standard Story says that the difference in anger thresholds is due to these differences in material and social conditions. If your property is portable and easily subject to theft, expression of anger and consequent aggression is an effective deterrent.

Philosophers are generally skeptical of honor, perhaps because it is implicated in divisions of class, race, and gender. Honor also undergirds morally regressive social practices, including dueling and honor killing (see Appiah, 2010). However, we should be suspicious when philosophers reject aspects of morality that are foreign to the social and political culture to which they belong (cf. Duarte et al., 2015). Furthermore, eradication of certain regressive social practices often depends not on eliminating honor but expanding its scope (Appiah, 2010; Kumar & Campbell, 2016). Might there be good reasons for people to be concerned with honor?

Let's return to the widely endorsed explanation for honor cultures in the American South and elsewhere. One face of the Standard Story is an explanation, but the other face is a justification. The Standard Story is a *vindication*. Assuming that it is worth retaining your property, all else being equal, it *makes sense* to acquire a lower threshold for anger in the conditions under which some honor cultures develop. Anger has greater instrumental value for herders than for farmers, and so this difference between the two communities fits their respective environments. This can be formulated more precisely in comparative terms: the relative difference in ease of theft between these two cultures justifies the introduction of a difference in thresholds for anger (even if the actual difference happens to be too extreme). The justification is not merely one of self-interest, i.e., not just prudential. Retaining one's property has immense social value for one's family and often for one's community as a whole.

Strikingly, however, there has been little empirical work that substantiates an explanation for *how* people's anger dispositions become attuned to material facts about property and social facts about enforcement of property rights. Without any such explanation, the Standard Story begins to look like a "just-so story" (Gould & Lewontin, 1979). That is, it is a plausible sounding tale about the existence of a trait, and how it is fitted to its environment, but lacks the empirical evidence needed to back it up. The problem with just-so stories is that they are so easy to construct. One can frequently tell a plausible adaptive story, for any given trait, even when it is not an adaptation. In that case, the likelihood that any particular adaptive story is true becomes quite low.

We find here another parallel between explanation and justification—in this case, between just-so stories and "just-why stories" (Kumar & Campbell, 2017). A just-so story for a behavioral trait offers an unsupported *explanation* for the existence of the behavior. A just-why story for a behavioral trait offers an unsupported *justification* for the behavior. The worry is that the story is just a rationalization. For example, it is possible that *any* emotional response or action displayed by Southerners and Northerners could be justified through one or another concocted narrative. As noted earlier, the easy availability of post hoc justifications casts doubt on any particular one. A story that rationalizes the status quo *whatever it happens to be* is just that—a rationalization and not a genuine justification. False positives abound, and so the probability is high that the behavior in question is not really justified at all.

Greene argues that deontological moral philosophy is, at base, one big rationalization. I suspect that this is true, to some extent, not just of deontology but of *any* philosophical theory that attempts to justify our moral attitudes and behavior while ignoring what explains them. The plausibility of a justification depends on the existence of an explanation that dovetails with it. For then we have reason to believe that the justification genuinely applies to the moral attitudes and behavior in question. Otherwise, we cannot trust the justification in the same way that we cannot trust unmoored adaptationism.

What then of honor? It is unlikely that biological evolution can explain honor cultures. Their origins are too recent. However, a mechanism *like* natural selection, one that explains the apparent fit between organism and environment, must be at play if the Standard Story is correct. We might hypothesize, then, that model-based learning is a key mechanism that vindicates honor cultures. This would fill in missing details in the explanation suggested by Nisbett and Cohen. Thus, people begin with a model of their environment that represents how liable their property is to theft, and this model generates expectations about the value of experiencing anger and aggression. People receive critical feedback, in the form of others stealing property, and this updates their model and expectations. Slowly and over time, people increase the value they assign to anger as their internal model tracks the statistical dependency between anger and retention of property. What this sketch of a theory offers is an explanation of how honor cultures arise in the first place. The process is incremental. Children imitate their parents, each generation learns a little more about their environment, and over time a significant difference between, say, farmers and herders arises. Developmental instruction and social norms reinforce the tendency, but it arises initially through experiential, model-based learning.

This empirical hypothesis is in need of testing. One way to do this is to expose participants to scenarios involving theft without restitution, and then probe whether they are likely to feel more anger in response to other, related moral violations. Boyd and Richerson (2005) offer an alternative hypothesis, in terms of imitation and cultural evolution, but it has not yet been substantiated. If the learning model offered here about the origins of honor cultures were to be empirically confirmed, it would add to our understanding of cross-cultural moral variation. But it would also lend support to the rationale for honor cultures suggested by the Standard Story.

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

The Standard Story would not merely provide a *just-why story* for the status quo. The development of a concern with honor makes sense, and not merely to minds drawn to rationalization, because it has actually been shaped by the very material and social conditions that justify it.

To be clear, the proposed vindication for certain honor cultures is defeasible, what philosophers call only a "prima facie" or "pro tanto" justification. There may be *other* reasons that count against lower thresholds for anger in honor cultures. Rates of violence are higher in such cultures, and an elevated concern with honor sometimes leads to spiraling cycles of revenge in the style of the Hatfields and the McCoys. Furthermore, the conditions that explain and justify lower thresholds for anger often disappear over time, with entrenched social norms giving them an inertia that they do not merit. When honor persists in conditions for which it is not suited, we have an explanation that debunks it.

So, empirical work may suggest a vindication of honor cultures generally, but a Greene-style debunking of their persistence in certain cases. A moral change toward a culture of honor is justified for some groups, but a further moral change in response to a reversal of environmental conditions would now be justified. Still, we find ourselves with the interesting and surprising result that empirical evidence can support a prima facie justification for honor cultures, even if the justification does not win the day.

## 7. Moral disgust

Near the beginning of the essay, we looked at debunking arguments. Since then, however, we've been occupied with another approach, similar in that it studies the causal history of moral attitudes, but different in that it attempts to vindicate them by showing that their causal history reflects an underlying rationality embodied in model-based learning. In moral vindications, explanation buttresses justification and, as we'll now see again, justification guides explanation. We'll continue to look at the role of negative emotions in moral thought, but our third and final case study takes a different focus.

Anger is the pre-eminent negative emotion in philosophical and psychological study of moral thought. Anger underlies blame and punishment, and, as we've seen, is also linked closely to respect and honor. However, another negative emotion has recently been accorded rising coverage, initially in moral psychology and related areas of cognitive science, but now too in moral philosophy: disgust (see Strohminger & Kumar, in press for a volume of essays on moral disgust by scientists and philosophers).

Disgust evolved to protect human beings from the threat of disease and infection. The most obvious elicitors of disgust are feces, grime, rotten food, and other noxious substances that are all vectors of disease and infection. Whereas anger is an "approach" emotion, disgust motivates withdrawal, avoidance, and distancing. This makes perfect sense as a response to vectors of disease. When one sees or smells rotten food, the best thing to do is to back away and/or rid oneself of it.

The primary function of disgust is disease-avoidance. However, disgust has also been recruited in moral cognition (Kelly, 2011; Rozin, Haidt, & McCauley, 2008; Tybur, Lieberman, Kurzban, & DeScioli, 2013). Disgust is commonly elicited by violations of ingroup norms and moral taboos. Outgroup members—people who speak, dress, and act differently—commonly evoke disgust. Some moral taboos—like those against meat eating and certain forms of "deviant" sex—proscribe substances and activities that are independently disgusting. But some do not, including certain religious taboos. Jonathan Haidt and his colleagues argue that disgust is entangled with a type of moral value that involves "impurity" (Graham et al., 2013; Haidt & Joseph, 2008). Often, it seems, behavior that "pollutes the soul" arouses moral disgust.

A striking fact about the burgeoning philosophical literature on moral disgust is that the most prominent work on the topic is skeptical (though see Jacobson, 2012; Plakias, 2013). Even among philosophers who admit that disgust plays a role in moral thought, the apparent consensus is that it *shouldn't* play any such role, i.e., that we should, as far as we can, eliminate disgust from moral thinking.

Kelly (2011) has led the empirical campaign against disgust in moral thought. He argues that the emotion is unreliable. With respect to disease and infection, Kelly points out, hypersensitivity pays. It is better to incur opportunity costs by avoiding seemingly foul substances that are actually benign than it is to interact with seemingly innocuous substances that actually harbor deadly pathogens. Consequently, the mechanisms underlying disgust generate frequent false positives for the sake of minimizing false negatives. Kelly argues that this hypersensitivity carries over into moral cognition. Thus, when we feel moral disgust, there is a strong possibility that we are experiencing a false positive, i.e., we feel repugnance toward something that is in fact unobjectionable.

Nussbaum (2004) is another prominent critic of moral disgust. She argues that the emotion should be expunged from moral thought not primarily because it is unreliable, but because it is harmful. Disgust is linked to ingroup norms, with the result that dominant groups tend to feel disgust toward people of color, LGBTQ individuals, the disabled, and the elderly. Because the emotion motivates withdrawal and avoidance, disgust tends to marginalize oppressed groups. Some researchers expose an apparent link between disgust and "depersonalization" (Harris & Fiske, 2007; Sherman & Haidt, 2011). When someone is "disgusting," evidence suggests that they are categorized as a mere thing rather than as a person. Worse, disgust also has the potential to facilitate genocide. In Nazi Germany, people of Jewish faith were cast as "vermin" and "parasites," thus fit to be excluded from society in ghettos and concentration camps and ultimately eliminated altogether. (Whether disgust promotes exile or violence—whether it motivates withdrawal or approach—may depend on to what extent it is mixed with more aggressive emotions like anger or hatred.)

Kelly and Nussbaum offer empirical support for skepticism about moral disgust, and their arguments are difficult to ignore. However, it is likely that philosophical skepticism is also fueled by a perceived connection between disgust and socially conservative norms and values. One of the few people to defend moral disgust is Kass (1997), who argues that moral disgust gives one insight into the moral unacceptability of new medical technologies like stem cell research and assisted reproduction. But a connection between disgust and conservativism is not merely anecdotal. Studies show that people who are highly sensitive to experiencing disgust are more likely to self-identify as conservatives (Inbar, Pizarro, & Bloom, 2009a; Inbar, Pizarro, Iyer, & Haidt, 2012), disapprove of gay people (Inbar, Pizarro, Knobe, & Bloom, 2009b), express punitive attitudes toward convicted criminals (Jones & Fitness, 2008), espouse ethnocentric prejudice (Navarette & Fessler, 2006), and even vote for John McCain in the 2008 U.S. presidential election (Inbar et al., 2012)

Liberal academics like Kelly and Nussbaum may be averse to disgust in part because of its apparent political affiliation. If so, then skeptical arguments against disgust threaten to be rationalizations, i.e., ways of reinforcing one's underlying political beliefs. Skeptics may reason implicitly: "Disgust leads to conservative opinions; conservative opinions are mistaken; so, disgust must be an untrustworthy force in moral thought." I'll suggest in a moment that the first, empirical premise is false: disgust does not lead to conservative opinions. Setting that aside, however,

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

empirically driven arguments against disgust might yet be sound even if critical attention to disgust is motivated by other reasons—that is, if disgust really is unreliable and harmful. Still, if debunking arguments against disgust are driven by rationalization of political beliefs, they deserve closer inspection.

Perhaps most telling: what sort of case could be made against anger, the other major negative emotion in moral thought, if similar critical attention were paid to it (see Kumar, in press-b)? First of all, is anger unreliable? Of course, anger can be misplaced, redirected onto easier targets. It's easier to "kick down" than it is to "kick up." Anger is also amplified by entitlement, which is notoriously modulated by racial, class, and gender privilege. Furthermore, is anger harmful? The emotion has a powerful connection to violence and injustice, likely stronger than disgust. Genocide typically stems primarily from anger towards an oppressed group. And yet, for all these problems, it seems unreasonable to recommend elimination of anger from moral thought. We feel resentment toward a friend for his betrayal, indignation toward a politician who sells out to lobbyists, outrage toward corporations that exploit their employees. These feelings of moral anger seem as appropriate as any moral feelings could be (for opposing views on anger see Nussbaum, 2016; Pereboom, 2001; Strawson, 1986).

A liberal cannot but find herself skeptical of psychological processes underlying socially conservative thought. However, it turns out that disgust is not an inherently conservative emotion. Some evidence does suggest a connection between disgust and ingroup norms and purity values. However, these moral categories are not exclusive to conservatives. Liberals simply have different characteristic outgroups (e.g., racists, nationalists), different sacred values (vegetarianism, environmentalism). Moreover, disgust is implicated in other types of moral norms, those that prohibit what I call "reciprocity violations" (Kumar, in press-b). People often feel disgust toward acts of dishonesty, cheating, and exploitation (Cannon, Schnall, & White, 2011; Chapman, Kim, Susskind, & Anderson, 2009; Gutierrez, Giner-Sorolla, & Vasiljevic, 2012; Hutcherson & Gross, 2011; Rozin, Lowery, Imada, & Haidt, 1999; Simpson, Carter, Anthony, & Overton, 2006; Tybur, Lieberman, & Griskevicius, 2009; for review see Chapman & Anderson, 2013). Of course, condemnation of these violations of reciprocity cuts across the political spectrum. In morality, disgust may have been initially recruited for ingroup norms and purity values, but disgust evinces much more flexibility than Kelly suggests (see Rozin, Markwith, & Stoess, 1997; Rozin & Singh, 1999). Kelly assumes that disgust is more "ballistic" than other emotions, and that this is partly why he distrusts the emotion. However, it is in virtue of disgust's flexibility that it has become tied to norms of reciprocity. Indeed, even in the domain of pathogens, disgust is flexibly attuned by relevant internal and environmental conditions. Just to take one example, women are more easily disgusted in the first trimester of pregnancy, a period during which their biological immune systems are suppressed (Fessler, Eng, & Navarette, 2005).

Anger and disgust are both elicited by moral violations, but if we look more closely, we find an interesting pattern (Kumar, in press-b). Anger is elicited principally by injury, theft, and coercion (call these "harm violations," for simplicity). Moral disgust, by contrast, is elicited principally by dishonesty, cheating, and exploitation ("reciprocity violations"). For example, many people are morally disgusted with a CEO who embezzles money from his company or with a corrupt politician who exploits her constituents. It's clear enough why each of these actions should evoke *some* negative moral emotion. But why should we feel anger in response to harm violations and disgust in response to reciprocity violations? That is, can we find any rationale for this pattern?

I believe that there is a rationale, and that it is grounded primarily in the different action tendencies that belong to each emotion. Anger is an approach emotion, and it makes sense to blame and confront people who inflict harm, since these people will continue to be a threat to you and to others unless they are confronted and some behavioral intervention is attempted. However, disgust is a withdrawal emotion, and it makes sense to avoid and exclude people who deceive, cheat, or exploit. People who commit reciprocity violations take advantage of trusting social relationships. Because disgust motivates withdrawal, avoidance, and exclusion, it not only protects victims from those who take advantage of them, but also punishes offenders by depriving them of social contact. Anger and disgust motivate different types of sanctions—direct vs. indirect—but both types of sanctions can serve to modulate immoral behavior. Pathogen disgust serves to protect human beings from pollution of the body, while moral disgust serves to protect human beings from pollution of constructive and beneficial social relationships.

Moral pollution, like bodily pollution, calls for quarantine. One reason that reciprocity violations threaten pollution is that they undermine the interpersonal trust that supports social relationships. Dishonesty, cheating, and exploitation inhibit trust. Another reason is that reciprocity violations tend to spread in a population. To a great extent, people tend to follow norms against deception, cheating, and exploitation only conditionally. If others around you begin to violate these norms, then you too are more likely to violate them. For example, participants recruited to play a public goods game tend to defect at higher rates once they notice that others around them are also defecting (Fischbacher, Gachter, & Fehr, 2001; Isaac & Walker, 1988). Disgust motivates quarantine, and therefore not only protects victims but also protects others from acquiring "moral diseases" (see Plakias, 2013).

We have now a justification for the role gained by disgust in moral thought. At this stage, however, we should not be too confident that disgust is a justified response to reciprocity violations. The reason, of course, is that the foregoing philosophical justification may well be a rationalization of our emotions. Perhaps no matter what pattern we had found reflected in our dispositions to feel anger and disgust, we would have been able to find some rationale for it. What's needed, then, is an *explanation* of this pattern. If the explanation dovetails with the proffered justification, we will have achieved more than a mere just-why story.

Learning studies in moral psychology have just begun and, so far, no one has carried out any work on the recruitment of moral disgust in response to reciprocity violations. We don't know yet precisely how flexible moral disgust is. However, what we know so far offers a hypothesis about how model-based learning can account for the pattern. Application of model-based learning in this context is similar to its application in the case of honor cultures. We construct an implicit model of our environment. This model assigns expected values for anger and disgust in response to various types of moral violations. We then experience these emotions, act on them, and the results of our actions are compared with the model's expectations. Discrepancies between expectations and results are used to revise the model. After a period of learning, we understand implicitly that anger has better results for some moral violations, disgust better results for others. In particular, we learn that disgust plays a functional role that is useful for tracking and responding to reciprocity violations. The learning is incremental, depends on other factors in development that reinforce it, and accrues to produce the observed pattern only after many generations.

A philosophical justification of moral disgust suggests a learning explanation of its etiology. Philosophical reasoning can be a source of empirical hypotheses, though whether we should give credence to this hypothesis about disgust depends on the results of empirical studies. If research were to confirm the hypothesis, it would vindicate the co-option of disgust toward reciprocity violations. But it would also support the instrumental justification of

this change, because it would show that the justification is supported by interaction with morally relevant aspects of the environment, and not just by our philosophical preconceptions.

## 8. Conclusion

In our first case study, an explanation for the psychology of moral luck buttresses a justification of moral luck in terms of the ability of outcome-oriented punishment to generate learning conditions that nurture pro-social feelings. In the second case study, an explanation of lowered anger thresholds in some honor cultures suggested a justification, and this justification in turn points to avenues for further scientific research. In the final case study, a justification for moral disgust leads to an empirical hypothesis about its source in learning. If these empirical hypotheses were to be substantiated, that would reinforce the corresponding philosophical justification of these innovations in moral thought. In moral learning, it would seem, explanation feeds justification and vice versa.

I have offered defeasible, instrumental vindications for three innovations in moral thought. The vindications are defeasible because although they justify a change in moral attitudes, they leave open whether further changes would have even better consequences. For non-ideal theorists, this sort of defeasible justification is the most we can hope for: we can understand why certain moral changes are justified; we can't hope to know what the ideal moral system is. Some philosophers may nonetheless believe that ideal theory plays an important role in moral philosophy. Even so, it is important to understand how moral progress is possible and how it can be better achieved.

Those unsympathetic to consequentialism may be disposed to reject the normative premise that producing better consequences is a reason for moral change. However, my arguments should put some pressure on non-consequentialists: if they are to avoid rationalization, they should argue that non-consequentialist justifications are reflected in the etiology of moral attitudes that they hope to justify. Otherwise, non-consequentialist justifications may be mere rationalizations, just as Greene alleges. My own vindications offered in this essay are consequentialist, but vindication arguments need not be so. If other researchers were to discover learning mechanisms that reflect non-consequentialist principles, the causal explanation for a new belief might support a non-consequentialist justification (see Gaus, 2011 for suggestions in this direction).

In some episodes of moral learning, explanation and justification coincide. Philosophers interested in justification should attend to empirical work that explains moral change. Otherwise, they might be pursuing rationalizations for the status quo. Likewise, psychologists interested in explanation should attend to philosophical work that justifies moral change. Some moral intuitions and feelings have a rationale, and if we think about why certain moral changes make sense, that can shed light on why they occurred.

Moral philosophy and moral psychology have had a rich and fruitful exchange of ideas over the past few decades. The joint exploration of moral vindications promises further profitable trade.

## References

Anderson, E. (2011). *The imperative of integration*. Princeton University Press.
Anderson, E. (2016). The social epistemology of morality: Learning from the forgotten history of the abolition of slavery. In M. Brady & M. Fricker (Eds.), *The epistemic life of groups*. Oxford University Press.
Appiah, K. (2010). *The honor code*. W.W. Norton.
Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy and Public Affairs, 37*, 293–329.
Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*, 171–195.
Boyd, R., & Richerson, P. (2005). *Not by genes alone*. University of Chicago Press.
Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352*, 220–224.
Buchanan, A., & Powell, R. (2016). Toward a naturalistic theory of moral progress. *Ethics, 126*, 983–1014.
Callaman, M., & Oakes, L. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development, 7*(2), 213–233.
Campbell, R., & Kumar, V. (2012). Moral reasoning on the ground. *Ethics, 122*, 273–312.
Cannon, P., Schnall, S., & White, M. (2011). Transgressions and expressions: Affective facial muscle activity predicts moral judgments. *Social Psychological and Personality Science, 2*, 325–331.
Carlsmith, K. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental and Social Psychology, 42*, 437–451.
Carlsmith, K. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research, 21*, 119–137.
Chapman, H., & Anderson, A. (2013). Things rank and gross in nature: A review and synthesis of moral disgust. *Psychological Bulletin, 139*, 300–327.
Chapman, H., Kim, D., Susskind, J., & Anderson, A. (2009). In bad taste: Evidence for the oral origins of moral disgust. *Science, 323*, 1222–1226.
Crockett, M. (2013). Models of morality. *Trends in Cognitive Science, 17*(8), 363–366.
Cushman, F. (2008). Crime and punishment: Differential reliance on causal and intentional information for different classes of moral judgment. *Cognition, 108*(2), 353–380.
Cushman, F. (2013a). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.
Cushman, F. (2015). Punishment: From intuitions to institutions. *Philosophy Compass, 10*(2), 117–133.
Cushman, F. (2011). Should the law depend on luck? In M. Brockman (Ed.), *Future science: 19 essays from the cutting edge*. Vintage.
Cushman, F. (2013b). The role of learning in punishment, prosociality, and human uniqueness. In R. Joyce, K. Sterelny, B. Calcott, & B. Fraser (Eds.), *Commitment and emotion, Vol. 2: Psychological and environmental foundations of cooperation*. MIT Press.
Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*, 6–21.
Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science, 17*, 1082–1089.
Demetriou, D. (2014). What should realists say about honor cultures? *Ethical Theory and Moral Practice, 17*, 893–911.
Domsky, D. (2004). There is no door. *Journal of Philosophy, 101*, 445–464.
Duarte, J., Crawford, J., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences, 38*, 1–13.
Dwyer, S. (2006). How good is the linguistic analogy? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind, Vol. 2: Culture and cognition*. Oxford University Press.
Fessler, D., Eng, S., & Navarette, C. (2005). Disgust sensitivity is elevated in the first trimester of pregnancy: Evidence supporting the compensatory prophylaxis hypothesis. *Evolution and Human Behavior, 26*, 344–351.
Fischbacher, U., Gachter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters, 71*, 397–404.
Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
Gaus, Gerald. (2011). *The order of public reason: A theory of freedom and morality in a diverse and bounded world*. Cambridge University Press.
Gopnik, A., & Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108.
Gottlieb, S. & Lombrozo, T. (in press). Folk theories in the moral domain. In K. Gray & J. Graham (Eds.), *The atlas of moral psychology*. Guilford Publications.
Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationism programme. *Proceedings of the Royal Society of London, Series B, 205*, 581–598.
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55–130.
Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.
Greene, J. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics, 124*, 695–726.
Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 3: The neuroscience of morality* (pp. 35–79). MIT Press.
Gutierrez, R., Giner-Sorolla, R., & Vasiljevic, M. (2012). Just an anger synonym? Moral context influences predictors of disgust word. *Cognition & Emotion, 26*, 53–64.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.
Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind*. Oxford University Press.
Harris, L., & Fiske, S. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience, 2*, 45–51.
Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias & philosophy: Volume 1, metaphysics and epistemology*. Oxford: Oxford University Press.

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005

Hutcherson, C., & Gross, J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*, 719–737.
Inbar, Y., Pizarro, D., & Bloom, P. (2009a). Conservatives are more easily disgusted than liberals. *Cognition and Emotion, 23*, 714–725.
Inbar, Y., Pizarro, D., Iyer, R., & Haidt, J. (2012). Disgust sensitivity, political conservativism, and voting. *Social Psychological and Personality Science, 5*, 537–544.
Inbar, Y., Pizarro, D., Knobe, J., & Bloom, P. (2009b). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*, 435–439.
Isaac, F., & Walker, J. (1988). Groups-size effects in public-goods provision—The voluntary contributions mechanism. *Qualitative Journal of Economics, 103*, 179–199.
Jacobson, D. (2012). Moral dumbfounding and moral stupefaction. In M. Timmons (Ed.). *Oxford Studies in Normative Ethics* (Vol. 2). Oxford University Press.
Jones, A., & Fitness, J. (2008). Moral hypervigilance: The influence of disgust sensitivity in the moral domain. *Emotion, 8*, 613–627.
Kass, L. (1997). The wisdom of repugnance. *New Republic*, 216.
Kelly, D. (2011). *Yuck! The nature and moral significance of disgust*. MIT Press.
Kumar, V. (2017). The weight of empathy (in preparation-a).
Kumar, V. (2017). Empirical vindication of moral luck (in preparation-b).
Kumar, V. (in press-a). The ethical significance of cognitive science. In Sarah-Jane Leslie & Simon Cullen (Eds.) *Current controversies in philosophy of cognitive science*. Routledge.
Kumar, V. (in press-b). Foul behavior. Philosophers Imprint.
Kumar, V. & Campbell, R. (2017). Why we are moral: The evolutionary foundations of moral progress. Unpublished book manuscript.
Kumar, V. & May, J. (in press). How to debunk moral beliefs empirically. In J. Suikkanen (Ed.), *New methods of ethics*.
Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology, 25*, 311–330.
Kumar, V., & Campbell, R. (2016). Honor and moral revolution. *Ethical Theory and Moral Practice, 19*, 147–159.
Mikhail, J. (2011). *Elements of moral cognition*. Cambridge University Press.
Mill, J. S. (1859). On liberty.
Moore, M. (1997). *Placing blame: A theory of criminal law*. Oxford: Clarendon Press.
Moore, M. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.
Nagel, T. (1979). *"Moral luck", in mortal questions*. Cambridge: Cambridge University Press.
Navarette, C., & Fessler, D. (2006). Disease avoidance and ethnocentrism: The effects of disease vulnerability and disgust sensitivity on intergroup attitudes. *Evolution and Human Behavior, 27*, 270–282.
Nichols, S. (2014). Process debunking and ethics. *Ethics, 124*, 727–749.
Nichols, S. (2015). *Bound: Essays on free will and responsibility*. Oxford University Press.
Nichols, S., Kumar, S., Lopez, T., Ayars, A., & H., Chan (2016). Rational learners and moral rules. *Mind and Language, 31*(5), 530–554.
Nisbett, R., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the south*. Westview Press.
Nussbaum, M. (2004). *Hiding from humanity: Disgust, shame, and the law*. Princeton University Press.
Nussbaum, M. (2016). *Anger and forgiveness: Resentment, Generosity, Justice*. Oxford University Press.
Pereboom, D. (2001). *Living without free will*. Cambridge University Press.
Pizarro, D., & Bloom, P. (2003). The intelligence of moral intuitions: Comment on Haidt (2001). *Psychological Review, 110*, 197–198.
Plakias, A. (2013). The good and the gross. *Ethical Theory and Moral Practice, 16*, 261–278.
Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813–859.
Richards, N. (1986). Luck and desert. *Mind, 65*, 198–209.
Royzman, E., & Kumar, R. (2004). Is consequential luck morally inconsequential? Empirical psychology and reassessment of moral luck. *Ratio, 17*, 329–344.
Rozin, P., Haidt, J., & McCauley, C. (2008). Disgust. In M. Lewis, J. Haviland-Jones, & L. Barrett (Eds.), *Handbook of emotions*. Guilford Press.
Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76*, 574–586.
Rozin, P., Markwith, M., & Stoess, C. (1997). Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological Science, 8*, 67–73.
Rozin, P., & Singh, L. (1999). The moralization of cigarette smoking in the United States. *Journal of Consumer Psychology, 8*, 339–342.
Sen, A. (2009). *The idea of justice*. Harvard University Press.
Sherman, G., & Haidt, J. (2011). Cuteness and disgust: The humanizing and dehumanizing effects of emotion. *Emotion Review, 3*, 1–7.
Simpson, J., Carter, S., Anthony, S., & Overton, P. (2006). Is disgust a homogeneous emotion? *Motivation and Emotion, 30*, 31–41.
Smith, A. (1759). The theory of moral sentiments.
Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind*. Cambridge: Cambridge University Press.
Strawson, G. (1986). *Freedom and belief*. New York: Oxford University Press.
Strohminger, N. & Kumar, V. (in press). *The moral psychology of disgust*. Rowmand Littlefield.
Thomson, J. J. (1993). Morality and bad luck. In D. Statman (Ed.), *Moral luck*. Albany: SUNY Press.
Tybur, J., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in the three functional domains of disgust. *Journal of Personality and Social Psychology, 97*, 103–122.
Tybur, J., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review, 120*, 65–84.
Walker, M. (1991). Moral luck and the virtues of impure agency. *Metaphilosophy, 22*, 14–27.
Wallace, R. J. (1996). *Responsibility and the moral sentiments*. Harvard University Press.
Wolf, S. (2001). The moral of moral luck. *Philosophic Exchange, 31*, 4–19.

Please cite this article in press as: Kumar, V. Moral vindications. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.05.005