Forthcoming: The Blackwell Companion to Experimental Philosophy

The adaptive logic of moral luck

Justin W. Martin & Fiery Cushman

Harvard University

Moral luck is a puzzling aspect of our psychology: Why do we punish outcomes that were not intended (i.e. accidents)? Prevailing psychological accounts of moral luck characterize it as an accident or error, stemming either from a re-evaluation of the agent's mental state or from negative affect aroused by the bad outcome itself. While these models have strong evidence in their favor, neither can account for the unique influence of accidental outcomes on punishment judgments, compared with other categories of moral judgment. Why might punishment be particularly sensitive to moral luck? We suggest that such sensitivity is easily understood from the broader perspective of punishment's ultimate adaptive goal: Changing others' behavior by exploiting their capacity to learn. This pedagogical perspective accounts for the exceptional influence of outcomes on punitive sentiments and makes predictions for additional moderators of punishment. We review evidence supporting the pedagogical hypothesis of punishment and discuss fruitful directions for future research.

On a January afternoon in 2014, Cynthia Garcia-Cisneros hit and killed two children with her car while driving home, but a few hours passed before she realized what she had done. In fact, Cynthia and her brother (a passenger in the car) only connected the dots when they learned about the deaths from the TV report. The children had been hiding in a leaf pile by the side of the road. Cynthia had driven through a pile of leaves on precisely the street where the children were killed. Both she and her brother noticed a jarring bump at the time, but each attributed it to sticks or compacted debris. In fact, what they had felt was 11-year-old Abigail and her 6-year-old sister Anna.[1]

When we described this story to an online sample of respondents, 94% indicated that Cynthia should be punished, assigning an average of 1-3 years in prison. But when we described a contrasting case in which there were no children in the leaf pile, and thus no harm done, 85% assigned no punishment at all. Not a single person assigned more than a small fine and probation. From a certain perspective, these judgments are remarkable: it is a matter of pure luck whether a leaf pile contains hiding children, and Cynthia's behavior was identical in both cases. Yet, the amount of punishment we assign is exquisitely sensitive to such chance variation in the magnitude of harm a person causes (Berg-Cross, 1975; Cushman, Dreber, Wang, & Costa, 2009; Cushman, 2008; Gino, Shu, & Bazerman, 2010; Mazzocco, Alicke, & Davis, 2004). Philosophers (Nagel, 1979; Williams, 1981) and legal theorists (Hall, 1960; Hart & Honore, 1959; McLaughlin, 1925) have long recognized this peculiar feature of human moral judgment, which is often termed moral luck. Our aim in this essay is to explain it.

In the philosophical literature, moral luck encompasses a broader range of types of luck, including resultant, circumstantial, constitutive and causal luck. Here, we focus in particular on resultant luck, or luck in the way things turned out. Currently, two explanations for moral luck prevail in the psychological literature. The first approach posits that accidental harms prompt us to reconsider whether a person acted reasonably in the first place. In other words, after we find out that Garcia-Cisneros' behavior lead to the death of two children, we think to ourselves, "It really isn't safe to drive through leaf piles; you never know if a child might be hiding in one." This explanation is often referred to as *hindsight bias*, because it attributes clairvoyant caution with the benefit of hindsight (Baron & Hershey, 1988; Tostain & Lebreuilly, 2008; see also Alicke & Davis 1989; Mazzocco et al., 2004; Young, Nichols, & Saxe, 2010). The second approach posits that the emotional salience of an accidental harm amplifies our moral judgments. In other words, the negative affect associated with the death of two children baises us to judge the harmdoer more harshly.

Both of these models have strong evidence in their favor, yet neither provides a complete explanation of

[1] Details taken from http://blogs.seattletimes.com/today/2014/01/teen-sentenced-to-probation-in-oregon-leaf-pile-hit-and-run/

moral luck. Critically, moral luck is not equally influential across different kinds of moral judgment: It plays an especially strong role in judgments of punishment and blame, but a significantly weaker role in judgments of moral wrongness and moral character (Cushman, 2008). Thus, in a case like Cynthia's most people would say that her behavior (i.e. driving through a leafpile) was minimally morally wrong, regardless of the outcome caused. But, as the data from our online respondents suggest, they would assign punishment in a manner sensitive to the outcome. This unique effect of accidental outcomes on punishment judgments is not explained by hindsight bias or negative affect, both of which ought to apply generally across different categories of moral judgment. It might suggest a comparable error or bias that would be predicted to apply uniquely to judgments of punishment. But, we pursue another possibility: perhaps moral luck is not a cognitive error or emotional bias at all, but rather an adaptive design feature tuned to the specific functional demands of punishment. Is there a hidden logic to moral luck?

Intention versus outcome in moral judgment

A convenient way to formalize moral luck is to contrast two putative sources of influence on moral judgment: A person's intended action, versus the outcome that they cause. For instance, consider the case of a potential poisoning (Young, Cushman, Hauser, & Saxe, 2007). Grace is making coffee for her friend. As she adds a white powder to the coffee, she either believes that it is sugar or rat poison (thus establishing her intent), and it either is sugar or rat poison (thus establishing the outcome). The key cases are those in which her beliefs do not match reality: Either she attempts to put rat poison in the coffee but actually adds sugar, or else she accidentally puts rat poison in the coffee believing it was sugar. By assessing moral judgments across these cases it is possible to establish the relative influence of intent versus outcome upon moral judgment.

This 2×2 design has the virtue of simplicity, but it obscures two important details. First, there is no model of moral judgment according to which the mere *occurrence* of a harmful outcome triggers moral condemnation; rather, what is required is the perception that a person is *causally responsible* for that outcome. So, when we speak of the influence of "outcomes" upon moral judgment, what we really mean is causal responsibility for the outcome, scaled by the degree of harm caused[2]. Establishing the relevant

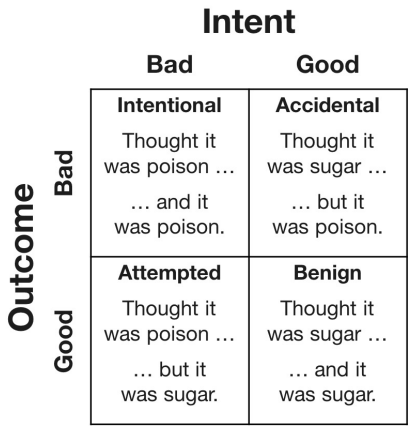

Figure 1: A factorial combination of intent and outcome yields four basic categories of conduct. Adapted from Young et al., 2007.

standard of causal responsibility is a matter of much research and controversy (Cushman & Greene, 2011; Hart & Honore, 1959; Lombrozo, 2010; Wolff, 2012; Appeals Court of New York, 1961) but it is undisputed that ordinary people rely on attributions of causal responsibility of some kind during moral judgment.

Second, in most circumstances a person's thoughts do not make them a target of punishment. For instance, a person who contemplates murdering his uncle after a rude comment, but who takes no steps towards carrying out this plan, is not yet punishable under the law—nor do most people think she should be (Robinson & Darley, 1995). What about for other categories of moral judgment: Was it *morally wrong* for her to form that intention? Perhaps, but it is unambiguously more wrong to act upon the intention. In other words, this connection between intent and action is clearly a target of condemnation. Thus, when we speak of contrasting "outcome" and "intent", what we really mean to describe is a contrast between causal responsibility for harm and an action undertaken with intent to harm.

This framework was first employed in seminal work by the developmental psychologist Jean Piaget. In one study, for instance, he described two characters and asked children who was naughtier: Marie, who as a result of well-intentioned behavior accidentally cut a large hole in her dress, and Margaret, who as a result of bad behavior cut a small hole in her dress (Piaget, 1965). That is, he asked children to evaluate whether an agent with bad intentions who caused a small harm was morally worse than an agent with good intentions who caused a large harm, pitting these two factors against each other. Below the ages of 7-10 years children tended to base their judgment on outcome (finding Marie naughtier, despite recognizing her good intent), but older children eventually based their judgments on intent (finding Margaret naughtier,

[2] In a more complex design these factors could be dissociated: A person could be causal responsible for a slight harm or a very extreme harm; or, a person could lack causal responsibility in either case.

despite the better outcome). From these results, Piaget influentially concluded that children initially focus on objective states of the world (outcomes) and only later in life, presumably with greater education and intelligence, come to appreciate the significance of subjective states of the mind (intentions). Recent research using more sophisticated methods has even extended these findings and suggests that intent-based judgment may emerge substantially sooner than Piaget's research indicated (Armsby, 1971; Farnill, 1974; Nobes, Panagiotaki, & Pawson, 2009; Yuill & Perner, 1988), even as early as the first year of life (Hamlin, Mahajan, Liberman, & Wynn, 2013; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Hamlin & Wynn, 2011; Hamlin, 2013).

Consistent with Piaget's findings about older children, many studies employing the outcome×intent framework find that intent plays a dominant role in adults' moral judgments (Cushman et al., 2009; Cushman, 2008; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Young, Cushman, Adolphs, Traniel, & Hauser, 2006; Young et al., 2007). Across a range of contexts and a diversity of studies, intentional harms are viewed as more deserving of condemnation than accidental ones. For instance, in a study asking participants to make different types of moral judgments and employing different contexts, differences in levels of intent accounted for between 63% and 84% of variability in participants' responses (Cushman, 2008).

The importance of intentions is further underscored by the activation profile of a network of brain regions responsible for inferring and assessing what other people think, believe and desire (theory of mind; see e.g. Castelli, Happé, Frith, & Frith, 2000; Fletcher, Happe, Frith, & Baker, 1995; Saxe & Kanwisher, 2003). Critically, these regions are robustly recruited during moral judgment, apparently by incorporating information about intentions. When we judge Grace in the situations above, the right temporoparietal junction (RTPJ) shows above baseline activity for both failed attempts to harm and accidental harms, consistent with its role in processing intentions, but greater levels of activation for the former compared to the latter case (Young et al., 2007). This reveals the specific nature of this region's role in moral judgment: The RTPJ is recruited more when participants need to use information about an agent's mental state to condemn them. The causal nature of this role has been confirmed by studies employing transcranial magnetic stimulation (TMS). TMS allows researchers to transiently impair the functioning of a brain region, providing a strong causal test of a region's role in a particular task or process. Here, applying TMS to the RTPJ selectively reduced moral condemnation in cases of attempted harm, compared with intentional and accidental harms (Young, Camprodon, et al., 2010).

In summary, current research emphasizes the dominant role of intent in moral judgment alongside a significant but much weaker role for outcomes (Cushman et al., 2009; Cushman, 2008; Young, Camprodon, et al., 2010; Young et al., 2006, 2007). Given the overwhelmingly dominant role of intention in moral judgment, it is not surprising that researchers have consistently advanced psychological models that frame the role of outcomes as a bias rather than as a design feature of moral judgment, and that they have often attempted to explain that bias in terms of mental-state reasoning. We consider the two families of bias model that have been most popular, and then ask whether they provide a complete account of moral luck.

Hindsight bias

The most influential psychological explanation of moral luck is hindsight bias. According to this model, when a bad outcome occurs as the result of somebody's behavior it causes us to reassess whether the person acted reasonably in the first place. As first noted by Walster (1966), there are two subtly distinct versions of this hypothesis. One version posits that people reassess the actual mental state of the harmdoer ("Cynthia probably considered the possibility of children playing in the leaves") while the other version posits that people reassess their standard of reasonable conduct against which a person's behavior is measured ("A reasonable person wouldn't take the risk of driving through a leaf pile").

Consistent with the first model of hindsight bias, Fincham (1982) found that outcome severity predicts mental state attributions (i.e. worse outcomes are viewed as more intentional). Recently, Nobes and colleagues (2009) leveraged this result to propose a reinterpretation of Piaget's finding that young children show greater sensitivity to outcomes in their moral judgments. While testing children between the ages of 3 and 8, Nobes and colleagues found that negligence played a large role in judgment: less than intentions but larger than outcomes. This is consistent with a hindsight bias model, and with other evidence that children use mental state information in social assessments earlier in life than previously thought (Hamlin, Mahajan, et al., 2013; Hamlin, Wynn, & Bloom, 2007; Hamlin & Wynn, 2011).

Walster's own research (1966) favored the second model of hindsight bias. Specifically, she found that adults judged an agent whose careless action lead to a worse outcome to be more responsible than an agent who caused a mildly bad outcome. And, they endorsed a harsher standard of the precautions an agent should have taken when he caused a severe versus mild outcome. Other work corroborates this, finding that outcomes are viewed as more predictable in retrospect

("a reasonable person should have expected that"), which leads to greater condemnation when the outcome is bad (Baron & Hershey, 1988; Tostain & Lebreuilly, 2008). Recent evidence further probed this relationship between the reasonableness of a person's beliefs and the outcomes they cause (Young, Nichols, & Saxe, 2010). Consider Mitch, who is getting his son ready for a bath when the phone rings. He tells his son to stay out of the tub and answers the phone. In one case, Mitch's son is obedient and so remains safe. In another condition, Mitch's belief that his son will remain where he left him is false and his son gets in the tub, thankfully remaining OK. In neither condition does a bad outcome obtain, but Mitch receives significantly more moral condemnation in the latter case, where he holds a false belief about what his son will do. In fact, the difference in blame between these cases is greater than the difference between the false belief case and a third case, where Mitch's belief is false, his son gets in the tub and drowns (Young, Nichols, et al., 2010). Here, the status of the agent's belief (true or false) appears to matter even more than if he causes a bad outcome, which some have taken to suggest that outcome bias is really more about the correspondence between the beliefs of the agent and the outcome (Young, Nichols, et al., 2010).

Motivated reasoning

Although less discussed in the psychological literature on moral luck, there is a second and quite distinct model for how outcomes influence moral judgment. Specifically, it may be the case that the negative affect produced by the outcome itself biases the process of moral judgment. The most direct evidence favoring this view comes from a series of studies by Carlsmith and colleagues (Carlsmith, Darley, & Robinson, 2002; Darley, Carlsmith, & Robinson, 2000). They find that more severe harms lead to greater degrees of punishment and that this effect is mediated by "moral outrage"—i.e., people feel outraged at the bad outcome, and this colors their assessment of the morally responsible party.

Carlsmith and colleagues' evidence is ambiguous between two interpretations. One possibility is that outcome effects are simply a bias. For instance, it may be that people use their affective reactions as a source of information when making moral judgments (see e.g. Damasio, 2008; Greene, 2008; Slovic, Finucane, Peters, & MacGregor, 2007) and thus misattribute the negative emotion associated with the harmful outcome to the behavior of the causally responsible agent. Alternatively, it might be that moral judgments are in fact designed to integrate information about outcomes, and emotions such as moral outrage directly motivate the assignment of blame to parties who are causally responsible for harm.

These approaches differ at a mechanistic level. While misattribution can occur with even minimal causal connection between an agent and an outcome, the second approach places relatively greater importance on such a connection: Only when the agent is sufficiently causal is emotion generated by a negative outcome a useful guide as to whether they should be punished or not. But, they also differ fundamentally at an adaptive level: the first approach makes no commitment to the adaptive value of outcome-based punishment, while the second approach assumes such value. As we will see, combining mechanistic and adaptive levels of analysis offers the opportunity for a new perspective on the psychological basis of moral luck.

The two process model

Moral luck can produce tremendous cognitive dissonance, and this fact requires explanation. Consider again the case of Cynthia Garcia-Cisneros. On the one hand, it feels unjust to let her go unpunished for killing two children; on the other hand, it feels unjust to punish her differently than another person who might have driven through a leaf pile harmlessly. This internal cognitive conflict animates both works of art (e.g. the role of unintended – and undesired – outcomes in *Oedipus*) and fuels philosophical debate (Lewis, 1987; Nagel, 1979; Williams, 1981). What, then, are its psychological origins?

On its face, cognitive conflict would seem to indicate that we have at least two distinct routes by which we arrive at moral condemnation, and that these yield opposing judgments in cases of moral luck. One of these would condemn on the basis of causal responsibility for harm, and another would condemn on the basis of intent to harm. The critical word here is "condemn", and it bears explaining why. As we have seen, it is entirely uncontroversial that both causal responsibility for harm and intent for harm contribute to the process of moral judgment. If both of these factors were necessary inputs into a mechanism of moral judgment wherein they were integrated in order to condemn, this would be a single process model. What makes the alternative a "two process" model is that it posits separate mechanisms that are each capable of condemnation on their own—one based solely on causal responsibility, the other based solely on mental states. According to this view, it is the fact that both processes are sufficient to achieve condemnation that allows them to conflict. Several lines of evidence favor such a model.

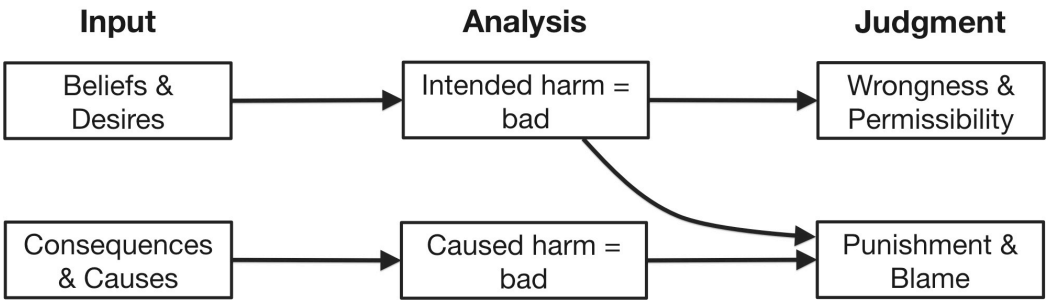

Figure 2: A two process model of moral judgment. Adapted from Cushman, 2008.

A key starting point is the fact that there is something special about the relationship between moral luck and punishment. Specifically, while judgments of punishment and judgments of wrongness both rely to a large extent on information about an agent's intentions, punishment determinations are additionally sensitive to the outcome caused (Cushman, 2008). This sensitivity to outcomes when judging deserved punishment is corroborated by numerous other studies (Berg-Cross, 1975; Cushman et al., 2009; Gino, Moore, & Bazerman, 2009; Gino et al., 2010; Mazzocco et al., 2004), as has the fact that judgments of wrongness depend principally on intent (Hebble, 1971; Imamoglu, 1975; Piaget, 1965; Wellman, Cross, & Bartsch, 1986; Young et al., 2007).

This discrepancy between judgments of punishment and other categories of judgment, like wrongness and character, casts immediate doubt upon hindsight bias and motivated reasoning as complete explanations for moral luck. After all, these accounts ought to apply equally to any category of moral judgment. Hindsight bias predicts either a re-evaluation of the subjective likelihood of an outcome after it is observed or an adjustment in the standard of responsible conduct, which in turn leads to greater condemnation. But, it is unclear why either a shift in likelihood or standard of conduct would occur more when assessing deserved punishment than when assessing how wrongly the person acted. Motivated reasoning suggests that observing a negative outcome leads to moral outrage. A side effect of this emotional reaction is a biasing of moral judgment. Again, however, such a bias should be felt equally for judgments of wrongness and moral character as for punishment. The same analysis applies to other putative explanations for moral luck, such as that outcomes are relied upon because they are more observable, and therefore more reliable, than intentions (Schächtele, Gerstenberg, & Lagnado, 2011).

Shortly, we will offer an adaptive explanation for why punishment judgments might be especially sensitive to accidental outcomes. For the moment, however, our point is that punishment judgments are more strongly influenced by the process of evaluating causal responsibility for harm, and this is at least consistent with the hypothesis that causal and mental state factors contribute to distinct processes of moral judgment.

This model has been extended to research with young children, asking whether the discrepancy between punishment and wrongness judgments emerges early in life (Cushman, Sheketoff, Wharton, & Carey, 2013). As first documented by Piaget, this study found that children based their judgments on outcomes more than intentions at 4 years old, but on intentions more than outcomes by 8 years old. Additionally, during this same period they also show increasing differentiation in the criteria for punishment versus wrongness judgments. But the most important evidence is that intent-based moral judgment emerged first in judgments of wrongness and only subsequently in judgments of punishment. In other words, children first begin to exculpate accidental harmdoers for one category of moral judgment (wrongness), and this subsequently constrains the way they make another category of moral judgment (punishment). This provides further evidence for the presence of two distinct processes—one present early on in life and another that arises relatively later.

Two other sources of evidence provide strong support for the hypothesis that causal and mental state factors are supported by distinct and competitive processes. First, the two-process model predicts that the causal process and the mental state process can operate simultaneously and thus competitively interact. Evidence for a competitive interaction comes from a contrast between two unlikely cases of attempted murder. In the first case, the would-be murderer sprinkles poppy-seeds on a rival's salad at a banquet, believing the rival is allergic to the seeds. But he was misinformed: the rival is allergic to hazelnuts, and so is unharmed. This is a standard case of attempted but failed harm. The second case is identical except that the salad happens to have been made with hazelnuts and thus, for reasons entirely unconnected to the would-be-murder, the rival happens to die. Remarkably, people assign significantly less punishment to the attempted murderer in the second case—despite

the fact that the main difference is the *addition* of a death! This result can be explained by competition between an intent-based judgment process (which would operate equally in both cases) and an outcome/responsibility-based judgment process. The latter process is exclusively engaged in the second case, because it involves a harmful outcome that does not occur in the standard attempt case. But, the process of assigning causal responsibility points *away* from the would-be murderer, competitively blocking intent based blame.

A second compelling source of evidence in favor of the two-process model comes from the study of moral judgment under cognitive load. Recent evidence suggests that effortful cognition is needed to integrate information about an agent's mental state when making a moral judgment, but not in order to integrate information about causal responsibility (Buon, Jacob, Loissel, & Dupoux, 2013). By having participants perform verbal shadowing (e.g. follow along with and repeating sentences aloud) while watching a series of videos involving moral situations, the authors were able to induce cognitive load. Under load, participants were selectively impaired in their ability to incorporate mental state information into moral judgment (either to condemn an intentional agent for his bad belief or to exculpate an accidental agent for her lack of bad belief). Importantly, this result was not due to a selective impairment in inferring intent: Participants were able to distinguish between agents who intended harm versus those who accidentally caused it. Rather, the impairment was specific to using such mental state information in moral judgment. This selective impairment localized to a process of intent-based moral judgment lends further support to its dissociation from an alternative causal process.

The logic of luck: A pedagogical hypothesis

The evidence for a distinct "causal" process of moral judgment, and for its unique role in determining judgments of deserved punishment, raises a deeper question: What is its adaptive value? In order to understand why moral luck is a distinctive psychological feature of punishment, however, we must start with a more basic question: Why punish at all?

In one manner or another, most adaptive analyses of punishment converge on the sensible view that we punish people in order to modify their future behavior (Clutton-Brock & Parker, 1995). Through punishment, we align others' behavior with our preferences, promoting future prosocial behavior and furthering our own interests (Boyd & Richerson, 1992; Fehr & Gächter, 2002; Henrich & Boyd, 2001). Critically, this explanation depends on the capacity of organisms to modify their behavior following punishment – in essence, to learn. Empirical evidence suggests that such a theory holds weight. For instance, Fehr and Gachter (2002) found that cooperation rates in a public goods game remain high throughout repeated interactions when punishment is possible. And, although cooperation rates fall when no threat of punishment exists, when the possibility of punishment is introduced midway through a multi-round session, rates of cooperation gradually increase, consistent with the idea that punishment (or the threat of it) causes behavioral change.

We refer to this theory as the pedagogical hypothesis of punishment. It suggests that the distinctive psychological structure of our intuitive punitive sentiments can be understood by appeal to punishment's function as a method of teaching. Two important points are worth highlighting. First, this is a claim specifically about the psychological foundations of punishment, not the design of current legal systems. Of course, it could be the case that legal statues dictating moral luck are grounded in ordinary people's punitive sentiments. Alternatively, moral luck could be a feature of law that stems from practical or policy considerations. In any event, we are concerned with the psychology of punishment. Second, we are not claiming that people actually compute the pedagogical value of punishment. The mechanistic basis for punishment may be quite different from its adaptive rationale. Indeed, past work suggests that punitive motivations are mostly retributive in nature (Carlsmith et al., 2002; Carlsmith, 2006; Darley et al., 2000), with little role for reasoning from a utilitarian perspective, either consciously or unconsciously. Rather, the pedagogical hypothesis attempts to explain the ultimate adaptive function of blind retribution in terms of its tendency to modify the behavior of social partners. In other words, retributive anger is an adaptive heuristic: It is engaged in circumstances that typically allow an offending party to learn from punishment.

This hypothesis provides a key insight into why punishment is more sensitive than other moral judgments to the presence of a bad outcome. Specifically, punishing bad outcomes—even when accidental—may be necessary in order to teach people to exercise greater care in the future. When a one year old throws her food on the floor, for instance, she is not trying to harm anyone and is not acting maliciously; indeed, she may not even know that she acted badly. But, the parent who refrains from punishment in this case has little hope of seeing his daughter's manners improve. In contrast, outcome-based punishment will cause the baby to adjust her future behavior to avoid such outcomes, even if this punishment is not "deserved" based on her lack of intent. Of course, punishing bad intent may *also* fit the functional demand of modifying others' behavior. And, as we have seen, punishment judgments are certainly sensitive to a person's intention alongside their accidental outcomes.

Critically, however, punishment can be effective and even necessary when a bad outcome occurs in the absence of a negative intent.

This allows us to easily reconcile the effect of moral outrage on punishment at a mechanistic level with the function of pedagogy at an adaptive level. Because moral outrage is directed at persons causally responsible for harm, is sensitive to their mental states, and scales with the degree of harm caused, it will tend to guide retributive punishment in a manner that effectively teaches social partners appropriate standards of conduct.

The pedagogical hypothesis makes further predications about when we might expect punishment to be employed and when it might be withheld. First, if punishment is to play an effective pedagogical role, then individuals who are punished must be able to associate their behavior with punishment. Without such an association, a costly punishment will not change the behavior in question and thus reap no reward. Thus, we might expect punishment to be endorsed more when a punishing agent makes explicit reference to the behavior motivating the punishment compared to when such a reference is absent. In a similar vein, we might expect punishment to be endorsed more when the agent originally harmed is the one doing the punishing and thus known to the perpetrator, facilitating the association between offending behavior and punishment. If Steve harms John, John's subsequent punishment of Steve is comprehensible and Steve will likely realize why he is being punished. But, if Steve harms John and then John pays Carl to punish Steve, Steve may not realize why he is being punished and will not learn to treat John better.

Similarly, we should expect punishment to be particularly sensitive to manipulations of time between the offending behavior and subsequent punishment. It is well known that feedback is most effective when presented at a minimal delay from behavior. If pedagogy is at the heart of punishment, then whether or not the agent is able to learn from punishment should change whether it is applied. Thus, after a long delay, assessed punishment will drop, mirroring legal statutes of limitation. Such a drop would be in contrast to judgments reflecting the agent's mental state or character. To the degree that these judgments do not serve a communicative function (but rather an evaluative role), assessments of an agent's mental state or character will be relatively less sensitive to any time lapse: An agent will be rated as having just as bad of character five years after a crime, but less deserving of punishment.

Finally, the pedagogical hypothesis predicts that decisions to punish will be sensitive to the degree to which the behavior in question is typical for an individual. A behavior performed once in a lifetime that results in a harm is in much less need of being changed than a behavior that will be performed day in and day out. To the degree that an agent will never engage in a particular harmful behavior again, we thus might expect punishment to decrease, as pedagogy is not as necessary. A worker on his final day before retiring may deserve less punishment for being clumsy and spilling some office supplies than a new worker who will face such a situation many more times in the future.

Testing pedagogy: Control and luck

We conclude with a case study: A recent experiment of ours that tests the adequacy of the pedagogical hypothesis as an account for moral luck. If moral luck is not merely a bias or error, but rather serves the adaptive function of pedagogy, then it should be restricted to cases where an agent's action is *controllable*. For instance, if a young child spills milk in a manner she could have controlled, then punishing this accident can successfully modify her future behavior. On the other hand, if a young child spills milk in a manner she could not have controlled—for instance, while sneezing—then punishing the accident would have little value.

Although it is well established that people assign less punishment to uncontrollable actions (Alicke, 2000; Cushman et al., 2009; Darley et al., 2000; Robinson & Darley, 1995), past research provides little empirical insight into whether non-controllable actions are less susceptible to moral luck, specifically. One plausible interpretation of these theories is that we forgive uncontrollable harms because a person who has no control over their action presumably did not *intend* to cause harm. Yet, moral luck arises in cases of accidents, where a person's intent is benign anyway—punishment is driven instead by mere causal responsibility for harm.

Thus, the key to our approach was to test for the influence of control in cases of *accidental harm*—where an agent's intention is good, but the outcome that they cause is bad. To illustrate, one vignette involved a doctor prescribing medicine for a patient (other vignettes involved an investment banker, workers on an oilrig, etc.). Two medications fit the patient's illness. One was relatively more likely to kill the patient and the other relatively more likely to cure him. However, the doctor would obtain a prestigious publication if the patient died. In our key case of interest, the doctor prescribes the good medication, but the patient is unlucky and the medication causes him to die anyway. Thus, the harm done to the patient was accidental. In order to create a case in which the doctor has no control over her choice, we simply stated that only the good medication was in stock at the hospital—thus, the doctor was forced to prescribe it.

If the doctor's lack of control primarily influences punishment because it diminishes the perception of intent, then a lack of control should be accompanied by *increased* punishment because it would obviate the agent's positive intent (recall that she chose the good medicine). However, the pedagogical hypothesis predicts that moral luck will be diminished in cases of accidental harm. From a pedagogical perspective, there is no utility in punishing the agent without control because the lesson they learn cannot influence their behavior.

Our results support this latter possibility: Across 3 studies, we found that agents causing an accidental harm received greater punishment when they had control than when they did not (Martin & Cushman, 2014). This result is striking: The agent with control has actually demonstrated good intentions (by making the prosocial choice) but ends up receiving greater punishment than an agent without control who has demonstrated no such positive intentions. Moreover, we found this counterintuitive result only for judgments of punishment, but not for judgments of moral character. This is consistent with the hypothesis that the punishment of accidental outcomes has a distinctive psychological structure that fits well with the general adaptive function of punitive behavior: To modify others' behavior in circumstances where they have sufficient control, and to do so by punishing those who are casually responsible for harm. Of course, such a concern is not warranted in this case: The doctor took all possible steps to bring about a good outcome and so no behavior needs to be modified. It is precisely this "misfire" of the system that illustrates the underlying representations people use when making punishment judgments, one of which is that those who cause accidental harm should be punished, but only if their behavior is controllable.

As we expected, further study of this effect showed that it is tightly integrated with the "causal process" of moral judgment (i.e., as distinct from the "mental-state process"), consistent with the strong influence of this process upon punishment judgments. Specifically, we found that our manipulation of control lead to changes in perceived causal role in the harm coming about, but no changes in perceived intent. In other words, we forgive a doctor who kills a patient with the only drug available not only because she didn't intend harm, but also because she seems not to have even been causally responsible for the harm. Consistent with this evidence, Alicke (2000) proposes two independent pathways for control to influence moral judgment: One by way of intent, and another by way of causation. In sum, then, the "causal process" of moral judgment appears to encompass a discrete set of computational properties that exhibit an adaptive fit to the pedagogical hypothesis.

Conclusion

There is a logic to luck. Although prevailing models of hindsight bias and motivated reasoning posit that moral luck is itself an accident, several lines of evidence demonstrate that these explanations cannot be complete. Moral luck is a distinctive feature of punishment, it arises from a discrete process of moral judgment focused on causal responsibility, and it exhibits a strong match to the adaptive function of using punishment to modify others' behaviors. In the case of Cynthia Garcia-Cisneros, presumably punishment on the basis of the outcome would lead her to be more cautious in the future. Moreover, punishing such accidents could allow others to learn vicariously. And yet, the idea of her being punished is still deeply unsettling.

A normative question still remains open: *Ought* we to punish others for their accidental behaviors? There is no obvious way to answer this question on the basis of scientific facts alone, but those facts may constrain the space of likely answers. The essence of our argument is that moral luck is a heuristic. We assign punishment retributively; but, this heuristic evolved because of the practical value of modifying others' behavior. This suggests two key avenues for philosophical inquiry. First, is pedagogy a legitimate basis for punishment? Second, if so, does the punishment of accidents remain the most effective path towards deterrence, or can we achieve the ultimate goals of punishment without resorting to moral luck?

Acknowledgments

We are grateful to Jake Davis, Mark Ho, Ryan Miller and Jonathan Philips for valuable feedback on earlier versions of this manuscript. This work was supported by National Science Foundation Award No. 1228380 to FC and by National Science Foundation Graduate Research Fellowship Grant No. DGE1144152 to JWM.

References

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574. doi:10.1037/0033-2909.126.4.556

Alicke, M., & Davis, T. (1989). The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology*, *25*(4), 362–377. doi:10.1016/0022-1031(89)90028-0

Appeals Court of New York. (1961). Palsgraf v. Long Island Railroad Company, 248 N.Y. 339, 162 N.E. In H. Morris (Ed.), *Freedom and Responsibility* (pp. 285–91). Stanford, CA: Stanford University Press.

Armsby, R. (1971). A reexamination of the development of moral judgments in children. *Child Development*, *42*(4), 1241–8.

Baron, J., & Hershey, J. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569–79.

Berg-Cross, L. (1975). Intentionality, degree of damage, and moral judgments. *Child Development*, *46*(4), 970–4.

Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *195*, 171–195.

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition*, *126*(2), 149–155. doi:10.1016/j.cognition.2012.09.006

Carlsmith, K. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. doi:10.1016/j.jesp.2005.06.007

Carlsmith, K., Darley, J., & Robinson, P. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. doi:10.1037/0022-3514.83.2.284

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*(3), 314–25. doi:10.1006/nimg.2000.0612

Clutton-Brock, T., & Parker, G. (1995). Punishment in animal societies. *Nature*, *373*, 209–216.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–80. doi:10.1016/j.cognition.2008.03.006

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, *4*(8), e6699. doi:10.1371/journal.pone.0006699

Cushman, F., & Greene, J. D. (2011). The philosopher in the theater. In M. Mikulincer & P. R. Shaver (Eds.), *Social psychology of morality: The origins of good and evil*. APA Press.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. doi:10.1016/j.cognition.2012.11.008

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. Putnam Publishing.

Darley, J., Carlsmith, K., & Robinson, P. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior*, *24*(6), 659–683.

Farnill, D. (1974). The effects of social-judgment set on children's use of intent information. *Journal of Personality*, *42*(2), 276–289. doi:10.1111/j.1467-6494.1974.tb00674.x

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–40. doi:10.1038/415137a

Fletcher, P., Happe, F., Frith, U., & Baker, S. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128. doi:10.1016/0010-0277(95)00692-R

Gino, F., Moore, D., & Bazerman, M. (2009). No harm, no foul: The outcome bias in ethical judgments.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless+harmless=blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93–101. doi:10.1016/j.obhdp.2009.11.001

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *The Neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). Cambridge, MA US: MIT Press.

Hall, J. (1960). *General principles of criminal law*. Indianapolis: Bobbs-Merrill.

Hamlin, J. K. (2013). Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*(3), 451–74. doi:10.1016/j.cognition.2013.04.004

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: infants prefer those who harm dissimilar others. *Psychological Science*, *24*(4), 589–94. doi:10.1177/0956797612457785

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226. doi:10.1111/desc.12017

Hamlin, J. K., & Wynn, K. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences*, *108*(50). doi:10.1073/pnas.1110306108

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559. doi:10.1038/nature06288

Hart, H. L. A., & Honore, T. (1959). *Causation in the law* (1st ed.). Oxford: Clarendon Press,.

Hebble, P. W. (1971). The development of elementary school children's judgment of intent. *Child Development*, *42*(4), 1203–15.

Henrich, J., & Boyd, R. (2001). Why people punish defectors. Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*(1), 79–89. doi:10.1006/jtbi.2000.2202

Imamoglu, E. O. (1975). Children's Awareness and Usage of Intention Cues. *Child Development*, *46*(1).

Lewis, D. (1987). The Punishment That Leaves Something to Chance. In *Proceedings of the Russellian Society* (pp. 81–97). University of Sydney.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–32. doi:10.1016/j.cogpsych.2010.05.002

Martin, J. W., & Cushman, F. (2014). Why we forgive what can't be controlled.

Mazzocco, P. J., Alicke, M., & Davis, T. L. (2004). On the Robustness of Outcome Bias: No Constraint by Prior Culpability. *Basic and Applied Social Psychology*, *26*(2-3), 131–146. doi:10.1080/01973533.2004.9646401

McLaughlin, J. A. (1925). Proximate cause. *Harvard Law Review*, *39*(2), 149.

Nagel, T. (1979). *Mortal questions*. Cambridge: Cambridge University Press.

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology*, *104*(4), 382–97. doi:10.1016/j.jecp.2009.08.001

Piaget, J. (1965). *The Moral Judgment of the Child. Psychoanalytic Review*. New York: Free Press.

Robinson, P., & Darley, J. (1995). *Justice, liability, and blame: Community views and the criminal law*. Oxford: Westview Press.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. doi:10.1016/S1053-8119(03)00230-1

Schächtele, S., Gerstenberg, T., & Lagnado, D. (2011). Beyond outcomes: The influence of intentions and deception. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. (2007). The affect heuristic. *European Journal of Operational Research*, *177*, 1333–1352. doi:10.1016/j.ejor.2005.04.006

Tostain, M., & Lebreuilly, J. (2008). Rational model and justification model in "outcome bias." *European Journal of Social Psychology*, *279*(October 2006), 272–279. doi:10.1002/ejsp

Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, *3*(1), 73–9.

Wellman, H. M., Cross, D., & Bartsch, K. (1986). Infant Search and Object Permanence: A Meta-Analysis of the A-Not-B Error. *Monographs of the Society for Research in Child Development*, *51*(3).

Williams, B. (1981). Moral luck. In *Moral luck* (pp. 20–39). Cambridge: Cambridge University Press.

Wolff, P. (2012). Causal Pluralism and Force Dynamics. In B. Copley, F. Martin, & N. Duffield (Eds.), *Forces in grammatical structures: Causation between linguistics and philosophy*. Oxford University Press.

Young, L., Camprodon, J. J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*(15), 6753–8. doi:10.1073/pnas.0914826107

Young, L., Cushman, F., Adolphs, R., Traniel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, *6*(1), 1–2. doi:10.1163/156853706776931312

Young, L., Cushman, F., Hauser, M., & 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. doi:10.1073/pnas.0701408104

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*(3), 333–349. doi:10.1007/s13164-010-0027-y

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology*, *24*(3), 358–365. doi:10.1037/0012-1649.24.3.358