Cognition 147 (2016) 133–143

Contents lists available at ScienceDirect

Cognition

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

Why we forgive what can't be controlled

Justin W. Martin *, Fiery Cushman

Department of Psychology, Harvard University, 33 Kirkland Street, William James Hall, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 29 May 2015
Revised 14 October 2015
Accepted 17 November 2015

Keywords:
Moral psychology
Morality
Punishment
Control
Intentional action
Causation

ABSTRACT

Volitional control matters greatly for moral judgment: Coerced agents receive less condemnation for outcomes they cause. Less well understood is the psychological basis of this effect. Control may influence perceptions of intent for the outcome that occurs or perceptions of causal role in that outcome. Here, we show that an agent who *chooses* to do the right thing but accidentally causes a bad outcome receives relatively more punishment than an agent who is *forced* to do the "right" thing but causes a bad outcome. Thus, having good intentions ironically leads to greater condemnation. This surprising effect does not depend upon perceptions of increased intent for harm to occur, but rather upon perceptions of causal role in the obtained outcome. Further, this effect is specific to punishment: An agent who chooses to do the right thing is rated as having better moral character than a forced agent, even though they cause the same bad outcome. These results clarify how, when and why control influences moral judgment.

1. Introduction

Charles Whitman murdered his wife and mother on a July night in 1966. The following morning he continued the killing spree, climbing a clock tower and using a large arsenal of rifles to indiscriminately murder passersby below. His spree left 13 dead and 32 wounded.[1] If Whitman was in control of his behavior then nobody could be more deserving of punishment. Yet, in an unusual suicide note Whitman professed love for his family and regret for the deeds he was about to commit. He described the recent onset of strangely violent thoughts, and requested an autopsy to determine whether there was an abnormality in his brain. There was: His autopsy revealed a growing tumor that impinged on a cluster of subcortical structures. Suppose, then, that Whitman's behavior was in some sense beyond his control. As heinous as his actions were, would this fact change our desire for retribution?

Many past studies suggest that it would: Agents who lack control over their behavior receive less condemnation for harms they cause (Alicke, 2000; Cushman, Dreber, Wang, & Costa, 2009; Darley, Carlsmith, & Robinson, 2000; Robinson & Darley, 1995). But, what is it about lacking control that lessens moral judgment? In other words, what is the psychological basis of this effect? One intuitively appealing possibility is that control impacts moral judgment through representations of intentionality. If a person strikes another during a seizure, for instance, their lack of control indicates that they likely did not cause harm intentionally. This inference follows because behavioral control implies a correspondence between intention and outcome, while a lack of control makes a mismatch possible. Returning to Whitman's case, potentially we forgive him because viewing his actions as uncontrollable leads us to assume that he lacked a culpable mental state—i.e., an intention, desire, motive, etc. to kill. Indeed, this connection between intentionality and control is well documented (Malle, Guglielmo, & Monroe, 2014; Weiner, 1995).

Yet, Whitman's case seems a poor example of this mechanism. His behavior was intentional in any ordinary sense of the word: He meticulously planned and then executed an attack on nearly four-dozen people, murdering 13 of them. Rather, it feels intuitively as if the tumor "made" Whitman murder, by forcing him or robbing him of alternative courses of action. In other words, Whitman's lack of control seems to deprive him of causal responsibility for the crime. It wasn't really Whitman who did it—his diseased brain did.

This illustrates an alternative possibility: That control influences moral judgment by modifying ascriptions of causal responsibility. Past research clearly demonstrates that moral judgment is sensitive to a person's role in causing harm, in addition to the role played by their intent to cause harm (Cushman, 2008; Guglielmo, Monroe and Malle, 2009; Piaget, 1965; Weiner, 1995; Young, Cushman, Hauser, & Saxe, 2007). Yet, there is less *prima face* appeal to the possibility that we forgive uncontrollable action because we don't hold an actor causally responsible for it. After all, we routinely apply the concept of causal responsibility to inanimate

* Corresponding author.
E-mail address: justinmartin@g.harvard.edu (J.W. Martin).

[1] Details taken from: http://www.theatlantic.com/magazine/archive/2011/07/the-brain-on-trial/308520/.

http://dx.doi.org/10.1016/j.cognition.2015.11.008
0010-0277/

objects and events to which the notion of "control" simply does not apply. For instance, we judge a storm to have caused a forest fire without positing that the storm has "control" over the lightning. By analogy, can't a person who lacks control over their behavior still be causally responsible for the harm that follows from it?

Answering this question depends upon a careful decomposition of behavioral control. At first blush, we might suppose that Whitman was a cause of the murders if, absent Whitman, the murders would not have occurred, in just the same way that lightning might cause a forest fire (no lightning, no fire). In this sense it is trivially true that Whitman was the cause. Following this logic, if Whitman is to be excused for his behavior on the basis of a brain defect, it could not be because he failed to *cause* the deaths of the people who he shot.

Yet, much past research indicates that people represent the causal pathway from a person to the world in a more nuanced manner. Specifically, they distinguish between the causal link from a person (e.g., Whitman) to their behavior (shooting) and the subsequent causal link from the behavior (shooting) to an outcome (deaths) (Alicke, 2000). According to this model, it is conceptually possible that Whitman's behavior caused the deaths, and yet "Whitman" did not cause his behavior. Clearly much hinges on the boundaries of personal identity. Is Whitman to be identified with his entire physical body, including the nervous system? If so, he clearly plays a causal role in the production of all of his behaviors. Or, alternatively, is Whitman to be identified with a limited portion of his mental capacity—specifically his will, the capacity for volitional control? On this view, it is possible for Whitman's body to have fired the shots without "Whitman" being the true cause of this behavior (construed as their "will"). And, of course, if he didn't cause his behavior, then he didn't cause the deaths that resulted from it. This latter conception formalizes the intuitive notion that we are not causally responsible for events over which we have no control.

Consistent with this possibility, Knobe and Nichols (2011) find that people judge an agent to be the cause of his own controllable actions (e.g., moving his hand away from a bee) but not to be the cause of his own uncontrollable actions (e.g., trembling in the presence of a bee). Applying a similar idea to the moral domain, Phillips and Shaw (2014) find that people attribute less blame to a person who is intentionally manipulated into performing a harmful action than to a person who is manipulated unintentionally. Critically, blame is reduced because people view the manipulated agent as less causally responsible for the harm she produced. Although Phillips and Shaw did not directly test ascriptions of control, their preferred interpretation is that people perceive the manipulated agent as being controlled by the other agent (and thus, presumably, lacking in control over themselves).

In sum, then, while it seems likely that there are cases in which we forgive uncontrollable actions because we do not think the agent intended harm, could it also be the case that we forgive such actions because we do not believe the agent is even causally responsible for them? Prior work has suggested that the intent and causation pathways are not mutually exclusive possibilities. For instance, Alicke (2000) proposes two independent pathways for control to influence moral judgment: One by way of intent, and another by way of causation. The causation hypothesis remains untested, however, because past studies have not successfully dissociated causation from intent when assessing the influence of control on moral judgment. Our aim is to accomplish this dissociation.

1.1. Experimental logic

Dissociating the influence of causation and intention requires situations of a particular type: an agent must be causally responsible for harm that they did not intend, and yet still be a viable target for moral judgment. Cases of moral luck, studied in both the psychological and philosophical literatures (Cushman, 2008; Nagel, 1979; Williams, 1981; Young et al., 2007), present such an opportunity.[2] In one variety of moral luck, a person acts with good intentions but accidentally brings about a bad outcome (Cushman, 2008; Young et al., 2007). Despite their good intentions, such agents are often held to deserve punishment (Cushman, 2008). This punishment of accidental outcomes depends on the attribution of causal responsibility to the agent. Here, we make use of such cases and explore how the punishment of accidental outcomes responds to greater versus lesser degrees of control.

The logic of our design is best illustrated through a specific example. Consider a doctor who can choose between two medications in order to save her patient. Without medication, the patient will certainly die. Medication A has only a 33% chance of killing the patient, while medication B has a 66% chance of death. The doctor will be able to publish in a prestigious medical journal if either medicine fails, however, so she has an incentive to choose the bad medication. Fortunately she is a good doctor and chooses the good medicine (A); unfortunately, the patient is among the unlucky minority who dies. Consistent with prior research, we expect that participants will assign some degree of punishment to this doctor because she is causally responsible for death, despite her choice of the best possible action. That is, we expect to observe the phenomenon of moral luck.

Against this backdrop, the critical question is how participants will judge a case that proceeds identically except for one detail: As the doctor is deciding, she finds out that only medication A is available in her office. This doctor still performs a good action (choosing medication A) that leads to a bad outcome (killing the patient), but her control over the outcome is diminished because she lacks an alternative course of action. (Note that we accomplish reduced control by a manipulation of counterfactual alternatives; below, we discuss the advantages and disadvantages of this approach in greater detail).

On the one hand, if control influences moral judgment exclusively by modifying participants' attributions of intent, then people should judge the doctor who lacks control as harshly—perhaps even more harshly—than the doctor who exercises control. After all, the doctor with a choice of medications has demonstrably good intentions: She chooses the best medication for the patient when a selfish alternative is available. In contrast, there is ambiguity about the intentions of the doctor who lacks control: Maybe she would have chosen the good medication, but on the other hand maybe she would have chosen the bad medication had it been available in order to boost her publication record.

On the other hand, if control influences moral judgment in part by modifying participants' attributions of causal responsibility, then people should judge the doctor with control more harshly than the doctor who lacks control. After all, both doctors contributed to the death of the patient—a highly negative outcome. Whoever is judged more causally responsible for this negative outcome will tend to receive more blame. In this case, it would be the agent who possesses control over her action.

This prediction of the causal responsibility hypothesis is so peculiar that it deserves special attention. Stated in the abstract, it seems appropriate that a person who causes harm would receive more punishment if she has more control over her action. Yet, in the specific case of an accidental harm, control is exercised in order to do the right thing: For instance, the doctor chooses the good drug. According to the causal responsibility hypothesis, it is pre-

[2] Although there are multiple types of moral luck (constitutive, circumstantial, causal and resultant), we focus on resultant moral luck, the type most often studied in the psychological literature.

cisely her free choice of the best possible course of action that dooms her to punishment, when this optimal choice leads by chance to an undesirable result. In contrast, a person who is *forced* to do the right thing is insulated from such punishment. Such a pattern of judgment would provide strong evidence uniquely in favor of the hypothesis that control can exert an effect on moral judgment without modifying participants' attributions of malicious mental states.

The logic of our experimental design requires participants to make moral judgments of a kind that depend both on attributions of intent and causal responsibility. Past research indicates that judgments of deserved punishment are sensitive to both of these factors (Berg-Cross, 1975; Cushman et al., 2009; Cushman, 2008; Gino, Shu, & Bazerman, 2010; Martin & Cushman, 2015; Mazzocco, Alicke, & Davis, 2004), while other categories of moral judgment, such as judgments of moral wrongness and moral permissibility, are not (Hebble, 1971; Imamoglu, 1975; Piaget, 1965; Wellman, Cross, & Bartsch, 1986; Young et al., 2007). Consequently, our studies focus on judgments of deserved punishment.

1.2. Control and counterfactual representation

As described above, our experimental design manipulates an agent's control over their behavior by depriving them of a possible course of action. For instance, we posit that an agent who has two routes to get home has more control than an agent who has only one route to get home. This approach is consistent with a large literature indicating a crucial role for counterfactual representation in causal judgment (Byrne, 2002, 2016; Roese, 1997; Walsh & Sloman, 2011; Wells & Gavanski, 1989), especially when assigning causal responsibility to intentional actions (Lombrozo, 2010). Nevertheless, it is a relatively weak and indirect manipulation of behavioral control. A stronger manipulation would involve directly intervening on the decision process of the agent; for instance, in Whitman's case, his brain defect apparently directly influenced his process of decision-making.

We adopted the weaker approach because it is also more conservative with respect to our central hypothesis. In the case of a brain defect, an alternative cause of behavior is explicitly introduced. The presence of a salient alternative cause may bias participants to discount the causal influence of the agent (i.e., of her will). In contrast, when an agent's behavioral control is manipulated by restricting the set of alternative behaviors available to them, this avoids the explicit introduction of an alternative cause. Thus it presents a particularly conservative test of the hypothesis that the manipulation of perceived volitional control influences moral judgment partially through an effect on causal representation.

Our experimental design shares certain features with past studies of moral judgment, control, and counterfactual representation. Most notably, a study by Williams and colleagues (Williams, Lees-haley, & Price, 1996) found that mutability of an outcome (the degree to which it could be mentally altered) influenced the causal role and blame assigned to an agent involved in it. One set of participants was told that a cab driver refused to pick up 2 paraplegic passengers and then safely drove over a bridge. When the couple later attempted to drive over the same bridge, it tragically collapsed and the couple drowned in the water below. Another set of participants was given the same information, with the exception that after refusing the couple, the cab driver did not make it safely across the bridge, but rather drove off of it (and survived). Again, the bridge later collapses when the couple attempts to drive over it, and they drown. That is, in one case the outcome could have been prevented, whereas in the other it was inevitable (i.e., the couple would have died regardless of the cab driver's behavior). Both groups of participants were asked about the degree of control exercised by the cab driver, how much of a causal role he played in the couple's deaths and the amount of punitive damages that should be levied against the cab driver. The difference in outcome mutability influenced all three judgments: When the outcome was avoidable, the cab driver was viewed as playing a greater causal role, as possessing more control and was assigned greater punitive damages (Williams et al., 1996).

In essence, this study and others like it (Walsh & Sloman, 2011; Williams et al., 1996) manipulate the influence of an agent's behavior on an event by altering whether the event would have occurred even absent the agent's behavior—that is, whether the event was overdetermined by the circumstances. They therefore use the manipulation of counterfactuals to adjust the perceived causal relationship between behavior and an outcome, whereas our study is designed to target the prior causal relationship between volition and behavior. In simple terms, it is a difference between whether *something else could have happened* (outcome mutability) and whether *something else could have been done* (volitional control). In Experiment 3 we make this distinction explicit and experimentally dissociate influences of each kind.

2. Experiment 1

Our first experiment assessed the influence of control in cases of moral luck. In order to do this, we designed stimuli in a fully crossed design manipulating three factors: The performance of a better vs. worse action, the occurrence of a bad vs. neutral outcome, and greater versus lesser control over action. In order to manipulate the level of control that the agent exerted over her action, we altered the availability of alternative courses of action to that agent. For instance, in the case described above, the doctor either has the choice of two medications to prescribe (full control), or else there is only one medication available (diminished control).

Our greatest interest is in two cells of this design: more versus less control in cases of accidental harm (a good action leads to a bad outcome). This case allows us to diagnose the potential role of control in modifying ascriptions of causal responsibility. Critically, any influence of control on the moral judgment of accidental harms could not be mediated by the attribution of malicious intent, because accidents are defined precisely by the lack of harmful intent.

Of course, this is not to say that control does not influence perceptions of intentionality, or that such an influence is inert on moral judgment. Rather, we would expect this relationship to manifest in case of attempted harm: those in which a bad action occurs but there is no bad outcome. Here, manipulations of control can only impact perceptions of the agent's intent, since there is no harm to be causally responsible for.

Thus, the design of Experiment 1 allows us to test for two independent and non-exclusive pathways by which control might influence moral judgment: Via the attribution of causal responsibility (in cases of accidental harm), and via the attribution of malicious intent (in cases of attempted harm).

2.1. Methods

2.1.1. Participants

Seven hundred thirty-one participants were recruited through Amazon Mechanical Turk and all provided informed consent. Participation was restricted to US residents who were native speakers of English. Level of education varied widely (from some high school to Ph.D), as did socioeconomic status. All participants received monetary compensation for their participation. All procedures were approved by the Brown University Institutional Review Board.

2.1.2. Materials

We designed vignettes according to a 2 (negative vs. positive action) × 2 (negative vs. positive outcome) × 2 (more control vs. less control) design, and then replicated this eight-cell design across four independent vignette contexts (see Supplemental Methods). To illustrate, the text of one of the scenarios is provided below. This scenario involves an accident, in which the outcome is negative and the action is positive; we then vary the level of control possessed by the agent.

> A doctor working in a hospital has a patient who is having hearing problems. This patient has two, and only two, treatment options. With Option A, there is a 66% chance the patient makes a full recovery. With Option B, there is a 33% chance that the patient makes a full recovery. Whichever option the doctor chooses, it happens that if the patient fails to make a full recovery then the doctor will be able to publish a very prestigious paper, which would be great for his career. On the other hand, if the patient makes a full recovery, the doctor won't be able to publish this paper at all, but his reputation will improve.

In the More Control (MC) condition, the next sentence read: "The doctor chooses to pursue Option A". In the Less Control (LC) condition, the sentence read: "While deciding, the doctor is informed that the patient is allergic to the medicine necessary for Option B and thus the doctor is forced to proceed with Option A." Thus, in the MC condition the agent had imperfect control over the situation – her choice impacted which outcome was most likely, but only imperfectly determined the eventual outcome. In the LC condition, the agent's available choices were relatively more constrained, and therefore she had reduced control over the eventual outcome. (In Experiment 2 we validate that participants in fact ascribe less control to the agent in such cases).

2.1.3. Procedure

Participants read 4 vignettes and rated how much punishment the protagonist deserved in each. Each of the 4 vignettes was presented in one of the 4 possible action × outcome combinations, the order of which was counterbalanced across participants. The agent either selected the option that was likely to help the target or to harm the target, and the target either was helped or was harmed. Each participant was randomly assigned to have all 4 vignettes presented in the MC or the LC condition.

For each trial, participants read the vignette and were asked to make a single judgment: How much punishment does [protagonist] deserve? They responded to this prompt on a 9-point scale anchored at 1 (No punishment at all) and 9 (Extreme punishment). For all ratings, reaction times were collected and used in participant screening. All participants also completed attention check questions after the last of the 4 scenarios, for use in participant screening. Based on our exclusionary criteria, data from 86 participants (out of 731; 11.7%) were discarded. Lastly, participants responded to a series of optional standard demographic items and were debriefed.

2.2. Results

To first examine responses to all conditions we used mixed effects regression. Fixed effects included action, outcome (both within-subjects factors) and control (between-subjects factor), as well as all possible two-way interactions and the three-way interaction. We included random intercepts for participant and scenario, as well as random slopes within each for action and outcome. This permitted us to model a maximal random effects structure and allow for correlation between random effects, given the number of observations per participant (Barr, Levy, Scheepers, & Tily, 2013). Because our data are discrete (a 1–9 integer scale) rather than continuous, we used ordinal mixed effects regression (specifically, we fit a cumulative link mixed model), implemented in R using the ordinal package (Christensen, 2015).

Results are provided in Table 1. Consistent with prior research on punishment judgments we found substantial contributions of both outcome and action on moral judgment. We additionally find a large and significant impact of control on judgments of punishment. Of greater relevance to our present study, the significant interactions between action and control, and also between outcome and control, motivate our selective attention to cases of attempted and accidental harms.

To assess the relationship between control and an agent's intentions, we focus on cases of attempted but failed harm. Here, participants assigned greater punishment to an agent who had more control over their behavior than one who had less control (More Control [MC]: mean = 3.26 (SEM = 0.13); Less Control [LC]: 1.69 (0.08); $t(643) = 10.1$, $p < 0.001$, 95% CI of mean difference [1.26, 1.89], $d = 0.80$, Fig. 1). In the absence of a bad outcome, such punishment likely stems from the fact that participants use the agent's choice as a way of understanding their intentions. An agent who freely chooses a bad action, thus exhibiting bad intent, is punished more than a forced agent, whose true intentions are more ambiguous. This accords well with prior work on the relationship between control and intent and is in line with the default route by which control is believed to influence moral judgment – by way of intent (Malle et al., 2014; Weiner, 1995)

We next test whether control can additionally influence moral judgment through attributions of causal responsibility by focusing on our primary case of interest: Accidental harms, in which an agent performed a good action that resulted in a bad outcome. We found that participants assigned greater punishment to accidental harms when the agent had more control over their action than when they had less control (MC: 2.93 (0.14); LC: 2.43 (0.11); $t(643) = 2.774$, $p < 0.05$, 95% CI [0.15,0.85], $d = 0.22$, Fig. 1). This suggests that, in the context of moral luck, control can exert an effect on punishment independently of intent.

2.3. Discussion

Our first study reveals a pattern of judgment consistent with two sources of influence of control on moral judgment: One through attributions of intent, and another through attributions of causal responsibility. More precisely, our findings show that the influence of control on moral judgment likely cannot be reduced to either one of these paths of influence alone. The finding that greater control leads to greater punishment in cases of attempted harm suggests an influence of control on attributions of malicious intent. Meanwhile, the finding that greater control leads to greater punishment in cases of accidental harm suggests an influence of control on attributions of causal responsibility.

It is possible, however, that control simply exerts its own, independent influence on moral judgment. According to this model, the effect of control is not mediated either by attributions of causal responsibility or by attributions of malicious intent. Our subsequent experiments aim to provide additional evidence that the manipulation of volitional control affects moral judgment via attributions of causal responsibility.

The effect that we report in Experiment 1 is small both numerically (approximately 0.5 points on a 1–9 scale) and in its effect size (Cohen's $d = 0.22$). This is expected because we manipulated the perception of volitional control indirectly, by limiting the set of behaviors available to the agent, rather than directly, by describing an agent operating without the capacity for volitional control. This was necessary in order to manipulate perceived control in a manner independent from intentional action. Nevertheless, we note

Table 1
Results from Experiment 1. Punishment response was regressed on predictors for action, outcome and control and all possible interactions. We find strong main effects for action, outcome and control, as well as interactions between control and action, and between control and outcome.

Fixed effects	Estimate	Std. error	*z* value	*p* value
Mixed-effects regression results				
Outcome	1.46	0.28	5.23	<0.001
Action	1.30	0.36	3.63	<0.001
Control	1.26	0.12	10.49	<0.001
Action × outcome	0.11	0.18	0.59	0.55
Action × control	1.96	0.19	10.44	<0.001
Outcome × control	0.61	0.17	3.51	<0.001
Action × outcome × control	0.48	0.34	1.42	0.15

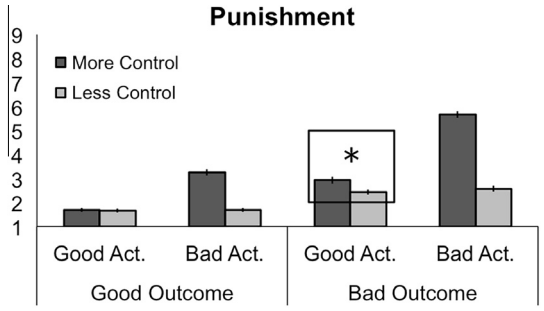

Fig. 1. Results from Experiment 1. Participants read scenarios drawn from each cell of a 2 (action) × 2 (outcome) design, and were assigned to either the More Control or Less Control condition. Plotted are averages for each cell of the full design. Boxed are the results from the critical moral luck condition. Error bars are SEM.

that approximately 40% of the overall effect of moral luck we observed in this design (that is, the punishment of an accidental outcome) was eliminated by our indirect manipulation of volitional control.

3. Experiment 2A

In Experiment 2A we directly probe participants' perceptions of causal responsibility and malicious intent. Our specific interest is how these attributions respond to manipulations of control over an agent's behavior in cases of accidental harm. As a manipulation check, we also probe whether our manipulation of control in fact influences participants' attributions of control.

A direct test of mediation requires analyzing the change in correlation between a predictor and a dependent variable when a mediating variable is included in the model – if the correlation is weaker when the proposed mediator is included versus absent, this is evidence in favor of that variable playing a mediating role (Baron & Kenny, 1986; Preacher & Hayes, 2008). However, such a technique cannot be applied in the current situation because there is a bi-directional relationship between the dependent variable (moral judgment) and the proposed mediators (attributions of intentionality and causality). Specifically, while differences in causality and intentionality influence moral judgment, making a moral judgment also influences how causal or intentional an agent is viewed (Alicke, 1992; Knobe & Fraser, 2008; Knobe, 2003, 2005).

Thus, we aim to develop more circumstantial evidence regarding the relationship between control, causation and moral judgment. For either intent or causation to serve as a mediator it must, by definition, show a relationship with control (the predictor variable). That is, manipulations of control must influence any mediating variable, in addition to the dependent variable. If attributions of either intentionality or causality are not affected by changes in an agent's control, then that variable is ruled out as a potential mediator. In other words, the approach we employ in Experiment 2 is to conduct an experimental test that could provide evidence unambiguously falsifying our model.

Our prediction is that the manipulation of control in cases of accidental harm will influence participants' attributions of causal responsibility, such that they assign greater causal responsibility in cases of greater control. Just as Whitman's tumor might reduce our perception that he is causally responsible for his actions, participants might conclude that the doctor forced to choose a particular drug was therefore less causally responsible for the patient's death, resulting in less punishment relative to the doctor with more control.

In contrast, we predict that if our manipulation of control has any influence on attributions of intent, it will lead to decreased attributions of malicious intent in cases of greater control. This is because the agent uses her control to engage in a manifestly beneficent action.

There is, however, some basis for the opposite prediction. As noted above, prior work shows that moral blame influences our inferences of intent (e.g. Knobe, 2003). Because participants in Experiment 1 assigned greater punishment to accidental harmdoers who acted with greater control, they may also increase their attribution of malicious intent to these agents, despite the doctor's choice of the beneficial medication. Indeed, prior work on hindsight bias has shown such an influence (Baron & Hershey, 1988; Tostain & Lebreuilly, 2008; Young, Nichols, & Saxe, 2010). According to this model, we would expect greater attributions of malicious intent as control increases. Of course, these models are not mutually exclusive, although their influences would tend to cancel each other out.

3.1. Method

Participants ($N = 1517$) were recruited under the same parameters as Experiment 1. Each read 4 vignettes, all in the critical moral luck condition (where the agent performed a good action that resulted in a bad outcome). Each participant was randomly assigned to view and respond to all 4 vignettes presented in the MC or the LC condition. Vignette order was counterbalanced across participants. Participants were randomly assigned to one of four rating types: Intent ($n = 385$, "To what extent did [protagonist] intend [outcome]?", 1 = "Did not intend at all", 9 = "Completely intended"), causal role ($n = 375$, "To what extent did [protagonist] cause [outcome]?", 1 = "Did not cause at all", 9 = "Completely caused"), degree of control ($n = 378$, "How much control did [protagonist] have over the outcome of this situation?", 1 = "No control at all", 9 = "Complete control") or punishment ($n = 379$, same as Experiment 1). All participants completed attention check questions after the last of the 4 scenarios, for use in screening. Based on our exclusion criteria, data from 124 participants (8.2%) was discarded. Participants finished by completing a series of demographic questions and being debriefed.

3.2. Results

We replicated our prior finding: Greater control lead to greater punishment in cases of accidental harm (More control [MC]: 2.73 (0.12); Less control [LC]: 2.33 (0.11); $t(352) = 2.395$, $p < 0.05$, 95% CI [0.07, 0.73], $d = 0.26$). We also demonstrated that our manipulation of control does indeed lead to differences in perceived control (MC: 4.72 (0.12); LC: 2.96 (0.11); $t(342) = 10.827$, $p < 0.001$, 95% CI [1.43, 2.07], $d = 1.17$). Critically, we found an interaction between control condition and judgments of causality versus intentionality ($F[1, 691] = 5.40$, $p < 0.05$). This difference was driven by the fact that the agent with more control was viewed as having a greater role in causing the harm (MC: 3.64 (0.13); LC: 2.93 (0.11); $t(348)$

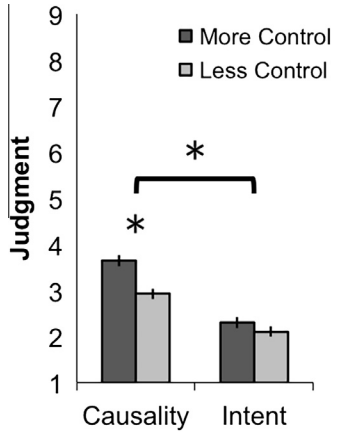

Fig. 2. Results from Experiment 2A. Participants read scenarios of accidental harm, all either in the More Control or Less Control condition. Plotted are average responses for the extent to which the agent caused the outcome to occur and the extent to which the agent intended the outcome to occur. Error bars are SEM.

= 4.251, $p < 0.001$, 95% CI [0.38, 1.04], $d = 0.45$), but not greater intent to harm (MC: 2.31 (0.10); LC: 2.10 (0.09), $t(343) = 1.446$, $p = 0.149$, 95% CI [−0.07, 0.48], $d = 0.16$, Fig. 2).

3.3. Discussion

As intended by our manipulation, we found that participants attributed greater control to the agent who chose between two available options compared with the agent who was forced to "choose" the only available option. We also replicated our principle finding, that an agent with greater control receives more punishment than one with less control in cases of moral luck. Consistent with the hypothesis that perceived control influences attributions of causal responsibility, we found that participants also assigned greater causal responsibility to the agent who had greater control over her action. In contrast, we did not find significantly greater attributions of malicious intent to the agent who had greater control over her action. Collectively, these data support the hypothesis that the influence of perceived control on the judgment of accidental harms is mediated at least in part by attributions of causal responsibility. The present research indicates that control modifies attributions of causal responsibility, and past research indicates that attributions of causal responsibility strongly influence judgments of deserved punishment (Berg-Cross, 1975; Cushman et al., 2009; Cushman, 2008; Gino et al., 2010; Mazzocco et al., 2004).

4. Experiment 2B

Notably, and contrary to our expectations, Experiment 2A revealed a slight trend toward greater attributions of malicious intent for the agent who had greater control over her actions. Two possible explanations for this finding stand out. First, greater attributions of intentionality might stem from (unprompted) moral judgments of the agent, which influence subsequent perceptions of the agent. That is, consistent with blame-early accounts of moral judgment (Alicke, 1992, 2000) and work on the relationship between moral judgment and intentions (Knobe, 2003, 2005, 2006), participants' negative moral evaluations of the agent may have influenced how intentional of a mental state they subsequently attributed to her. (On this hypothesis, increased blame directed at the agent with control might arise due to a direct effect of control on moral judgment, a mediated effect due to causal attribution, or other processes).

Alternatively, the specific phrasing we chose for our dependent measure may have influenced the results. Participants rated to what extent the agent in each situation either intended for the outcome to occur, or caused the outcome to occur. For causality, the scale is clear: The agent could be maximally causal, or not causal at all (or somewhere in between). When assessing intentionality, however, such a scale may have constrained participants' responses. In theory, the agent could have fully intended the bad outcome to occur, or could not have had a harmful intention, or could indeed have possessed an intention for the *good* outcome to occur. On our scale, however, participants were unable to express this latter possibility. That is, they were precluded from indicating that the agent had a good intent. This may be especially problematic in the case of the agent who chooses the good drug with control over her actions, because a natural inference is that her intentions were positive.

Thus, in Experiment 2B, we explore whether the tendency to attribute slightly more malicious intent to an agent with more control is an artifact of the method used in Experiment 2A, by employing a scale that allows participants to express a wider range of possibilities.

4.1. Methods

Participants ($N = 401$) were recruited under the same parameters as Experiment 2. Each read 4 vignettes, all in the critical moral luck condition (where the agent performed a good action that resulted in a bad outcome). Vignette order was counterbalanced across participants. Each participant was randomly assigned to have all 4 vignettes presented in the MC or the LC condition. Participants evaluated the protagonist in each vignette using a 1–9 scale with the following text above the indicated points: 1 = "Definitely intended for [positive outcome]", 3 = "Possibly intended for [positive outcome], 5 = "Not sure", 7 = "Possibly intended for [negative outcome]", 9 = "Definitely intended for [negative outcome]". All participants completed attention check questions after the last of the 4 scenarios, for use in screening. Based on our exclusion criteria, data from 26 participants (6.5%) was discarded. Participants finished by completing a series of demographic questions and being debriefed.

4.2. Results

Allowing participants to respond about a protagonist's intentions for either a negative or a positive outcome created a different pattern of results than those seen in Experiment 2A. Whereas previously an agent with greater control was slightly more likely to be viewed as intending a negative event, here we found the opposite: An agent with more control was viewed as having greater intent for the good outcome than the agent with less control (MC: 2.70 [0.09]; LC: 3.63 [0.08]; $t(373) = 7.83$, $p < 0.001$, 95% CI [0.70 1.17], $d = 0.81$). Further, we found that the mean for both groups was significantly different from the midpoint (MC: $t(186) = 25.11$, $p < 0.001$, 95% CI [2.52 2.88]; LC: $t(187) = 17.87$, $p < 0.001$, 95% CI [3.48 3.79]), indicating that agents with greater and lesser control were both viewed as more likely to have intended the good outcome than the bad outcome.

4.3. Discussion

In Experiment 2A, we found that agents with more control over an accidental harm were slightly more likely to be viewed as intending the bad outcome than agents with less control ($p = 0.15$). This is surprising given that both agents took the same

action and caused the same bad outcome. As predicted, Experiment 2B provides evidence that this finding was due to our use of a scale that only allowed participants to evaluate intent for the bad outcome. Using a scale that allowed participants to attribute both negative and positive intentions, we found that agents with more control were rated as having better intent than agents with less control. Moreover, we found that agents with both more and less control were rated as having greater intent for the good outcome to occur, suggesting that the scale used in Experiment 2A was indeed problematically constrained.

Importantly, the issue of a restricted range does not apply to judgments of causality. An agent can either be completely causally responsible for a harm, not at all causally responsible, or somewhere in between. She cannot, however, be causally responsible for a harm that did not occur.

Taken together, Experiments 2A and 2B strongly suggest that changes in an agent's degree of control over an accidental harm lead to larger attributions of causal responsibility for the harm but lesser attributions of intending to harm. Given past research indicating a role for causal attribution in moral judgment, this implies that one route by which attributions of control influences moral judgment is via their effect on attributions of causal responsibility. In Experiment 4 we provide a further and more stringent test of this hypothesis.

5. Experiment 3

Some past research links counterfactual representation to moral judgment in a manner that contrasts with our own model. These studies demonstrate that we hold a person more causally and morally responsible for their harmful action when things could have gone differently, had they acted differently (Williams et al., 1996). This property is called "outcome mutability". In the introduction we presented an example: A taxi driver is held more responsible for harm to a couple he strands in the case that they would have been unharmed in his cab, compared with a case where they would have been equally harmed in his cab. There is an intuitive sense in which an agent has more control over the *outcome* in cases where their behavior renders that outcome mutable. Note, however, that the agent does not seem to have any more or less control over their *behavior* in such cases. In other words, a taxi driver is equally in control of their choice to pick up the passengers, whether or not this leads to subsequent injury. Our present interest is in this latter form of control: The volition control that a person has over her own behavior.

Nevertheless, one explanation for the influence of control on moral judgment found in Experiments 1 and 2 is that volitional control is confounded with outcome mutability. That is, the participant is confronted with a situation in which an agent is involved in harming another. In one case, the agent had an alternative course of action (More Control condition) that might have resulted in a better outcome. Thus, participants may represent an upward counterfactual associated with the alternative course of action. The presence of outcome mutability causes the agent to be rated as having more control over the situation, as playing a greater causal role in the situation and as more deserving of punishment than when the agent had no alternative course of action (Less Control condition). This explanation hinges on the degree to which the actually obtained outcome is mutable, and not whether the agent had any control their behavior: We view the agent as more worthy of punishment not because the agent had control over their behavior, but because alternative courses of action were available, and thus a different outcome could have obtained.

These influences can be distinguished when an agent has a choice of multiple options (high volitional control) that would all yield identical outcomes (low outcome mutability). In Experiment 3, we test this possibility using vignettes from Experiments 1 and 2 modified only so that the agent's unchosen option would have lead to the same outcome as their realized choice—i.e., a failed outcome that causes harm. Thus, for instance, a doctor chooses a medication likely to succeed but it fails; yet, her only alternative option was a placebo that never could have succeeded. In this case an agent has some volitional control in the sense that multiple options are available. Yet, although the agent's plan failed, no upward counterfactual is available—the only available alternative would have yielded the same outcome. As in our prior experiments, we manipulate whether or not this alternative behavior is available to the agent (i.e., in some cases the placebo is not in stock). Thus, we manipulate the degree of control that the agent has over their behavior, while preserving the (non-)mutability of the outcome across cases.

5.1. Methods

Participants ($N = 1636$) were recruited under the same parameters as Experiments 1 and 2. Vignettes for this study were modified from 3 out of the 4 vignettes used in Experiment 2. Now, rather than a choice between a good versus bad action, each agent chose between a good versus "placebo" action – an action that would have resulted in the same outcome as obtained when the agent chose the good action. To illustrate, we present the modified version of the vignette used as an example when describing Experiment 1, with differences italicized (but not in the experimental version):

> A doctor working in a hospital has a patient who is having hearing problems. This patient has two, and only two, treatment options. With Option A, there is a 66% chance the patient makes a full recovery. *Option B is a placebo pill and so there is no chance that the patient makes a full recovery, but it will make the patient feel as though they are being treated.* Whichever option the doctor chooses, it happens that if the patient fails to make a full recovery then the doctor will be able to publish a very prestigious paper, which would be great for his career. On the other hand, if the patient makes a full recovery, the doctor won't be able to publish this paper at all, but his reputation will improve a little bit.

In the MC condition, the next sentence read: "The doctor chooses to pursue Option A". In the LC condition, the sentence read: "While deciding, the doctor checks the medicine cabinet where both treatments are kept and discovers that the hospital is all out of the placebo pill – Option B – and thus the doctor is forced to proceed with Option A." Thus, the agent in the MC condition continues to possess a choice between a good action, likely to help another person, and a bad action, here certain to harm them.

Each participant read 3 vignettes, all in the critical moral luck condition. Vignette order was randomized across participants. Each participant was randomly assigned to have all 3 vignettes presented in the MC or the LC condition. Participants were randomly assigned to one of three rating types: Causal role ($N = 427$, same as Experiment 2A), degree of control ($N = 415$, same as Experiment 2A) or punishment ($N = 794$, same as Experiments 1 and 2A). All participants completed attention check questions after the last of the 3 scenarios, for use in screening. Based on our exclusion criteria, data from 147 participants (9.0%) was discarded. Participants finished by completing a series of demographic questions and being debriefed.

5.2. Results

Using vignettes that reduce the salience of a counterfactual outcome obtaining, we replicate all three of our prior effects. We find that, even in these modified vignettes, agents with more control

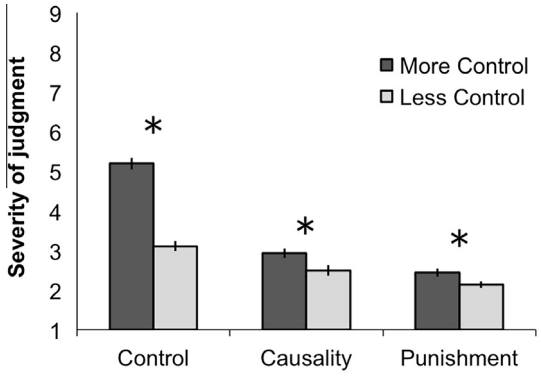

Fig. 3. Results from Experiment 3. Participants read scenarios of accidental harm, modified so that the obtained outcome would have also been reached from the agent's alternative option, all either in the More Control or Less Control condition. Plotted are average responses for the extent to which the agent had control, the extent to which the agent caused the outcome to occur and the amount of punishment the agent deserves. Error bars are SEM.

are rated as having greater control (MC: 5.19 (0.15); LC: 3.11 (0.13); $t(374) = 10.73$, $p < 0.001$, 95% CI [1.70, 2.46], $d = 1.11$, Fig. 3). Critically, they are also viewed as having a greater causal role in the obtained outcome (MC: 2.93 (0.12); LC: 2.50 (0.14); $t(387) = 2.38$, $p < 0.05$, 95% CI [0.07 0.80], $d = 0.24$). And, greater control leads to greater punishment (MC: 2.43 (0.10); LC: 2.13 (0.09); $t(722) = 2.32$, $p < 0.05$, 95% CI [0.05 0.56], $d = 0.17$).

5.3. Discussion

Experiment 3 tests whether outcome mutability fully accounts for the influence of our "control" manipulation in Experiments 1 and 2. We modified vignettes such that the agent's alternative choice could not have lead to a better outcome, and yet we continue to find that agents with more control are assigned greater punishment and causal responsibility. This suggests that outcome mutability cannot fully account for the results from Experiments 1 and 2.

6. Experiment 4

Thus far, we have demonstrated a link between the degree of behavioral control an agent possesses and attributions of both intent for a harm to occur and causal role in a harm coming about. We have focused on these relationships in the context on judgments of punishment, which are sensitive to both intentions and causal responsibility for outcomes (Berg-Cross, 1975; Cushman et al., 2009; Cushman, 2008; Gino et al., 2010; Martin & Cushman, 2015; Mazzocco et al., 2004) and could therefore shed light on our hypothesized link between control and causality.

In contrast, other categories of moral judgment show substantially less sensitivity to manipulations of causal responsibility, and instead rely especially heavily on attributions of malicious intent (or intentional action; Hebble, 1971; Imamoglu, 1975; Piaget, 1965; Wellman et al., 1986; Young et al., 2007). For instance, recent work of ours has demonstrated a striking divergence between decisions of how much to punish a partner for defecting in a economic interaction and decisions of whether or not that partner is likely to be cooperative on another round of play (Martin & Cushman, 2015). Whereas punishment decisions showed the expected sensitivity to both a partner's intentions and the outcome that they brought about, decisions of whether or not to interact with that partner again were sensitive almost exclusively to that partner's intentions and not the outcome that they caused. In other words, when deciding how much to punish, we care both about a person's intent and what they caused, whereas when assessing what kind of person they are – judging their character – we care much more about their underlying intentions.

This divergence in the sensitivity of character and punishment judgment allows a unique and precise test of whether control influences moral judgment in part via the attribution of causal responsibility. If, as we expect, character judgments are found to be relatively insensitive to attributions of causal responsibility, any influence of control on moral judgment due to the attribution of causal responsibility should be severely diminished for judgments of character. Thus, for instance, when people evaluate an agent's moral character in cases where she causes accidental harm, they should be insensitive to the amount of control she exerts over her behavior. Indeed, if anything, they should rate her to have a *superior* moral character in cases where she has control and uses it to choose the best available action. In line with the results of Experiment 2, she has demonstrated more helpful intentions, and judgments of character may be especially sensitive to information about intentions.

We first validate our use of character as a contrast judgment in a pilot study, by having participants judge each case of the action × outcome design, all in the MC condition. Participants assess either character or punishment and we examine the extent to which these judgments are influenced more by actions taken or outcomes caused. Then, we proceed to the main experiment, where we contrast judgments of character with punishment in our critical moral luck case, in each case contrasting judgments across the MC and LC conditions.

6.1. Pilot study – methods

Participants ($N = 100$) read 4 vignettes, each presented in one of the 4 possible action (positive, negative) × outcome (positive, negative) conditions. All vignettes were presented in the MC condition. Participants were randomly assigned to judge either the protagonist's personal moral character ($n = 49$, "How would you rate [protagonist's] personal moral character?", 1 = "Great moral character", 9 = "Horrible moral character") or deserved punishment ($n = 51$, same as Experiment 1). Vignette order was counterbalanced across participants. Participants completed attention check and demographic questions after the last scenario, and were then debriefed. Based on our exclusionary criteria, data from 5 (5.0%) participants was discarded.

6.2. Pilot study – results

We ran a 2 × 2 × 2 mixed ANOVA, with action (negative vs. positive action) and outcome (negative vs. positive outcome) as within-subjects factors and response type (character vs. punishment) as a between-subjects factor. Here, we found a significant 3-way interaction, $F(1,93) = 5.00$, $p < 0.05$, indicating that the impact of actions taken and outcomes caused did differ across judgments. To unpack this interaction, we separately conducted 2 (negative vs. positive action) × 2 (negative vs. positive outcome) repeated-measures ANOVAs on subjects' ratings of characters and punishment. As expected, when making character judgments, demonstrated intentions exerted greater influence than outcomes (Intentions: $F[1,43] = 70.57$, $p < 0.001$, $\eta^2 = 0.32$; Outcomes: $F[1,43] = 44.37$, $p < 0.001$, $\eta^2 = 0.11$), while the reverse was true when assessing deserved punishment (Intentions: $F[1,50] = 42.08$, $p < 0.001$, $\eta^2 = 0.13$; Outcomes: $F[1,50] = 78.69$, $p < 0.001$, $\eta^2 = 0.26$).[3] Thus, while both judgments are sensitive to both factors,

[3] We report eta squared because we are interested in the amount of variability explained by both intent and outcome relative to the total variability between our two response conditions. Such a comparison is reasonable here because total variability was very similar between conditions ($SS_{total\ (punishment)} = 1490.29$, $SS_{total\ (character)} = 1413.48$.

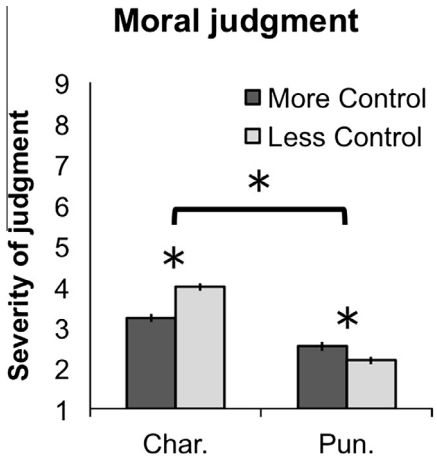

Fig. 4. Results from Experiment 4. Participants read scenarios of accidental harm, all either in the More Control or Less Control condition. In addition, they were assigned to make a judgment of how much punishment is deserved or a judgment of the agent's moral character. Plotted are average responses for the MC and LC conditions for participants rating punishment or character. Error bars are SEM.

judgments of character rely relatively more on information about intentions and determinations of punishment rely more on information about outcomes caused.

6.3. Methods

Participants ($N = 857$) read 4 vignettes, each presented in the critical moral luck condition, with each participant randomly assigned to read all MC or all LC vignettes. Participants were randomly assigned to judge either the protagonist's personal moral character ($n = 426$, Same as pilot study) or deserved punishment ($n = 431$, same as Experiment 1) for all 4 vignettes. Vignette order was counterbalanced across participants. Participants completed attention check and demographic questions after the last scenario, and were then debriefed. Based on our exclusionary criteria, data from 45 (5.3%) participants was discarded.

6.4. Results

Using our now validated judgments, we first replicated the impact of control on punishment, finding for a fourth time that greater control leads to greater punishment (MC: 2.53 (0.10); LC: 2.18 (0.10); $t(416) = 2.36$, $p < 0.05$, 95% CI [0.06, 0.64], $d = 0.23$, Fig. 4). Critically, greater control had exactly the opposite effect on judgments of character: Greater control lead to perceptions of *better* moral character (MC: 3.23 (0.10); LC: 3.99 (0.10); $t(392) = 5.481$, $p < 0.001$, 95% CI [0.49, 1.03], $d = 0.55$). To confirm this interaction, we used linear mixed-effects regression, including fixed-effects predictors for control condition, response type and their interaction, as well as a random-effects predictor for vignette. As expected, the interaction between control and judgment type was significant ($F[1,808] = 29.802$, $p < 0.001$, $\eta_p^2 = 0.036$).

6.5. Discussion

Experiment 4 provided a strong test of the hypothesis that control can influence moral judgment through representations of causality in addition to its influence on inferences of intentionality. We found that for judgments sensitive to outcomes caused (e.g. punishment), greater control over an accidental outcome leads to greater punishment, presumably by virtue of the greater perceived causal role the agent plays in the bad outcome that occurs. This effect was reversed for judgments that are less sensitive to outcomes caused (e.g. character): Greater control leads to judgments of better moral character, likely as a result of better perceived intent. These findings provide especially strong evidence that control sometimes influences moral judgment by modifying attributions of causal responsibility. Thus, we can explain why an agent with greater control is viewed as more worthy of punishment but also of better moral character.

Our experiment was limited to judgments of punishment and character; an important direction for future research is to establish whether comparable results can be obtained for judgments of blame (which, like punishment, depends substantially on attributions of causal responsibility) and for judgments of moral wrongness (which, like character, show a much greater reliance on attributions of intent or intentional action).

These results accord well with the findings in Experiment 2B, in which participants were more likely to view the agent with greater control as intending the good outcome compared to the agent with less control. In Experiment 4, participants indicated that the agent with greater control had a superior moral character. Together, these results strongly suggest that the impact of intentions found in Experiment 2A is a byproduct of the scale employed in that experiment.

7. General discussion

Why do we forgive a person for behavior beyond their control? Is it only because we view their behavior as unintentional, or could it also be because we view them as less causally responsible for the harm? To test this, we experimentally dissociated an agent's intent to cause harm, their causal role in harm coming about, and the degree of control they had over their behavior. Validating past models (Alicke, 2000; Malle et al., 2014; Weiner, 1995), we show that control can influence perceptions of intent. And, this model appears to be sufficient to explain the influence of control on judgments of moral character. But, when control and intent are effectively dissociated, control has a unique impact on punishment, and our evidence suggests that it depends on representations of causal responsibility.

These two routes by which control can impact moral judgment – through intent and causal role – are mirrored in prior work dissociating the contributions of intent to harm and causal responsibility for harm in moral judgment (Buon, Jacob, Loissel, & Dupoux, 2013; Cushman, Sheketoff, Wharton, & Carey, 2013; Cushman, 2008; Phillips & Shaw, 2014). One appealing model of the cognitive architecture of moral judgment is a single process that incorporates information about causation and intention in a sequential manner: Did she cause harm? If so, did she intend the harm? However, the influence of these factors is best explained by a two-process model in which each is supported by a unique cognitive process that competitively interacts with the other (Cushman, 2008). Further support for this architecture comes from the unique developmental trajectory of these two processes: In young children, the influence of intentions is first seen in judgments supported by the intent-based process (e.g. judgments of wrongness), and only later influences judgments supported by the causation-based process (Cushman et al., 2013), consistent with the idea that intentions and causal responsibility are processed separately. And, these processes rely on effortful cognition to different extents (Buon et al., 2013). Under cognitive load, judgments reliant on differences in agents' intentions (e.g. distinguishing between an intentional and accident harm) are impaired, whereas judgments that rely on differences in causal role (e.g. distinguishing between an intentional and attempted harm) are preserved. Finally, each process appears to influence some moral

judgments more than others – compared with judgments of wrongness, permissibility or character, punishment and blame judgments depend relatively more on an agent's causal connection to harm (Cushman, 2008) Our results support this dissociation in two ways. First, we demonstrate that control can influence moral judgment independently through perceptions of intent and perceptions of causal role. Second, we confirm that punishment judgments depend to a greater extent on causal responsibility for harm than judgments of character.

Beyond their implications for our understanding of behavioral control as a criterion for moral judgment, our results speak to a larger debate over the mechanisms underlying moral luck and its influence on punishment. Previous psychological research has characterized moral luck in terms of outcome bias (Alicke & Davis, 1989; Carlsmith, Darley, & Robinson, 2002; Darley et al., 2000) or hindsight bias (Baron & Hershey, 1988; Tostain & Lebreuilly, 2008; Young et al., 2010). According to the outcome bias model, the mere presence of a bad outcome generates negative affect in the perceiver, which subsequently biases moral judgment. According to the hindsight bias model, a bad outcome causes us to reassess whether the agent acted reasonably in the first place. In either case, moral luck is a general error or bias operating within a system designed to make moral judgments on the basis of intent alone.

The results of Experiment 4 are not consistent with either of these accounts. We found that an agent with more control who causes an accidental harm receives more punishment than an agent with less control, but is viewed as having better moral character. Yet, outcome bias would predict a biasing effect on any moral judgment, including character. There is no *a priori* reason to expect a misattribution of negative affect to judgments of punishment alone. Meanwhile, hindsight bias predicts that an agent causally responsible for a bad outcome will be viewed as having a *more* culpable mental state, which should increase negative character attributions. However, we find just the opposite effect. This same argument allows us to rule out other potential explanations of our findings. For instance, it is possible that outcomes influence moral judgment because they are often the best indicator of an underlying, unobservable intention. However, this account also fails to explain why, in the case of accidental outcomes, an agent with more control is punished more but is rated as having better moral character. If the accidental outcome is a signal of a negative underlying intention, character ratings should track punishment ratings, which we do not find. Thus, these previous explanations of moral luck cannot account for our results.

Instead, we favor an adaptive account of moral luck (Martin & Cushman, in press). Punishing based on outcomes will, on the whole, lead those causing bad outcomes to change their behavior in beneficial ways. Critically, such punishment should be sensitive to whether the person has the capacity to change the behavior in question – If the behavior was uncontrollable (e.g. generated by an epileptic seizure), then punishment cannot cause that behavior to change, and punishment should be reduced. Our data is consistent with this: Despite viewing agents causing accidental harms with reduced control as having worse character (because their "choice" of the good action was forced), we punish them less.

Our juxtaposition of moral character and punishment judgments illustrates both the power and limits of the "person-centered" approach to moral judgment (Uhlmann & Zhu, 2013; Uhlmann, Pizarro, & Diermeier, 2014). Illustrating the power of this perspective, we find evidence that the moral assessment of character is well-tuned to the task of predicting future behavior. It is chiefly sensitive to intent, and credits individuals for controllable actions that are diagnostic of prosocial motives. It is limited, however, in its ability to explain the distinctive reliance of punishment on accidental outcomes, and the concomitant influence of the attribution of control in such cases. This reflects a more general division between two organizing principles of adaptive function in the moral domain: Partner choice (deciding whom to interact with, based presumably on assessments of moral character) and partner control (influencing the behavior of others via reward and punishment) (Baumard, André, & Sperber, 2013; Martin & Cushman, 2015).

A key finding of our studies is that control moderates the effect of moral luck specifically through via its influence on the attribution of causal responsibility. This accords with other work suggesting that judgments of causal responsibility are designed not only to reflect statistical or generative relations between events, but also to efficiently guide judgments about moral blame (Knobe & Fraser, 2008; Knobe, 2005, 2009; Roxborough & Cumby, 2009). In colloquial terms, if the human mind is characterized by the operation of intuitive theories, then these are the theories not only of a scientist, but also of an engineer and a lawyer (Knobe, 2010; Pinker, 1999; Tetlock, 2002).

Our work points towards several unresolved issues. First, to what extent does an intuitive versus deliberative mindset influence the role that control plays? Above, we suggest an adaptive account of moral luck, one based on an adaptive structuring of retributive motivations (Martin & Cushman, in press). To the extent that this account explains our results, we might expect the influence of control in cases of accidental harm to be driven more by heuristics than rational, deliberative thought, but this remains to be tested. Second, are our results specific to sanctions imposed because of negative outcomes, or might they generalize to reward of positive outcomes? Given that both punishment and reward have the potential to change behavior, we may find that reward is similarly sensitive to the degree of control an agent possess when they cause good outcomes: I am only willing to spend resources on encouraging behavior that is under an agent's control and is therefore able to be increased. Of course, differences in how people learn from positive versus negative feedback may lead reward and punishment to be sensitive to distinct factors, including control.

To conclude, we provide evidence that control influences moral judgment not only by influencing the attribution of intent, but also by independently influencing the attribution of causal responsibility. This underscores the fundamental importance of both intentional and causal attribution to moral judgment, and it provides further evidence that they make differential contributions to distinct categories of moral judgment (e.g., punishment versus character). In Whitman's case, we may be tempted to forgive his heinous actions not because he didn't intend to cause harm, but rather because it was never really *he* who caused the harm at all.

Acknowledgements

We wish to thank Jorie Koster-Hale, Ryan Miller, Jonathan Phillips and Steven Sloman for helpful comments on earlier versions of this manuscript and Michael Frank, Bertram Malle and the Moral Psychology Research lab for valuable feedback on this work. This material is based upon work supported by National Science Foundation Award No. 1228380 to FC and by National Science Foundation Graduate Research Fellowship Grant No. DGE1144152 to JWM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.11.008.

References

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63* (3), 368–378. http://dx.doi.org/10.1037/0022-3514.63.3.368.

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574. http://dx.doi.org/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. http://dx.doi.org/10.1016/0022-1031(89)90028-0.

Baron, J., & Hershey, J. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*(4), 569–579.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(1173–1182), 1173–1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *The Behavioral and Brain Sciences, 36* (1), 59–78. http://dx.doi.org/10.1017/S0140525X11002202.

Berg-Cross, L. (1975). Intentionality, degree of damage, and moral judgments. *Child Development, 46*(4), 970–974.

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

Byrne, R. M. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences, 6*(10), 426–431. http://dx.doi.org/10.1016/S1364-6613(02)01974-5.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology, 67*(1). http://dx.doi.org/10.1146/annurev-psych-122414-033249.

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. http://dx.doi.org/10.1037/0022-3514.83.2.284.

Christensen, R. (2015). Ordinal – Regression models for ordinal data.

Cushman, F. A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. http://dx.doi.org/10.1016/j.cognition.2008.03.006.

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

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

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

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. http://dx.doi.org/10.1016/j.obhdp.2009.11.001.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52*(5), 449–466. http://dx.doi.org/10.1080/00201740903302600.

Hebble, P. W. (1971). The development of elementary school children's judgment of intent. *Child Development, 42*(4), 1203–1215.

Imamoglu, E. O. (1975). Children's awareness and usage of intention cues. *Child Development, 46*(1).

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.

Knobe, J. (2005). Cognitive processes shaped by the impulse to blame. *Brooklyn Law Review*.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130*(2), 203–231. http://dx.doi.org/10.1007/s11098-004-4510-0.

Knobe, J. (2009). Folk judgments of causation. *Studies in History and Philosophy of Science Part A*.

Knobe, J. (2010). Person as scientist, person as moralist. *The Behavioral and Brain Sciences, 33*(4), 315–329. http://dx.doi.org/10.1017/S0140525X10000907. Discussion 329–365.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: two experiments. *Moral Psychology*.

Knobe, J., & Nichols, S. (2011). Free Will and the Bounds of the Self, (1984), pp. 1–24.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*(4), 303–332. http://dx.doi.org/10.1016/j.cogpsych.2010.05.002.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(May 2014), 147–186. http://dx.doi.org/10.1080/1047840X.2014.877340.

Martin, J. W., & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PLOS ONE, 10* (4), e0125193. http://dx.doi.org/10.1371/journal.pone.0125193.

Martin, J. W., & Cushman, F. A. (in press). The adaptive logic of moral luck. In J. Sytsma & W. Buckwalter (Eds.), *The Blackwell companion to experimental philosophy*, Wiley-Blackwell.

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. http://dx.doi.org/10.1080/01973533.2004.9646401.

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

Phillips, J., & Shaw, A. (2014). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science, 203*, 1–48. http://dx.doi.org/10.1111/cogs.12194.

Piaget, J. (1965). *The moral judgment of the child*. New York: Free Press (Psychoanalytic Review).

Pinker, S. (1999). How the mind works. *Annals of the New York Academy of Sciences*.

Preacher, K., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. http://dx.doi.org/10.3758/BRM.40.3.879.

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

Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin, 121*(1), 133–148. http://dx.doi.org/10.1037/0033-2909.121.1.133.

Roxborough, C., & Cumby, J. (2009). Folk psychological concepts: Causation 1. *Philosophical Psychology, 22*(2), 205–213. http://dx.doi.org/10.1080/09515080902802769.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review, 109*(3), 451–471. http://dx.doi.org/10.1037//0033-295X.109.3.451.

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

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2014). A person-centered approach to moral judgment. *Perspectives on Psychological Science*.

Uhlmann, E. L., & Zhu, L. [Lei] (2013). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science, 5*(3), 279–285. http://dx.doi.org/10.1177/1948550613497238.

Walsh, C. R., & Sloman, S. a. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language, 26*(1), 21–52. http://dx.doi.org/10.1111/j.1468-0017.2010.01409.x.

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

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).

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*(2), 161–169. http://dx.doi.org/10.1037/0022-3514.56.2.161.

Williams, B. A. O. (1981). *Moral luck: Philosophical papers, 1973–1980*. Cambridge University Press.

Williams, C., Lees-haley, P. R., & Price, J. R. (1996). The role of counterfactual thinking and causal attribution in accident-related judgments. *Journal of Applied Social Psychology, 26*(23), 2100–2112. http://dx.doi.org/10.1111/j.1559-1816.1996.tb01789.x.

Young, L., Cushman, F. A., 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. http://dx.doi.org/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. http://dx.doi.org/10.1007/s13164-010-0027-y.