


# Counterfactuals, Control, and Causation: Why Knowledgeable People Get Blamed More

Personality and Social  
Psychology Bulletin  
2015, Vol. 41(5) 643–658  
© 2015 by the Society for Personality  
and Social Psychology, Inc  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/0146167215572137  
pspb.sagepub.com  


Elizabeth A. Gilbert<sup>1</sup>, Elizabeth R. Tenney<sup>2</sup>,  
Christopher R. Holland<sup>1</sup>, and Barbara A. Spellman<sup>1</sup>

## Abstract

Legal and prescriptive theories of blame generally propose that judgments about an actor's mental state (e.g., her knowledge or intent) should remain separate from judgments about whether the actor caused an outcome. Three experiments, however, show that, even in the absence of intent or immorality, actors who have knowledge relevant to a potential outcome will be rated more causal of that outcome than their ignorant counterparts, even when their actions were identical. Additional analysis revealed that this effect was mediated by counterfactual thinking—that is, by imagining ways the outcome could have been prevented. Specifically, when actors had knowledge, participants generated more counterfactuals about ways the outcome could have been different that the actor could control, which in turn increased causal assignment to the actor. These results are consistent with the Crediting Causality Model, but conflict with some legal and moral theories of blame.

## Keywords

attribution, counterfactual, causation, knowledge, mental state

Received February 28, 2014; revision accepted January 18, 2015

Imagine that you just killed someone. What are you guilty of—first degree murder, manslaughter, reckless homicide, or nothing at all? Under U.S. law, the answer depends at least in part on your mental state at the time of the killing (e.g., Brown, 2012; Raz, 2010). You may be guilty of first degree murder if you killed intentionally; you may be guilty of nothing at all if you killed out of reasonable fear for your life. And, somewhere in the middle, you may be guilty of manslaughter or reckless homicide if you did not intend to kill but you knew or should have known that your actions were likely to lead to the fatal outcome. These graded levels of guilt are consistent with most prescriptive theories of blame (e.g., Heider, 1958; Shaver, 1985). And they are codified in criminal and tort laws requiring that judges and juries consider at least two separate factors when assessing blame for a bad outcome: (a) whether the actor in question *caused* the bad outcome (i.e., *actus reus*) and (b) whether the actor had a requisite blameworthy mental state (i.e., *mens rea*, e.g., intent, knowledge).

In this article, we argue that, in contrast with moral and legal theories, people do not keep these two factors separate. Instead, consistent with models of causal reasoning, we believe that an actor's mental state—here, her knowledge—may affect her blameworthiness because knowledge, itself, makes the actor seem more causal. We further believe that

knowledge increases causal attribution through counterfactual thinking—that is, by the way people imagine how the outcome could have turned out differently. In the example above, for instance, if you knew your actions were likely to cause the death, it might be easier for others to imagine how you could have avoided the killing, which would in turn increase the causation attributed to you.

## Causation and Counterfactuals

### *Why Mental State Judgments Might Affect Causation Judgments*

We are not the first to propose that mental states may affect causation judgments. Many studies have already shown that actors are judged to be more causal and blameworthy for bad outcomes when they have bad intentions (e.g., Alicke, 1992; Knobe, 2005) or could have foreseen that their actions could

<sup>1</sup>University of Virginia, Charlottesville, USA

<sup>2</sup>University of Utah, Salt Lake City, USA

#### Corresponding Author:

Elizabeth A. Gilbert, University of Virginia, 102 Gilmer Hall, P.O. BOX 400400, Charlottesville, VA 22904, USA.

Email: eag4yx@virginia.edu

lead to the bad outcomes (Lagnado & Channon, 2008; Robinson & Darley, 1995, Study 8). Intent and foreseeability may be rationally related to causation judgments, insofar as they provide evidence that the actor was behaving in a way that would cause the outcome (see Heider, 1958; Kelley, 1973). Prior research, however, suggests that the actor's mental state can influence others' judgments of how causal she is even when her causal behavior is held constant. Most previous theorists thus propose a top-down, morality-based account for these effects. In particular, they suggest that people use mental states to evaluate an actor's morality, and in turn, these morality judgments influence causation and blame judgments (Alicke, 2000; Knobe, 2010; Pizarro & Tannenbaum, 2012).

We, however, propose an alternative multi-step bottom-up process. Specifically, we suggest that an actor's mental state influences whether and what counterfactuals observers consider, which affects assessments about how much the actor increased the likelihood of the outcome, which in turn affects judgments about how causal (and blameworthy) the actor was. This process is consistent with the view that people are reasoning not like moralists but rather like lay scientists when assessing causation.

### The Crediting Causality Model

Most causal models proposing that people reason like lay scientists have been limited to assessing general causation for repeated events—such as whether smoking causes cancer or whether fertilizer causes plants to grow. These mathematical models try to capture the process of “causal learning” (i.e., how people decipher causal relationships) from multiple recurring events (see Holyoak & Cheng, 2011, for a review). In addition to looking at multiple repetitions of events, the events in question tend to deal with physical, chemical, or biological causation. These models generally claim that a cause is something that raises the probability of an effect above some baseline probability of occurrence when the cause is absent.

A largely separate line of research deals with how people attribute causation for a particular outcome—such as whether smoking caused *this* person's cancer or who caused *this* particular car accident. This “causal attribution” research, which is usually more applicable to the legal world than the research on causation for repeated events, tends to look at events that involve human agents as potential causes of one-time events. And it generally applies non-mathematical theories of attribution. Instead of presuming that people act like lay scientists and apply mathematical models to assess causation,<sup>1</sup> this research has often claimed that people use heuristics to make attributions, including studying whether voluntary actions are more causal than environmental causes (Lagnado & Channon, 2008; McClure, Hilton, & Sutton, 2007), whether first or last events in a causal chain are most causal (see Spellman, 1997, for an older review), the morality of the

actions (Alicke, 1992; Knobe, 2005), and how the similarity of cause and effect (e.g., “correspondence bias”) influences causal judgments (Jones & Davis, 1965).

However, one theory of causal reasoning about particular events proposes that, when computing causation for a single event outcome, people use a mathematical process similar to that for repeated events in scientific causation. The Crediting Causality Model (Spellman, 1997; also called Spellman's probability-updating account, or SPA; see Mandel, 2003) states that

- i. Causes (c) are evaluated as to how much they change the probability (P) of the event outcome (e) based on what has previously occurred. Causal strength is thus a function of the difference between the probability of the outcome after the target cause has occurred minus the probability of the outcome before the target cause has occurred, or  $P(e|c) - P(e|\sim c)$ .<sup>2</sup>
- ii. The person or event that changes the outcome<sup>3</sup> most is deemed “the cause” when a single selection is necessary.

The Crediting Causality Model thus proposes that people assess causality for a specific case (e.g., how much smoking caused a particular person's cancer) using the same types of estimation as is used in scientific, repeated-events causal assessment.

### The Role of Counterfactuals

Of course, there is a problem with applying a science-type general causation model to particular outcomes: How can the probability of an outcome for events that happen only once be computed? Where do the probabilities for *before* and *after* come from? According to the Crediting Causality Model (as extended in Spellman & Kincannon, 2001, and Spellman, Kincannon, & Stose, 2005), one source of probabilities is from the pre-existing knowledge of the person making the judgment (e.g., knowledge that car accidents are more common during storms). A second source of probabilities is counterfactual thinking—that is, imagining ways that the outcome could have unfolded differently “if only” some preceding event had been different. For instance, imagine that someone shoots a victim dead after the victim pulled out a lead pipe and scared the shooter. You might think, *if only* the shooter had run away or called the police, then the victim would still be alive. These thoughts would likely increase your judgment of how causal the shooter is. Alternatively you might think, *if only* the victim had not scared the shooter, he would still be alive. This counterfactual would likely increase your judgment of how causal the victim is (see Robinson & Darley, 1995, Study 5; also Branscombe, Owen, Garstka, & Coleman, 1996).

Indeed, such counterfactual thinking is common. People spontaneously generate counterfactuals in response to bad

events, particularly when the events are the result of exceptional or controllable actions (McEleney & Byrne, 2006; Roese, 1997).

### *Importance of Counterfactual Type: Changing the Outcome, Potency, and Control*

**Changing the outcome.** Many studies have established that counterfactual thinking affects causation judgments (Branscombe et al., 1996; Wells & Gavanski, 1989). Specifically, even when two actors complete the exact same behavior, one may be judged as more causal when it is easy to imagine how a change in his or her behavior would have led to a different outcome (as in the shooter example above).

However, not all counterfactuals affect judgments of causation in the same way. Most counterfactual research focuses on “if-only” counterfactuals that “undo” an outcome—that is, counterfactuals in which one imagines both an earlier event (“antecedent”) being different and, as a result, a later event (“consequent”) being different. In the example of the shooter and victim, you may imagine that changing either actor’s behavior could have prevented the death.

However, it is also possible to imagine that a change in some antecedent would *not* change the outcome. For instance, you could imagine that even if the victim had not scared the shooter, the shooter would have killed him anyway (if perhaps the shooter were delusional or trigger-happy). Such “even if” thinking—sometimes called “semifactual” thinking—may reduce causal attributions to the victim (e.g., Branscombe et al., 1996; Byrne, 2007; McCloy & Byrne, 2002).

**Potency.** In the real world, of course, often, it is not known with certainty whether changing an antecedent would change a consequent outcome. Moreover, even if an imagined counterfactual would very likely change an outcome if it occurred, the counterfactual might be unrealistic or improbable. The construct of “potency” takes into account both of these concerns.

Counterfactual potency was proposed by Petrocelli, Percy, Sherman, and Tormala (2011) as a measure of the effectiveness of counterfactuals. Potency depends on two characteristics of the counterfactual. The first characteristic is the “if likelihood”—that is, how likely it is that the counterfactual could have actually happened. It seems likely that someone could call the police, but it seems unlikely that the laws of gravity would be suspended. The second characteristic is the “then likelihood”—that is, if the counterfactual actually did happen, how likely it is that the outcome in question would have been different. Note that the Then likelihood is thus a conditional likelihood. Potency is the product of the If likelihood multiplied by the Then likelihood. This measure of “how much people actually *believe* in the counterfactuals that they generate” (Petrocelli et al., 2011, p.43) predicts the

amount of causation that people assign to the subject of a given counterfactual. Thus, for instance, imagining that a Martian had come down and saved the shooting victim will likely not influence someone to judge the death as being caused by lack of benevolent Martian. However, imagining something reasonable and effective, say that the shooter had retreated to safety rather than shooting or that the victim had not scared the shooter, may increase someone’s judgment of how causal the shooter or victim was, respectively (see Robinson & Darley, 1995, Study 5 using a dependent measure of prison sentences). The potency construct is consistent with the Crediting Causality Model, in that they rely on the notion that the causal strength depends on how much a potential cause actually changes the likelihood of the outcome (see Appendix A.)

**Control.** When considering human actors, how much control an actor has over his actions and their effects may also be relevant to how much a counterfactual influences causal attribution. Prior models propose that an actor’s control over an outcome is relevant to assigning blame (Alicke, 2000; Guglielmo, Monroe, & Malle, 2009; Weiner, 1995) and causation (McGill & Tenbrunsel, 2000). And people are more likely to generate counterfactuals about controllable antecedents than uncontrollable ones (Markman, Gavanski, Sherman, & McMullen, 1995; McEleney & Byrne, 2006). In one study, participants read a story about a man who was delayed by four events on his way home, which led to him being too late to save his wife, who had a heart attack while he was out (Giroto, Legrenzi, & Rizzo, 1991). Participants were more likely to generate counterfactuals about the man choosing to stop by a pub for a beer than about, for instance, the man having to wait for a tree blocking the road or a slow-moving truck. Similarly, when actors make a choice that affects the outcome (thus exerting control), they may be judged as more causal or blameworthy than when they do not have reasonable alternatives to choose from (e.g., Goldinger, Kleider, Azuma, & Beike, 2003). For instance, participants judged a man to be more causal of a break-in to his car when they believed he could have parked his car in a safer area than when there were no safer spots available (McGill & Tenbrunsel, 2000).

That generating controllable counterfactuals increases causal attribution more than uncontrollable counterfactuals is consistent with a functional theory of counterfactual thinking, which states that focusing on counterfactuals is useful for learning how to avoid making the same mistakes in the future (e.g., Epstude & Roese, 2008). Imagining how someone could have acted differently to avoid a negative outcome (e.g., a bad grade) in ways someone could have controlled (e.g., if only I had studied) rather than in ways someone could not have controlled (e.g., if only the teacher had been lenient) may be especially useful for learning to avoid such outcomes in the future (McEleney & Byrne, 2006). The

degree of control in a counterfactual is thus likely an important feature of the counterfactual.

## Present Studies

In this article, we aim to (a) demonstrate one way that human reasoning differs from legal theories of blame (Experiments 1-3); (b) provide support for a bottom-up (more scientific than heuristic) process of causal attribution consistent with the Crediting Causality Model (Experiments 2 and 3); and (c) show the importance of “counterfactual control” when measuring the effectiveness of a counterfactual thought (Experiment 3).

Specifically, our experiments address whether an actor’s knowledge affects causation judgments, and if so why. In each experiment, participants read a vignette involving actors with or without knowledge related to a potential bad outcome. Participants judge the actors’ causation (and make other judgments that differ across studies). Experiment 1 establishes that actors who have knowledge of a risk are rated more causal than actors who do not have that knowledge. Experiment 2 replicates that finding and provides preliminary evidence that the mechanism is counterfactual thinking. It suggests that if an actor takes reasonable action to avoid the bad outcome given his knowledge (thus eliminating an obvious potent and controllable counterfactual), then people no longer judge the knowledgeable actor to be especially causal. Finally, Experiment 3, a replication using the same stimuli as Experiment 1, explicitly measures counterfactual thinking and shows that counterfactual thinking about things the actor could control mediates the relationship between knowledge and causation. Together, these data support the hypothesis that when actors have knowledge, people are more likely to think of counterfactuals that the actor could control, and such counterfactuals increase the causality attributed to the actors.

For these experiments, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures. The data files, materials, and supplementary information addressing statistical assumptions are available online at the Open Science Framework at <https://osf.io/n2ecf>

## Experiment 1: Establishing That Knowledge Increases Causal Attributions

Experiment 1 investigated whether an actor’s knowledge of a potential risk factor affects judgments of causation. Participants read a vignette about a woman who either did or did not have prior knowledge that her lawnmower might be likely to malfunction. In each story, while she mowed, the mower began to rumble and her tulips were destroyed. Participants evaluated the woman’s causality for the destruction of the tulips. The study also attempted to rule out any

influence of morality or intent, as the outcome negatively affected only the actor herself, and follow-up questions assessed perceptions of the actor’s intent. We predicted that when the woman had knowledge of the potential for the lawnmower to malfunction, she would be judged as more causal of the destruction of the tulips than when she did not have this knowledge. This finding would support the hypothesis that, inconsistent with theories of law and moral blameworthiness, people do not intuitively separate a person’s mental state from analysis of her causal action.

## Method

**Participants.** Participants were 202 people<sup>4</sup> (71 women; *Mdn* age = 27) who completed the experiment online via Amazon Mechanical Turk for US\$0.35. Access was limited to people in the United States with at least 95% approval rating on the website.

**Design, materials, and procedure.** Participants were randomly assigned to the Knowledge or No Knowledge condition.

**Scenario.** All participants read a scenario about a woman who destroyed her prize-winning tulips with a malfunctioning lawnmower. In only the Knowledge condition did her mechanic warn her that the mower was defective. The full scenario is below, with the additional Knowledge condition information in italics.

Julia is mowing her lawn using an old motorized lawnmower while listening to headphones. *Her lawnmower mechanic had told her weeks before to buy a new model because hers was prone to malfunction.* Halfway through her work, her lawnmower malfunctions and begins to rumble back and forth wildly. When she finishes, she looks back at her lawn and realizes that her prize-winning tulips have been completely destroyed.

**Main causal judgment.** Participants first responded to the question, “Who or what would you say was the main cause of the destruction of the tulips?” in an open, free-response text box. On the following page, they self-coded their free responses. They were reminded of what they had written in the text box and were asked to summarize it using the option that best captured what they wrote: *Julia*, *Lawnmower*, or *Other* (with a text box available to explain *Other*). The options were presented in random order with *Other* always last. They could select more than one option and were instructed to do so in the case of a tie.

**Other measures and demographics.** The extent to which participants thought that Julia and Lawnmower caused the outcome was measured on separate scales from 0 (*not at all*) to 7 (*very much*). Next, participants answered whether they thought Julia had wanted the tulips to be destroyed by selecting *yes*, *no*, or *don’t know*. As a comprehension check, participants answered one question asking what happened to

the tulips by selecting *they thrived*, *they died*, or *nothing happened to them*. Finally, participants reported their age and sex and had space to write anything they wanted to tell us about the survey or their answers.

## Results and Discussion

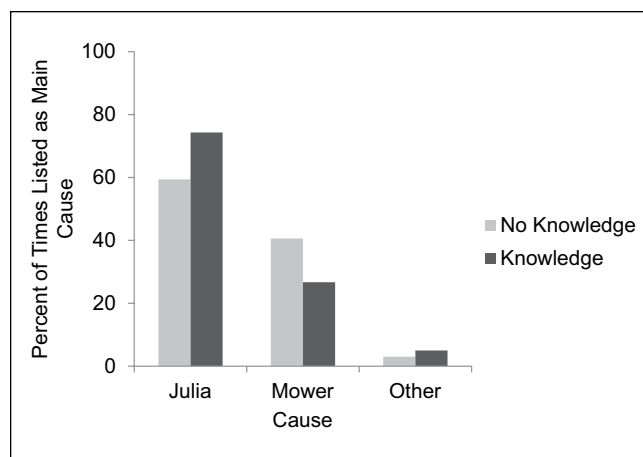
Attributions of cause differed across knowledge conditions in the predicted direction. Although almost all participants believed that Julia had not intended to destroy her tulips, they judged Julia as being more causal on average when she had knowledge that her lawnmower might malfunction than when she did not.

**Comprehension check.** Almost all participants (96.5%) correctly reported that the tulips died. Excluding participants who did not pass this comprehension check did not meaningfully affect the results, and their data were retained in analyses.

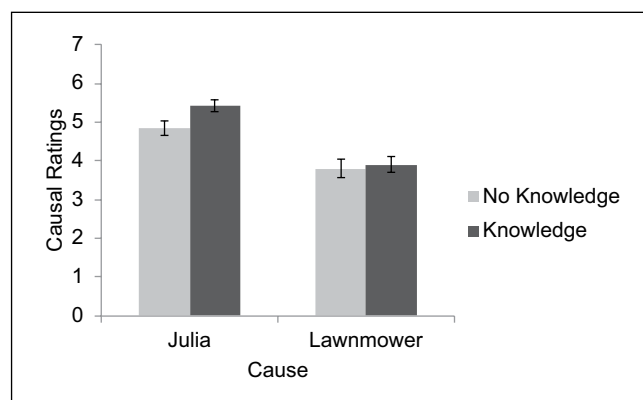
**Main cause.** Participants could select more than one option as the main cause of the destruction of the tulips, creating non-independence among the observations. Therefore, we calculated the proportion of the time that a participant selected Julia rather than something else as the main cause. Participants who selected just Julia ( $n_{\text{No Knowledge}} = 57$ ,  $n_{\text{Knowledge}} = 70$ ) scored 1. Those who selected Julia and another option ( $n_{\text{No Knowledge}} = 3$ ,  $n_{\text{Knowledge}} = 5$ ) scored 0.5. Those who did not select Julia at all ( $n_{\text{No Knowledge}} = 41$ ,  $n_{\text{Knowledge}} = 26$ ) scored 0. Overall, participants selected Julia as the main cause more frequently in the Knowledge condition than in the No Knowledge condition,  $t(200) = 2.12$ ,  $p = .035$ ,  $d = 0.30$ , 95% confidence interval (CI) = [0.02, 0.58]. Thus, participants were more likely to consider Julia the main cause when she had knowledge that the lawnmower was defective than when she did not, although her actions were exactly the same (see Figure 1).

**Causal ratings.** The causal ratings corroborated the pattern of results found with the free responses about who or what was the main cause (see Figure 2). A MANOVA with causal ratings of Julia and Lawnmower as the two outcome variables and knowledge (No Knowledge = 0, Knowledge = 1) as an independent variable revealed, as expected, that Knowledge condition affected causal ratings,  $F(1, 200) = 5.05$ ,  $p = .026$ ,  $d = 0.32$ , 95% CI = [0.04, 0.60]. Julia was rated more causal in the Knowledge condition ( $M = 5.4$ ,  $SD = 1.6$ ) than the No Knowledge condition ( $M = 4.8$ ,  $SD = 2.1$ ),  $t(200) = 2.26$ ,  $p = .025$ ,  $d = 0.32$ , 95% CI = [0.11, 0.67]. As might be expected, causal ratings of Lawnmower were similar across conditions ( $M = 3.8$ ,  $SD = 2.3$  and  $M = 3.9$ ,  $SD = 2.0$ , respectively),  $t(200) = 0.33$ ,  $p = .744$ ,  $d = 0.05$ , 95% CI = [-0.23, 0.32].

**Intent.** Almost all participants (98%) reported that Julia had not wanted to destroy the tulips. Thus, the results are a



**Figure 1.** Main cause: Percentage of responses attributing cause to Julia, the lawnmower, and other causes, across conditions.



**Figure 2.** Causal ratings (0-7) for Julia and the lawnmower across conditions (with standard errors).

reflection of what was known about Julia's knowledge, not what was believed about her intent.

**Summary of Experiment 1.** Participants attributed cause based on differences in a person's knowledge of a potential risk. Although her actions were exactly the same, when Julia knew that her lawnmower might malfunction, participants tended to judge her as being more causal of the destruction of her tulips than when she had no knowledge that her lawnmower might malfunction. These findings conflict with legal and moral theories, which dictate that analysis of a person's actions (and whether he or she actually caused an outcome) remains separate from analysis of the person's mental state.

## Experiment 2

Experiment 1 established that people may be judged as more causal of an undesirable outcome if they knew of a related potential risk. However, does knowledge alone make people

seem more causal, or can someone avoid this attribution by acting reasonably given the knowledge?

According to our hypothesis, knowledge leads to increased causal attribution through counterfactual thinking. We proposed that, compared with an ignorant actor, when an actor has knowledge, it is easier for others to generate counterfactuals about how the actor could have taken steps, within his or her control, to avoid the negative outcome. Such counterfactual thinking should increase causal attribution to the knowledgeable actor. For example, in Experiment 1, Julia could have taken several different actions to prevent the destruction of the tulips, such as getting the mower fixed or being particularly careful—things that she reasonably could do and perhaps should have done given knowledge of the risk of malfunction, but things she would be less likely to be expected to do when she did not have that knowledge.

If this hypothesis is true, then if a person takes reasonable precaution given his or her knowledge of the situation, doing so may eliminate an obvious counterfactual about how the person could have prevented the outcome. Eliminating an obvious counterfactual should, in turn, decrease causal assignment compared with a knowledgeable actor who did not take any reasonable precautions.

Experiments 2 and 3 further investigate when and why knowledge makes people seem especially causal and test the hypothesis that knowledge increases causal attribution through counterfactual thinking. Experiment 2 replicates Experiment 1 using a different scenario (driving instead of mowing) and explores what happens to ratings of causality when a person with knowledge takes a reasonable precaution to prevent the bad outcome.

In Experiment 2, there is a potentially dangerous pre-existing risk (bad brakes in a car). The owner of the car, Josh, lends his car to Sarah. Josh either (a) does not know about the bad brakes (i.e., No Knowledge), (b) knows and fails to warn Sarah (i.e., Knowledge), or (c) knows and does warn Sarah (i.e., Knowledge + Action). In every case, Sarah drives “a little recklessly” and gets into a car accident. We predicted that Josh would be rated as more causal of Sarah’s accident when he had knowledge about the brakes than when he did not (replicating Experiment 1), but that this difference would diminish when Josh took some precaution to try to prevent the outcome from occurring (and thus eliminated an obvious counterfactual<sup>5</sup>).

## Method

**Participants and procedure.** Participants were 210<sup>6</sup> undergraduates from introductory psychology classes at the University of Virginia who participated in partial fulfillment of a course requirement. Participants were given 1 hr to complete this and other short reasoning experiments. Up to three participants were in the lab simultaneously. No participant took the entire time to complete the experiments.

**Materials and design.** Participants read one of three versions of a scenario about Josh who lent his car to Sarah. In all conditions, the brakes in Josh’s car were problematic, Sarah drove recklessly, the brakes failed, and Sarah got into a car accident. Knowledge varied across the conditions: Either no one knew the brakes were bad (No Knowledge condition), Josh knew and did not mention it to Sarah (Knowledge condition), or Josh shared his knowledge with Sarah such that they both knew (Knowledge + Action condition). We can therefore investigate how possessing knowledge affects causal attribution and how taking preventive action given knowledge affects causal attribution. The three versions began differently:

| Condition          | <i>n</i> | Language  |
|--------------------|----------|---|
| No Knowledge       | 69       | (Nothing)   |
| Knowledge          | 72       | Josh knew that the brakes on his car were not working properly. He lent the car to Sarah but did not mention the brake problem. |
| Knowledge + Action | 69       | Josh knew that the brakes on his car were not working properly. He lent the car to Sarah and mentioned the brake problem.       |

The conditions all continued,

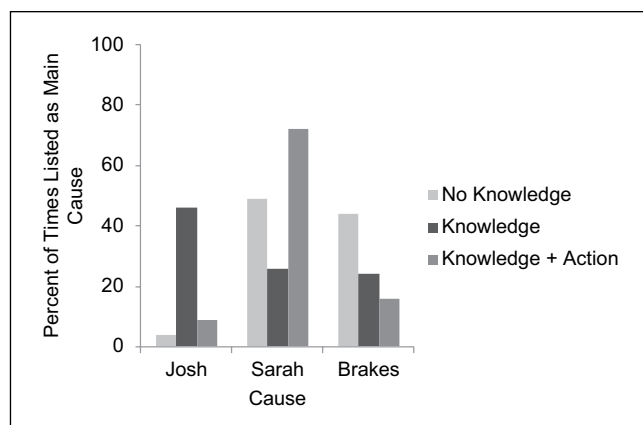
As Sarah drove a little recklessly along a twisting road, the brakes failed, and she crashed into another car. She was not injured, but the other driver was badly hurt in the accident.

The dependent measures were the same as in Experiment 1 but adapted to the current scenario. Participants answered a free-response question regarding the main cause of the car accident (but did not score the response themselves as they had in Experiment 1). Participants also rated the causality of three people or things: Josh, Sarah, and the brakes.<sup>7</sup> We added two additional hypothetical exploratory measures asking about whether Josh thought Sarah would get into an accident and whether he should have thought Sarah would get into an accident.

## Results and Discussion

**Main cause.** Josh’s and Sarah’s knowledge of the brakes affected the free-response judgments about the main cause of the car accident (see Figure 3). Two independent raters read each free-response answer and coded it into one of six categories: Josh, Sarah, Brakes, Josh and Brakes (both), Sarah and Brakes (both), and Other. There was initially 95.2% agreement between ratings; disputes were resolved by discussion.

**Josh.** As in Experiment 1, we used the free responses to see the proportion of the time that a participant selected

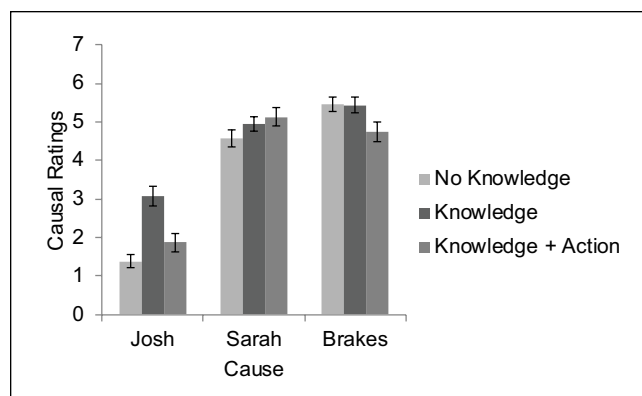


**Figure 3.** Percentage of free responses listing Josh, Sarah, and the car brakes as the main cause across conditions in Experiment 2. Note. In the No Knowledge condition, no one knew the brakes were bad. In the Knowledge condition, only Josh knew the brakes were bad. In the Knowledge + Action condition, Josh warned Sarah that the brakes were bad.

Josh as the main cause rather than something else. A participant's response of just Josh ( $n_{\text{No Knowledge}} = 1$ ,  $n_{\text{Knowledge}} = 32$ ,  $n_{\text{Knowledge+Action}} = 5$ ) was scored 1. A response of Josh and another item ( $n_{\text{No Knowledge}} = 4$ ,  $n_{\text{Knowledge}} = 2$ ,  $n_{\text{Knowledge+Action}} = 3$ ) was 0.5. A response without Josh ( $n_{\text{No Knowledge}} = 64$ ,  $n_{\text{Knowledge}} = 38$ ,  $n_{\text{Knowledge+Action}} = 61$ ) was 0.

A one-way ANOVA showed that how often participants listed Josh as the main cause of the car accident varied by condition,  $F(2, 207) = 30.85$ ,  $p < .001$ ,  $\eta^2 = .230$ . In the No Knowledge condition, participants almost never selected Josh as the main cause. However, in the Knowledge condition, they selected Josh more frequently. In the Knowledge + Action condition, again participants rarely selected Josh as the main cause. Fisher's least significant difference post hoc tests revealed that the Knowledge condition was significantly different from the No Knowledge and Knowledge + Action conditions ( $ps < .001$ ,  $d = 1.14$ , 95% CI = [0.79, 1.50] and  $d = 0.94$ , 95% CI = [0.59, 1.28], respectively), and the latter two were not significantly different from each other ( $p = .386$ ,  $d = 0.22$ , 95% CI = [-0.56, 0.11]). Thus, participants frequently believed Josh was the main cause of the car accident when he had knowledge of the faulty brakes, but they were much less likely to list him as the main cause when he passed that knowledge on to Sarah.<sup>8</sup>

**Causal ratings.** The causal ratings showed the same pattern of results as the free-response main cause answers (see Figure 4). Josh's knowledge made him, but not others, seem more causal on average. A MANOVA with causal ratings of Josh, Sarah, and Brakes as three outcome variables and knowledge (No Knowledge vs. Knowledge) as a predictor revealed that knowledge significantly affected causal ratings,  $F(1, 139) = 29.66$ ,  $p < .001$ ,  $\eta_p^2 = .131$ . Josh was rated more causal in the Knowledge condition ( $M = 3.1$ ,  $SD = 2.1$ ) than in the No Knowledge condition ( $M = 1.4$ ,  $SD = 1.5$ ),  $t(139) = 5.45$ ,



**Figure 4.** Average causality ratings (0-7, with standard errors) for Josh, Sarah, and the car brakes across conditions in Experiment 2.

Note. In the No Knowledge condition, no one knew the brakes were bad. In the Knowledge condition, only Josh knew the brakes were bad. In the Knowledge + Action condition, Josh warned Sarah that the brakes were bad.

$p < .001$ ,  $d = 0.92$ , 95% CI = [0.56, 1.26]. As might be expected, Josh's knowledge did not significantly affect causal attributions of Sarah or Brakes,  $t(139) = -1.34$ ,  $p = .184$ ,  $d = 0.23$ , CI = [-0.11, 0.56] and  $t(139) = 0.22$ ,  $p = .909$ ,  $d = 0.02$ , CI = [-0.31, 0.35], respectively.

Josh's preventive action (passing his knowledge to Sarah) also affected his causal rating—by reducing it. An analogous MANOVA with condition (Knowledge vs. Knowledge + Action) as a predictor showed that action affected causal ratings,  $F(1, 139) = 12.67$ ,  $p = .001$ ,  $\eta_p^2 = .084$ . Josh was rated less causal in the Knowledge + Action condition ( $M = 1.9$ ,  $SD = 1.9$ ) than in the Knowledge condition,  $t(139) = 3.56$ ,  $p = .001$ ,  $d = 0.60$ , 95% CI = [0.26, 0.94]. Sarah's causal ratings, perhaps unexpectedly, were similar in these conditions,  $t(139) = -0.51$ ,  $p = .61$ ,  $d = 0.08$ , CI = [-0.24, 0.42], possibly because it was easy to generate counterfactuals about how she should have driven carefully regardless of her knowledge. Moreover, Josh's causal rating was only somewhat higher in the Knowledge + Action condition than in the baseline No Knowledge condition,  $t(136) = -1.66$ ,  $p = .099$ ,  $d = 0.28$ , CI = [-0.05, 0.62]. This latter finding supports the idea that knowledge alone does not necessarily substantially increase causality; using one's knowledge to take preventive action (and thus eliminating an obvious counterfactual) might protect a person from increased causal attributions.

**Exploratory hypothetical questions.** Across conditions, only a small minority (<10% overall) of participants reported believing that Josh thought Sarah would probably or definitely get into an accident. However, participants were quite likely to think that Josh *should have* thought Sarah would get into an accident. (For full details, see Appendix B, Tables B1 and B2.) Moreover, in the Knowledge + Action condition, the more participants believed Josh should have thought



Sarah would get into an accident, the more causal they rated him,  $r = .45$ ,  $p < .001$ . This relationship, however, did not extend to the No Knowledge,  $r = .095$ ,  $p = .438$ , or Knowledge,  $r = -.083$ ,  $p = .489$ , conditions. Thus, it likely does not explain the relationship between knowledge and causal attribution.

Only one participant in any condition (Knowledge + Action) thought that Josh intended that Sarah have an accident. Thus, the results here are a reflection of what was known about Josh's knowledge, not what was believed about his intent.

### Experiment 3

Experiments 1 and 2 showed that an actor who had more knowledge about a potential risk factor was judged to be more causal of a bad outcome than one who did not. Experiment 2 also showed that when an actor took preventive action given his knowledge, his causal attribution decreased. This latter finding is consistent with the hypothesis that knowledge increases causal attribution because it induces people to imagine ways the knowledgeable actor could have prevented the bad outcome (and by taking preventive action, the actor eliminates an obvious way the outcome could have been prevented). For Josh and Sarah (Experiment 2), for instance, it is easier to generate such counterfactuals about Josh when he did not share his knowledge with Sarah compared with when he did. One can generate the counterfactual "if only Josh had told Sarah about the brakes, she would have been more careful and not gotten into the accident." Experiment 2 implicitly supports the hypothesis that such counterfactual thinking is the mechanism by which knowledge increases causal attribution; Experiment 3 explicitly tests that hypothesis.

Importantly, this hypothesis highlights that not all counterfactuals are equal; instead, certain *types* of counterfactuals are more likely to affect an actor's causation level. Specifically, we suggest that the effect of a counterfactual on an actor's causality depends on two criteria. First, the counterfactual must plausibly prevent the outcome (i.e., be *potent*). Petrocelli et al.'s (2011) research suggests that effective, or "potent," counterfactuals are those about things that could reasonably happen and that could actually change the outcome.

Second, at least for human actors, the counterfactual must be controlled by the actor whose causation is in question. Although potency provides a basis for assessing the impact of a given counterfactual, it does not directly address who or what will be judged more or less causal. For instance, imagining that Sarah had driven more carefully would likely be a potent counterfactual, but we predict it would not affect Josh's causation (except perhaps indirectly by increasing Sarah's causal attribution). Thus, we propose that this second factor, *control*, must be considered when assessing how a counterfactual will affect a particular actor's causal attribution.<sup>9</sup>

This proposal is consistent with prior research showing that people are more likely to generate counterfactuals about controllable, rather than uncontrollable, antecedents (e.g., Girotto et al., 1991; McEleney & Byrne, 2006), and that control is relevant to causation and blame assignment (e.g., Alicke, 2000; McGill & Tenbrunsel, 2000).

Experiment 3 tests the hypotheses that (a) compared with when an actor does not have relevant knowledge, when an actor has knowledge, people generate counterfactuals that are more potent and more under the control of the knowledgeable actor and (b) this counterfactual thinking is the mechanism by which knowledge increases causation. Participants read a vignette from Experiment 1. However, unlike in Experiment 1, prior to evaluating causality, participants were asked to list counterfactuals and rate their potency (i.e., their If likelihood and Then likelihood) and controllability. The mediation effects of such counterfactual thinking were analyzed.

### Method

**Participants and procedure.** Participants were 142 undergraduates<sup>10</sup> from the same participant pool who had not participated in the previous experiment. Experiment 3 was run in the same manner as Experiment 2.

**Materials and design.** The materials were similar to Experiment 1: A character named Julia is mowing her lawn, the lawnmower begins to rumble back and forth wildly, and her tulips are destroyed. The only differences were that there was no Main Cause question, and participants answered questions about counterfactuals in addition to questions about causality and intent.

The first counterfactual question read, "List four ways that things could have turned out differently, so that the tulips would *not* have been destroyed." Participants were provided lines beginning with "The tulips would not have been destroyed *if only* . . ." and space to write in their counterfactuals.

Participants were then asked to provide If and Then likelihoods (see Petrocelli et al., 2011) and a control rating, as measured by these questions:

- i. If: "How likely do you think it is that what you wrote ACTUALLY could have happened" (from 0 to 10, where 0 = *no chance at all* and 10 = *easily could happen*),
- ii. Then: "IF what you wrote ACTUALLY DID HAPPEN, how likely is it that the tulips would have been preserved?" (from 0 to 10, where 0 = *definitely ruined anyway* and 10 = *certainly preserved*), and
- iii. Control: "Do you believe that this is something that Julia could have controlled?" (from 0 to 10, where 0 = *definitely could NOT control* and 10 = *Definitely could control*).



Participants then did the same causality ratings and measure of intent as in Experiment 1. (See Appendix C for notes on material design, piloting, and data analysis for Experiment 3.)

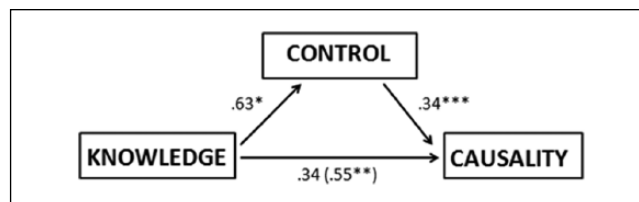
## Results and Discussion

Again, Julia was rated more causal when she had knowledge than when she did not. Moreover, counterfactual thinking—in particular, the production of counterfactuals over which Julia had control—mediated these results in Experiment 3.

**Main effect of knowledge on cause and meta-analytic effect size estimates.** The results from Experiments 1 and 2 were replicated. When Julia had prior knowledge of possible mechanical trouble, she was rated more causal ( $M = 5.7$ ,  $SD = 1.1$ ) than when she had no knowledge ( $M = 5.1$ ,  $SD = 1.3$ ),  $t(140) = 2.67$ ,  $p = .009$ ,  $d = 0.45$ , 95% CI = [0.11, 0.78]. Meta-analysis of the effect size of knowledge condition on causation ratings across all studies using the tulips vignette (Experiments 1 and 3 as well as the two preliminary studies reported in Footnote 4) estimated a moderate effect, ( $d = 0.40$ , standard error [SE] = 0.09, 95% CI = [0.23, 0.56]; random effect model). Meta-analytic estimates for causation ratings incorporating the car borrowing vignette (Experiment 2) as well as meta-analysis of the main cause results (Experiments 1 and 2) similarly suggest a moderate to large effect size of knowledge condition ( $d = 0.51$ ,  $SE = 0.11$ , 95% CI = [0.29, 0.73]; random effect model; and  $d = 0.87$ ,  $SE = 0.43$ , 95% CI = [0.03, 1.71]; logit analysis, random effect model, respectively; see Appendix D).

**Potency.** Consistent with Petrocelli et al. (2011), participants' answers to the If and Then questions were multiplied for each counterfactual then averaged across the four counterfactuals to form a measure of counterfactual potency ("potency") with possible scores of 0 to 100. (Note that the Then rating is a conditional probability; it assumes the occurrence of the If event.) As predicted, combining across knowledge conditions, a hierarchical linear regression (used because potency is an interaction term; see Cohen, 1978) revealed that, after accounting for mean If and Then ratings separately (step 1), potency (step 2) predicted Julia's causation rating,  $b = .12$ ,  $SE = 0.035$ ,  $p = .001$ , 95% CI = [0.05, 0.19]. As potency increased, so did Julia's causation.

Exploring across knowledge conditions, however, participants in the Knowledge condition ( $M = 64$ ,  $SD = 15$ ) rated their counterfactuals on average as only slightly (and non-significantly) more potent than participants in the No Knowledge condition ( $M = 61$ ,  $SD = 13$ ),  $t(140) = 1.58$ ,  $p = .117$ ,  $d = 0.26$ , 95% CI = [-0.06, 0.59]. The similarly high potency scores may have occurred because the vignette lent itself to multiple potent counterfactuals regardless of condition (e.g., "if only the lawnmower hadn't malfunctioned") or because participants were explicitly instructed to generate counterfactuals that could have changed the outcome.



**Figure 5.** Diagram showing that control mediated the relationship between knowledge and causality.

Note. Decimals represent  $b$  weights.

\* $p < .05$ .

Moreover, after accounting for If and Then ratings, when Knowledge condition (No Knowledge = 1, Knowledge = 2) was added to the hierarchical linear regression in the same step as potency, knowledge ( $b = .49$ ,  $SE = 0.19$ ,  $p = .012$ , 95% CI = [0.11, 0.87]) and potency ( $b = .12$ ,  $SE = 0.03$ ,  $p < .001$ , 95% CI = [0.06, 0.19]) predicted causality independently. Thus, although potency affected causal judgments, it does not appear to explain the relationship between knowledge and causality.

**Control.** Mean control rating had a positive relationship with Julia's causality,  $b = .34$ ,  $SE = 0.06$ ,  $p < .001$ , 95% CI = [0.22, 0.46]. In addition, the mean control rating in the Knowledge condition ( $M = 8.2$ ,  $SD = 1.5$ ) was higher than in the No Knowledge condition ( $M = 7.6$ ,  $SD = 1.5$ ),  $t(140) = 2.52$ ,  $p = .013$ ,  $d = 0.42$ , 95% CI = [0.09, 0.76], indicating that participants generated more counterfactuals about things Julia had more control over when she had knowledge compared with no knowledge.

Further analysis revealed that control acted as a mediator of the relationship between knowledge and cause. When control and knowledge were entered into a simultaneous regression predicting causality, control remained a significant, positive predictor of causality ( $b = .31$ ,  $SE = 0.06$ ,  $p < .001$ ), but the effect of knowledge decreased and was marginally significant ( $b = .34$ ,  $SE = 0.19$ ,  $p = .071$ ). Using bootstrapping analysis (Preacher & Hayes, 2004) with 5,000 bootstrap samples, we tested the indirect effect of knowledge on causality through control. The estimated indirect effect was 0.20 ( $SE = 0.09$ ), 95% CI = [0.04, 0.40], indicating that control mediated the relationship between knowledge and causality in this experiment (see Figure 5). These results support the following explanation: Compared with when Julia did not have knowledge, when Julia had knowledge, participants thought of more ways the outcome could have been different that were under her control. In turn, such counterfactual thinking increased her causation ratings.

Moreover, potency does not appear to mediate or moderate the relationship between control and causality. Multiple mediation analysis using Hayes's (2012) PROCESS Model 6 ( $X$  = condition,  $M_1$  = Control,  $M_2$  = If ratings,  $M_3$  = Then ratings,  $M_4$  = Potency,  $Y$  = Julia's causality) revealed that a

model incorporating potency did not have a reliable indirect effect on cause (Total Indirect Effect = 0.16,  $SE = 0.12$ , 95% CI =  $[-0.05, 0.41]$ ). Moreover, control mediated the relationship between knowledge condition and causality even when potency was relatively low. When If and Then ratings were a standard deviation below the mean, control continued to be a reliable mediator (Indirect Effect = 0.15,  $SE = 0.10$ , 95% CI =  $[0.01, 0.42]$ ). This may have occurred because, as noted above, mean potency ratings were moderately high ( $M_s = 61$  and  $64$  on the 100-point scale) and did not significantly differ by knowledge condition. Indeed, less than 5% (7/142) of mean potency ratings fell below 40 in either condition. In addition, the interaction between potency and control was not a better predictor than potency and control alone. When entered into a hierarchical linear regression, after accounting for If, Then, and control ratings separately (Step 1) as well as potency, If  $\times$  control, and Then  $\times$  control (Step 2), potency  $\times$  control (Step 3) did not reliably predict ratings of Julia's causality,  $r^2$  change = .002,  $p = .533$ ,  $b = -.015$ ,  $SE = 0.025$ , 95% CI =  $[-0.064, 0.033]$ .

In summary, potency and control were each good predictors of causality ratings. However, the results of the mediation analyses suggest that knowledge affects causation ratings by increasing the generation of controllable counterfactuals rather than by influencing counterfactual potency. That is, regardless of Julia's knowledge, participants tended to generate reasonably potent counterfactuals about how the outcome could have been prevented. However, when Julia had knowledge, participants were more likely to generate counterfactuals that she had more control over, and these more controllable counterfactuals mediated the relationship between knowledge and causation.

## General Discussion

These three experiments found that people with knowledge were judged to be more causal of bad outcomes than their more ignorant counterparts. Moreover, these experiments provide evidence that the mechanism for the relationship between knowledge and cause is counterfactual thinking. Specifically, our findings support the hypothesis that when an individual has knowledge relevant to a bad outcome, people are more likely to generate counterfactuals that the knowledgeable individual could control. And such counterfactual thinking in turn increases causation ratings for the individual.

These findings suggest that the way people assess causation conflicts with philosophical and legal theories of blame (e.g., Brown, 2012; Heider, 1958; Kelley, 1973), which generally prescribe that causal judgments and mental state assessments should occur separately. Moreover, these findings suggest that the relationship between mental state and blame is not necessarily due to a top-down process in which blame assessment occurs first and in turn affects causal judgments (see, for example, Alicke, 2000). Instead, these findings are consistent with the theory that mental state assessment can

affect blame judgments via a bottom-up process based on counterfactual and causal reasoning. That is, these findings are consistent with the following explanation: When an actor has knowledge relevant to a bad outcome, people are more likely to generate counterfactuals that the knowledgeable actor could control, such counterfactuals increase causal attribution to the knowledgeable actor, and these causal judgments in turn increase blame judgments (see, for example, Malle, Guglielmo, & Monroe, 2014).

## Support for Potency and the Crediting Causality Model

These findings replicate prior research showing that the more potent a counterfactual is rated—in other words, the more it actually could have occurred, and the more its occurrence could have changed the outcome—the more effect the counterfactual is likely to have on causation judgments (Petrocelli et al., 2011). They are also consistent with the Crediting Causality Model of causation, which proposes that an actor is judged to be causal to the extent that he or she increased the probability of the bad outcome occurring over baseline.

Counterfactual thinking can provide information about how likely an outcome would be normally and how likely it would be given the potential cause. For instance, consider the role of Julia and the lawnmower in the destruction of Julia's tulips. Compared with when she did not have knowledge that her lawnmower might malfunction, when Julia had knowledge, participants were more likely to imagine that she could have acted in ways that would have prevented the outcome. This counterfactual thinking affected her assigned causation.

## Support for the Importance of Control

Our results also further clarify how counterfactuals affect causation judgments by illuminating the importance of counterfactual control. We suggest that whereas potency affects *how much* a counterfactual influences causation, control helps explain *who or what* will be assigned that causation. And, at least for humans, counterfactuals might only affect causation ratings when they are reasonably controllable. For instance, in the case of Julia, a potent counterfactual might be "if only it had been raining, Julia would not have mowed, and the flowers would be preserved." However, because Julia has no control over rain, this counterfactual would be unlikely to affect causal ratings of her.

Moreover, control alone mediated the relationship between knowledge and causation. Together, these findings suggest that knowledge increases causation ratings of the knowledgeable actor because people generate counterfactuals that the actor could control more when the actor has knowledge of a potential risk than when she does not have knowledge. Future research, however, should further examine whether potency might moderate the effects of control. In our study, perhaps because they were instructed to do so or because the vignette lent itself to multiple

effective counterfactuals, participants generated reasonably potent counterfactuals regardless of knowledge condition. However, theoretically, a very controllable but non-potent counterfactual should have little effect on causal ratings.

That controllable counterfactuals are particularly likely to increase causation ratings is consistent with functional theories of counterfactual thinking. According to such theories, counterfactual thinking helps people think through ways to behave that will improve future performance (e.g., Epstude & Roese, 2008; Roese & Olson, 1995). For the Julia story, for example, focusing on counterfactuals that Julia could control (such as fixing the lawnmower or being more careful) would improve her ability, or the ability of someone else in her position, to prevent destroying flowers in the future.

It is even possible that the current results underestimate the importance of control on causation, because previous research suggests that people are more likely to spontaneously mutate controllable antecedents in the first place (e.g., McEleney & Byrne, 2006). Future work should address more complex real-world situations; in such cases, control may play a dual role in affecting causation—first by spurring the generation of automatic counterfactual thinking about controllable antecedents, and then by magnifying the effect of those antecedents on causation judgments.

### *Alternative Explanations: Intent and Morality*

Whereas we show a relationship between mere knowledge and causation (see Robinson & Darley, 1995, Experiment 8, on knowledge and liability), previous work suggests that immorality and intent also may increase causation ratings (Alicke, 1992; Knobe, 2005; Pizarro, Uhlmann, & Salovey, 2003). Immorality and intent could even be rationally related to causal ratings if they increase the chance the actor actually behaved in a way that (at least she believed) would cause the outcome. Thus, one explanation for our results could be that knowledgeable actors are simply judged to be more immoral or to have more bad intent. If, for instance, someone had knowledge that could prevent a bad outcome but failed to act on it, then she might be viewed as more immoral or people might believe she intended to cause the bad outcome.

However, these experiments were designed to eliminate effects of morality or intent. In Experiments 1 and 2, when specifically asked, hardly any participants reported believing that the actor intended the bad outcome. In addition, in Experiment 1 (and 3), the bad outcome only affected the actor herself (i.e., Julia destroyed her own tulips). Thus, we believe that it is indeed knowledge of a risk, and not intent or morality, that is affecting causation ratings.

### *Findings Inconsistent With Prescriptive Theories of Blame and With Law*

Philosophical and psychological theories of blame traditionally presume that people use a stepwise process, in which

whether an actor caused an outcome is a completely separate consideration from the actor's mental state (see, for example, Heider, 1968; Malle et al., 2014; Shaver, 1985). Similarly, the U.S. legal system<sup>11</sup> generally requires two distinct "elements" to be proven before someone can be convicted of a crime: First, the person must have perpetrated the act that caused the outcome, and, separately, the person must have had the requisite mental state. Thus, in psychological theories of blame and in law, determining whether an actor caused an outcome is a *separate* step from considering an actor's prior knowledge or mental state. The findings of these experiments challenge the assumptions behind these models by suggesting that people do not really keep those "elements" separate (see also Spellman & Gilbert, 2014).

Of course, it seems reasonable that one's behavior can influence judgments of his or her mental state. For instance, if a defendant slowly and repeatedly stabbed a victim, this could serve as evidence that the defendant knew his or her actions were likely to lead to the death of the victim (and that the death was probably intended). Our findings, however, suggest the relationship between action and mental state may automatically go the opposite direction as well. That is, if someone believes a defendant merely *knew* of a potential risk related to a bad outcome, then people may believe the defendant is more likely to have actually *caused* the bad outcome. And arguably, there is no logic to this reverse relationship; either the actor caused the outcome, or she did not. In the field of psychology and law, future work could examine whether confounding knowledge with cause might unreasonably prejudice defendants, and whether jury instructions or rules of evidence can be drafted to limit any prejudicial effects of such thinking.

### *Findings Consistent With Theories of Negligence and Recklessness Law*

Our findings, however, do not suggest that knowledge of a potential risk alone increases causal attribution; instead, knowledge increases causal attribution through counterfactual thinking. Thus, it is possible that knowledgeable actors may escape causal attribution by taking preventive action that eliminates obvious controllable counterfactuals. Experiment 2, for instance, showed that an actor (Josh) was considered more causal of a car accident when he knew the car had brake problems than when he did not. However, when he shared his knowledge with the car's driver (Sarah), his causal ratings dropped significantly, such that they were only marginally higher than when he had no knowledge at all. This finding is consistent with the theory that when Josh had knowledge, his causation increased because people could easily imagine the controllable counterfactual "if only Josh had told Sarah she would have been more careful or otherwise avoided the accident." However, by telling Sarah of the brake problem, this counterfactual—and the increase in causation that it would likely confer on him—is eliminated.

That taking reasonable preventive action may decrease causal attribution is also consistent with legal concepts of negligence and recklessness. Negligence (failing to take the kind of precaution that a reasonably prudent person would) and recklessness (knowing the possibility of specific risks but taking a risk anyway) rely on the idea that people should take reasonable care to prevent foreseeable harm. The importance of foreseeability is further highlighted by Experiment 2's exploratory finding that the more participants thought the actor should have predicted the outcome, the more causal they rated him (similar to objective foreseeability in Lagnado & Channon, 2008).

## Conclusion

Together, these experiments support the extended Crediting Causality Model, showing that knowledge increases causal assignment, and it does so through affecting the types of counterfactuals people generate. Moreover, these findings build off Petrocelli et al.'s (2011) findings showing that not all counterfactuals are equally important—and that in the case of knowledge, and perhaps in all analyses of human actors, the ability of an agent to have created an alternative state of affairs is particularly important.

## Appendix A

### *Relation Between the Extended Crediting Causality Model (Spellman, Kincannon, & Stose, 2005) and Counterfactual Potency (Petrocelli, Percy, Sherman, & Tormala, 2011)*

Spellman et al. (2005). Causal ratings depend on how much someone (or something) changed the probability of an outcome from before the event (or action) to after the event (or action) in question.

$$\text{Causality} = \text{function of (probability of outcome after)} - (\text{probability of outcome before}) \quad (1a)$$

In the book chapter, it is expressed like this:

$$C \approx p(O_{\text{after}}) - p(O_{\text{before}}) \quad (1b)$$

Then, it is expanded to Equation 2 below. However, for this appendix, we ignore the denominators (which in simple cases equal 1).

$$C \approx \frac{p(O_{\text{after}})}{p(O_{\text{after}}) + p(\sim O_{\text{after}})} - \frac{p(O_{\text{before}})}{p(O_{\text{before}}) + p(\sim O_{\text{before}})} \quad (2)$$

Given that the events in question are one-time events, how can people estimate the probability that an outcome would happen? One way would be to consider all of the ways that the world could unfold. For example, in the Robinson and Darley (1995) shooter scenario, after the victim threatens him with a weapon, the shooter could have run away (and not shot), or stayed and talked (and maybe shot), or just shot.

For each of those ways, we can sum up, for each way, the product of the likelihood that that "way" would happen [ $p(\text{way})$ ] and the probability of the outcome of getting shot dead given that the world had unfolded in that way [ $p(O|\text{way})$ ]. We do that for both after and before the causal event in question.

$$C \approx \frac{\sum_{\text{ways}} p(\text{way}) \times p(O_{\text{after}}|\text{way})}{\sum_{\text{ways}} p(\text{way}) \times p(O_{\text{after}}|\text{way}) + \sum_{\text{ways}} p(\text{way}) \times p(\sim O_{\text{after}}|\text{way})} - \frac{\sum_{\text{ways}} p(\text{way}) \times p(O_{\text{before}}|\text{way})}{\sum_{\text{ways}} p(\text{way}) \times p(O_{\text{before}}|\text{way}) + \sum_{\text{ways}} p(\text{way}) \times p(\sim O_{\text{before}}|\text{way})} \quad (3)$$

Unpacking the big equation (again, just looking at the numerators because here, the denominators equal 1) is illustrated by the questions below and Table A1.

What was the probability of getting shot dead *before* the victim attacked the shooter? There might be some probability of getting shot, for example, .1.

What was the probability of getting shot dead *after* the victim attacked the shooter?

- The shooter could have run away. Probability of that happening = .6, and if that happened, probability of getting shot = 0.
- The shooter could have stayed and talked. Probability of that happening = .1, and if that happened, probability of getting shot = .5.
- Or the shooter could have just shot. Probability of that happening = .3, and if that happened, probability of getting shot = 1.

So after he attacked the shooter, we sum all of those:

$$(.6 \times 0) + (.1 \times .5) + (.3 \times 1) = .35 \text{ probability of getting shot.}$$

Then, we subtract the probability before he attacked, which is .1. And we see that by attacking the shooter, the victim raised the probability of getting shot by .25.

Where do these possible scenarios and probabilities come from? As we have said, they come from pre-existing knowledge and counterfactual reasoning.

Petrocelli et al. (2011). Petrocelli et al. asked their participants to generate counterfactuals to a scenario (Studies 1 and 3) or provided their participants with a specific counterfactual (Studies 2 and 4) that could have changed an outcome. They then asked participants to rate each counterfactual on a scale for its If likelihood and Then likelihood. The specific question to measure each likelihood varied slightly across the four studies, but to measure the If likelihood, participants were generally asked to rate the likelihood of the counterfactual “actually” occurring.<sup>12</sup> To measure the Then likelihood, participants were asked to rate how likely it was that the outcome could have been changed or avoided.<sup>13</sup>

In our Experiment 3, for the questions using their method, we asked,

- If: “How likely do you think it is that what you wrote ACTUALLY could have happened” (from 0 to 10, where 0 = *no chance at all* and 10 = *easily could happen*).
- Then: “IF what you wrote ACTUALLY DID HAPPEN, how likely is it that the tulips would have been preserved?” (from 0 to 10, where 0 = *definitely ruined anyway* and 10 = *certainly preserved*).

These questions evoke the same two important pieces of information as described in Spellman et al. (2005): the probability of the world unfolding a specific “way” (i.e., the If likelihood) and the conditional probability of the outcome given that the world unfolded that way (i.e., the Then likelihood). Multiplying If and Then gives “potency” for each counterfactual—and that product is exactly what is labeled “Product” in Table A1.

(There are some complications when the sum of a participant’s Ifs is greater than 100% or when experimenters use average potencies, but the idea of how counterfactuals affect causal judgments is fundamentally the same in the two articles.)

**Table A1.** Comparing Potency (Petrocelli et al., 2011) and Causality (Spellman et al., 2005) Evaluations for Counterfactuals in the Robinson and Darley (1995) Example in the Text Above.

|                                       | Probability of the world unfolding that “way”? (“If”) | Probability of getting shot given that “way”? (i.e., conditional probability: “Then”) | Product   |
|---------------------------------------|---|---|-----------|
| Before attacking shooter              |   |   |           |
| Crazy shooter might have shot for fun | .1  | 1   | .1        |
| Shooter just walks past               | .9  | 0   | 0         |
| Total before                          | (1)   |   | Sum = .1  |
| After attacking shooter               |   |   |           |
| Shooter runs away                     | .6  | 0   | 0         |
| Shooter stays and talks               | .1  | .5  | .05       |
| Shooter just shoots                   | .3  | 1   | .3        |
| Total after                           | (1)   |   | Sum = .35 |

Note. The potency for each counterfactual is represented by a product. The overall CCM causality is represented by the difference between the two sums.

## Appendix B

### Experiment 2 Results Regarding Josh’s Beliefs

**Table B1.** Judgments of Whether Josh Thought Sarah Would Get Into an Accident.

| Condition          | Percent who thought Josh thought Sarah would get into an accident (raw count in parentheses) |         |          |            |
|--------------------|--|---------|----------|------------|
|                    | Not at all   | Maybe   | Probably | Definitely |
| No Knowledge       | 32 (22)  | 55 (38) | 13 (9)   | 0 (0)      |
| Knowledge          | 42 (30)  | 54 (39) | 4 (3)    | 0 (0)      |
| Knowledge + Action | 30 (21)  | 58 (40) | 10 (7)   | 1 (1)      |

**Table B2.** Judgments of Whether Josh Should Have Thought Sarah Would Get Into an Accident.

| Condition          | Percent who thought Josh should have thought Sarah would get into an accident (raw count in parentheses) |         |          |            |
|--------------------|--|---------|----------|------------|
|                    | Not at all   | Maybe   | Probably | Definitely |
| No Knowledge       | 9 (6)  | 35 (24) | 33 (23)  | 23 (16)    |
| Knowledge          | 6 (4)  | 14 (10) | 39 (28)  | 42 (30)    |
| Knowledge + Action | 7 (5)  | 32 (22) | 35 (24)  | 26 (18)    |

## Appendix C

### Notes on Material Design, Piloting, and Data Analysis for Experiment 3

Prior to Experiment 3, pilot testing used the prompt, “Julia was upset about her tulips being destroyed. She started thinking . . . IF ONLY . . . [Line break] How do you think Julia might have finished that sentence? List three ways she might have completed it.” However, eyeballing pilot responses revealed that when this prompt was used, participants tended to focus on only things that related to Julia. Thus, the prompt was changed to ensure that participants did not assume Julia’s perspective only. Previous research has used both types of prompts (see, for example, Mandel & Lehman, 1996 [“if only” prompts]; Giroto, Legrenzi, & Rizzo, 1991 [listing ways the outcome could be “undone”]). We believe that “she/he began thinking if only” prompts induce participants to focus on the actor’s thoughts, thus artificially making counterfactuals related to the actor more available.

Consistent with Petrocelli, Percy, Sherman, and Tormala (2011), mean potency of all four counterfactuals was used because research suggests that the most influential

counterfactual may occur at any point in a list (Spellman & Gilbert, 2012). However, analyzing only the first counterfac-

tual did not change the results for any regression or mediation analyses.

## Appendix D

### Meta-Analytic Effect Size Estimates

**Table D1.** Meta-Analytic Summary of Effect Sizes of Knowledge Condition (No Knowledge vs. Knowledge) on Causal Ratings Across Experiments.

| Study  | <i>n</i> by condition<br>(No Knowledge,<br>Knowledge) | <i>M</i> (No<br>Knowledge,<br>Knowledge) | <i>t</i> | Weight | <i>d</i> | SE of <i>d</i> | CI around <i>d</i> |
|--|---|--|----------|--------|----------|----------------|--------------------|
| Tulips experiments only  |   |  |          |        |          |                |                    |
| Experiment 1   | 102, 100  | 4.83, 5.43                               | 2.26     | 49.8   | 0.32     | 0.14           | [0.04, 0.59]       |
| Unreported lab experiment (Footnote 4, tulips, lab)                                | 37, 37  | 4.22, 5.16                               | 2.52     | 17.7   | 0.59     | 0.24           | [0.12, 1.05]       |
| Unreported MTurk experiment (Footnote 4, tulips, MTurk, no free-response question) | 71, 69  | 4.79, 5.41                               | 2.09     | 34.5   | 0.35     | 0.17           | [0.02, 0.69]       |
| Experiment 3   | 72, 73  | 5.14, 5.68                               | 2.76     | 34.6   | 0.45     | 0.17           | [0.11, 0.78]       |
| Random effect estimation for Tulips Experiments only (same as fixed)               |   |  |          |        | 0.40     | 0.09           | [0.23, 0.56]       |
| Tulips experiments and car experiment combined                                     |   |  |          |        |          |                |                    |
| Experiment 2   | 69, 72  | 1.39, 3.07                               | 5.45     | 31.9   | 0.92     | .18            | [0.56, 1.26]       |
| Fixed effect estimation  |   |  |          |        | 0.50     | 0.08           | [0.23, 0.54]       |
| Random effect estimation   |   |  |          |        | 0.51     | 0.11           | [0.29, 0.73]       |

Note. Random effect estimation was used when analyzing these data even when analyzing tulips vignettes only, as small population and procedure differences between studies (i.e., MTurk vs. lab populations) suggest there could be differences in underlying effect size between the studies. Results for the tulips studies only, however, did not differ between random and fixed effect analysis. CI = confidence interval; MTurk = Amazon Mechanical Turk.

**Table D2.** Meta-Analysis of Main Cause (Treated as a Proportion, Where 0 = Not *Julia/Josh* and 1 = *Julia/Josh*).

| Study   | <i>n</i> by condition<br>(No Knowledge,<br>Knowledge) | Proportion<br>main cause (No<br>Knowledge,<br>Knowledge) | $\chi^2$ | Weight | <i>d</i> (logit) | SE of <i>d</i> | CI around <i>d</i> |
|---|---|--|----------|--------|------------------|----------------|--------------------|
| Tulips experiments only                                 |   |  |          |        |                  |                |                    |
| Experiment 1 <sup>a</sup>                               | 102, 100  | 0.57, 0.73   | 5.02     | 36.45  | 0.37             | 0.17           | [0.05, 0.69]       |
| Unreported lab experiment <sup>b</sup><br>(Tulips, lab) | 37, 37  | 0.84, 0.95   | 2.24     | 4.53   | 0.67             | 0.47           | [-0.25, 1.59]      |
| Random effect estimation (same as random)               |   |  |          |        | 0.40             | 0.16           | [0.10, 0.71]       |
| Tulips experiments and car experiment combined          |   |  |          |        |                  |                |                    |
| Paper Experiment 2                                      | 69, 72  | 0.04, 0.46   | 30.93    | 7.59   | 1.66             | 0.36           | [1.04, 2.26]       |
| Fixed effect estimation                                 |   |  |          |        | 0.60             | 0.14           | [0.32, 0.88]       |
| Random effect estimation                                |   |  |          |        | 0.87             | 0.43           | [0.03, 1.71]       |

Note. Random effect estimation was used when analyzing these data even when analyzing tulips vignettes only, as small population and procedure differences between studies (i.e., MTurk vs. lab populations) suggest there could be differences in underlying effect size between the studies. Estimates for the tulips studies only, however, did not differ between random and fixed effect analysis. CI = confidence interval.

<sup>a</sup>Answers coded as both *Julia* and *Other* were given a value of 0.5 when calculating the proportion.

<sup>b</sup>Participants in this study did not self-code the main effect. Instead, the first and second authors did. They found it challenging to determine what counted as *Julia* (hence, why the study was re-run with participants coding their own responses). These results reflect what coding the authors agreed on after discussion.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. Kelley's (1973) covariation model proposes that people attribute causation based on whether a potential cause and effect occur together over time. That model, however, is conceptual rather than quantitative.
2. Appendix A unpacks this equation.
3. Note that it is important to specify what is meant by "the outcome" (see Mandel, 2003; Spellman, Kincannon, & Stose, 2005).
4. Sample size here (aiming for 100 per condition) was chosen to increase power for a medium effect size, after previously running one lower-powered study and one study that did not include the main cause question. Meta-analyses including the two otherwise unreported studies are provided in Experiment 3 and Appendix D.
5. In a check of this assumption, 61 Amazon Mechanical Turk (MTurk) participants read the Knowledge condition scenario and listed five counterfactuals. Ninety-seven percent of the participants listed that Josh could have warned Sarah. Participants rated this counterfactual as high in potency ( $M = 70$  out of 100) and controllability ( $M = 8.9$  out of 10).
6. Sample size for Experiments 2 and 3 was limited to the number of participants who could reasonably complete the study in one semester.
7. Questions about enabling the outcome were also asked but not analyzed for this article.
8. Analysis of Sarah's causal attribution complemented these findings. A one-way ANOVA on the proportion of the time that Sarah was selected as the main cause revealed that Sarah was listed as the main cause more often when she had knowledge of the problematic brakes (the Knowledge + Action condition) than when she did not have knowledge (the No Knowledge and Knowledge conditions),  $F(2, 207) = 19.22, p < .001$ .
9. Importantly, control may function independently from potency, and we propose that some baseline levels of both are necessary to affect causal attribution. Thus, imagining that a snowstorm stopped Josh from loaning his car to Sarah may be a potent counterfactual to avoid the car crash, but it is out of Josh's control and would thus be unlikely to change his causal attribution. Conversely, imagining that Sarah could have worn different clothes that day may be completely in her control but would not be potent in preventing the crash.
10. Although 145 participants began the study, 3 skipped questions, and their data were excluded prior to analysis.
11. Legal systems of many other countries around the world do to, especially those stemming from common law traditions (Badar & Marchuk, 2013).
12. In Study 3, however, which manipulated participants' perceptions of how unique or common their counterfactuals were, participants were asked to rate how "confident" they were

that their counterfactuals could have occurred and would have changed the outcome (Petrocelli et al., 2011).

13. Scale anchors matched the specifics of the study, such that 1 represented *extremely unlikely* (Studies 1 and 2), *not at all confident* (Study 3), or *I gave no consideration to picking the other [option]* (Study 4), and 7 or 9 represented *extremely likely* (Studies 1 and 2), *I very nearly picked the other [option]* (Study 3), and *it is certain [the other outcome would have occurred]* (Study 4).

## Supplemental Material

The online supplemental material is available at <http://pspb.sagepub.com/supplemental>.

## References

- Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63, 68-378.
- Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126, 556-574.
- Badar, M. E., & Marchuk, I. (2013). A comparative study of the principles governing criminal responsibility in the major legal systems of the world. *Criminal Law Forum*, 24, 1-48.
- Branscombe, N. R., Owen, S., Garstka, T. A., & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome? *Journal of Applied Social Psychology*, 26, 1042-1067.
- Brown, D. K. (2012). Federal mens rea interpretation and the limits of culpability's relevance. *Law and Contemporary Problems*, 75, 109-131.
- Byrne, R. M. (2007). Precis of the rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, 30, 439-452.
- Cohen, J. (1978). Partialled products are interactions: Partialled powers are curve components. *Psychological Bulletin*, 85, 858-866.
- Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12, 168-192.
- Giroto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78, 111-133.
- Goldinger, S. D., Kleider, H. M., Azuma, T., & Beike, D. R. (2003). "Blaming the victim" under memory load. *Psychological Science*, 14, 81-85.
- Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry: An Interdisciplinary Journal of Philosophy*, 52, 449-466.
- Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling* [White paper]. Retrieved from <http://www.personal.psu.edu/jxb14/M554/articles/process2012.pdf>
- Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: John Wiley.
- Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, 62, 135-163.
- Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219-266). San Diego, CA: Academic Press.



- Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28, 107-128.
- Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, 9, 357-359.
- Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33, 315-329.
- Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108, 754-770.
- Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, 25, 147-186.
- Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual, and covariational reasoning. *Journal of Experimental Psychology: General*, 132, 419-434.
- Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, 71, 450-463.
- Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1995). The impact of perceived control on the imagination of better and worse possible worlds. *Personality and Social Psychology Bulletin*, 21, 588-595.
- McCloy, R., & Byrne, R. M. (2002). Semifactual "even if" thinking. *Thinking & Reasoning*, 8, 41-67.
- McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, 37, 879-901.
- McEleney, A., & Byrne, R. M. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning*, 12, 235-255.
- McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal election. *Journal of Personality and Social Psychology*, 79, 677-689.
- Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, 100, 30-46.
- Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91-108). Washington, DC: American Psychological Association.
- Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, 14, 267-272.
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, 36, 717-731.
- Raz, J. (2010). Responsibility and the negligence standard. *Oxford Journal of Legal Studies*, 30, 1-18.
- Robinson, P. H., & Darley, J. M. (1995). *Justice, liability, and blame: Community views and the criminal law*. Boulder, CO: Westview.
- Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121, 133-148.
- Roese, N. J., & Olson, J. M. (1995). Outcome controllability and counterfactual thinking. *Personality and Social Psychology Bulletin*, 21, 620-628.
- Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer-Verlag.
- Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, 126, 323-348.
- Spellman, B. A., & Gilbert, E. A. (2012). *Counterfactual and causal reasoning: Similar, different, and related*. Unpublished manuscript.
- Spellman, B. A., & Gilbert, E. A. (2014). Blame, cause, and counterfactuals: The inextricable link. *Psychological Inquiry*, 25, 245-250.
- Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems*, 64, 241-264.
- Spellman, B. A., Kincannon, A., & Stose, S. (2005). The relation between counterfactual and causal reasoning. In D. R. Mandel, D. J. Hilton, & P. Catellani (Eds.), *The psychology of counterfactual thinking* (pp. 28-43). London, England: Routledge Research.
- Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford.
- Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, 56, 161-169.