Running Head: EXTERNAL VALIDITY IN MORAL PSYCHOLOGY

Revisiting External Validity:

Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology

Christopher W. Bauman

University of California, Irvine

A. Peter McGraw

University of Colorado Boulder

Daniel M. Bartels

University of Chicago

Caleb Warren

Texas A&M University

Word count = 7117 (not including the appendices)

Abstract

Sacrificial dilemmas, especially trolley problems, have rapidly become the most recognizable scientific exemplars of moral situations; they are now a familiar part of the psychological literature and are featured prominently in textbooks and the popular press. We are concerned that studies of sacrificial dilemmas may lack experimental, mundane, and psychological realism and therefore suffer from low external validity. Our apprehensions stem from three observations about trolley problems and other similar sacrificial dilemmas: (i) they are amusing rather than sobering, (ii) they are unrealistic and unrepresentative of the moral situations people encounter in the real world, and (iii) they do not elicit the same psychological processes as other moral situations. We believe it would be prudent to use more externally valid stimuli when testing descriptive theories that aim to provide comprehensive accounts of moral judgment and behavior.

Revisiting External Validity:

Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology

Research on morality has experienced a major resurgence over the past decade. A shift away from rationalist theories that dominated the literature for many years created new theoretical space, prompted new questions, and called for new empirical methods. New stimuli created for laboratory studies has spurred research activity and led to many contributions to our understanding of morality. However, we believe it is now important to revisit the methodological principle of external validity. We question whether behavioral scientists who study morality should be concerned that they have become desensitized to potential limitations of stimuli that have risen in prominence over the past several years. To the extent that researchers seek to develop general theories of morality, their study stimuli must engage the same psychological processes that operate in everyday situations (Aronson, Wilson & Brewer, 1998; Mook, 1983).

The scholarly literature on moral judgment increasingly features studies that examine people's reactions to "sacrificial dilemmas," (Bartels & Pizarro, 2011) or brief scenarios where the only way to prevent a calamity from affecting a group of people would be to harm someone else or some smaller group. The tradeoff in sacrificial dilemmas is not problematic in-and-of-itself. Researchers can learn a great deal from the way people approach tough choices that put different moral considerations in conflict. Our concern, however, is that many sacrificial dilemmas are set in fanciful, sometimes absurd contexts, and these artificial settings may affect the way people approach the situation and decide what to do. Moral psychology has developed a sophisticated understanding of how people respond to sacrificial dilemmas (Waldman, Nagel, & Wiegmann, 2012; Bartels, Bauman, Cushman, Pizarro, & McGraw, in press), but we worry that the judgment and decision making processes people use in these unusual situations may not accurately reflect moral functioning in a broader set of situations. To be clear, our focus in the current paper is on aspects of commonly used sacrificial dilemmas that make them seem frivolous and different from more realistic moral situations; we find little fault with studying moral dilemmas per se.

External validity refers to how well the results of a given study generalize and explain a range of other situations (Campbell, 1957). We contend that the results of experiments that examine people's responses to artificial sacrificial dilemmas may suffer from low external validity because artificial sacrificial dilemmas often lack experimental, mundane, and psychological realism (Aronson et al., 1998). Experimental realism is how well the situation meaningfully engages participants and causes them to take the study seriously. Mundane realism refers to how likely it is that the events in a study resemble those participants confront in their everyday lives. Psychological realism involves whether the same mental processes operate during an experiment and real-world analogues. We suspect that many—and especially the most popular—sacrificial dilemmas score relatively low on all three types of realism, which reduces the extent to which people's choices about the dilemmas can inform general theories of morality. In the absence of external validity, researchers may collectively be building a science of how people respond to a select set of stimuli that capture only a narrow and perhaps distorted view of moral phenomena rather than generating a comprehensive theory of how people make moral judgments across the full range of moral situations they encounter in their daily lives.

To illustrate our concerns about the artificial settings of many sacrificial dilemmas, we examine trolley problems. Trolley problems are the most prominent examples of sacrificial dilemmas. They have been used extensively in experiments, and they acted as the catalyst that brought sacrificial dilemmas into mainstream moral psychology. We first explain the origin of trolley problems and other sacrificial dilemmas to contrast the purposes for which they were originally conceived by philosophers with how they are currently used by psychologists. We next discuss why psychologists and philosophers generally use different methods and call attention to ways in which experiments that use artificial sacrificial dilemmas may not be externally valid. We also present three observations about trolley problems that illustrate in concrete terms why we are concerned about external validity. Finally, we conclude by calling for researchers to be mindful of external validity when choosing stimulus materials.

Before proceeding, we wish to state explicitly that we are not suggesting that researchers completely abandon all sacrificial dilemmas or disregard theories that have been strongly influenced by them. We believe that sacrificial dilemmas can be a legitimate source of data, provided that researchers (i) recognize the limitations of unrealistic stimuli and (ii) do not rely on them exclusively. However, the popularity of some sensational examples of sacrificial dilemmas appears to have drawn attention away from external validity by somehow blurring the line between rhetorical devices and scientific stimuli. Moreover, over-reliance on any one class of stimuli can lead to common method variance that can cause the observed relationship between variables to differ from their natural association across a wider range of situations (Campbell & Fiske, 1959). Therefore, we believe it is important consider the effects of our collective methodological choices.

Trolley Problems

Trolley problems have quickly become a familiar part of literature on morality in the behavioral sciences. Since 2000, at least 136 papers published in behavioral science outlets explicitly discussed trolley problems in some way, and 65 of those reported original studies that used trolley problems as experimental stimuli (see Figure 1). Research on trolley problems has not only been plentiful, it also has been highly visible. Papers on trolley problems have been published in top journals and have received attention from major media outlets. For example, Greene, Sommerville, Nystrom, Darley, & Cohen's (2001) paper in *Science* has been cited over 968 times. Additionally, the *New York Times* has run multiple features on empirical research on trolleys and other sacrificial dilemmas (e.g., Pinker, 2008; Wade, 2007), and psychology textbooks now include trolley problems in lessons on moral judgment (e.g., Myers, 2010; Schacter, Gilbert, & Wegner, 2011). Taken together, it is clear that trolley problems—and the larger class of sacrificial dilemmas of which they are the most prominent examples—are the focus of a considerable amount of scholarly activity, and they represent one means for teaching students and the general public about moral psychology. But, are they representative of the methodological rigor and sophistication that behavioral scientists typically use? A brief overview of how trolley problems and other sacrificial dilemmas became popular may help explain why concerns about their external validity have not been raised before.

Trolley Problems as Thought Experiments in Philosophy

Thought experiments are imaginary scenarios designed to explore the implications of a principle or theory. Thought experiments have been a fixture in scholarly discourse since classical antiquity and have "led to enormous changes in our thinking and to an opening up of most important new paths of inquiry" (Mach, 1897/1976, p. 138). Across a wide range of disciplines, including physics, mathematics, economics, and philosophy, thought experiments have helped scholars identify the logical implications of a set of premises (Cooper, 2005) and call attention to anomalies (Kuhn, 1964). In moral philosophy, thought experiments often are used to compare broad theoretical propositions with situation-specific moral judgments (Brower, 1993).

Trolley problems are the most well-known thought experiments in the field of ethics. Foot (1967) introduced the original version of the trolley problem as one in a series of thought experiments she designed to punctuate her argument about whether the permissibility of an action should depend on whether harmful consequences are desired by the actor or occur as a foreseen but unintended side effect (i.e., the doctrine of double effect; Aquinas, 13th century/1918; Quinn, 1989). In her version, the driver of a runaway tram must choose whether to steer from a track with five men working on it to another with one man working on it. Foot expected readers to agree "without hesitation" that it is morally acceptable for the driver to turn to the track with one worker because the one worker's death is not an essential part of the driver's plan to save the five (p. 8). However, she then undermined certainty behind this initial judgment by contrasting the tram scenario with others where it does not seem morally acceptable to intend to kill one person even if five would be saved as a result (e.g., killing and harvesting the organs of an unwilling donor to save five people who need transplants). Through these deliberately constructed examples, she illustrated why she believed that the doctrine of double effect is less important than the distinction between avoiding injury and bringing aid. In other words, Foot used the trolley problem and other thought experiments as "intuition pumps" that helped her audience understand and embrace her position (Dennett, 1984).

Thomson (1976, 1985) modified Foot's original scenario to explore the notion that people feel less obligated to do something that saves lives than avoid doing something that kills people (i.e., positive and negative duties; Rawls, 1971/1999). In the process, she created the most well known versions of the trolley problem, each of which involved actions that sacrificed one person to save five. In the "Bystander at the Switch" version, an actor could flip a switch to divert a trolley from a track with five workers onto a track with one worker. In the "Fat Man" version (aka “footbridge” version), an actor could push a fat man off a bridge to stop a trolley before it ran over five workers on the track ahead. Thomson’s trolley problems captivated scholars who, in turn, created even more variants to examine a number of other moral principles and how their applicability changes as a function of seemingly subtle differences across situations (e.g., Unger, 1996).

In sum, philosophers developed trolley problems as *rhetorical devices* that could help them articulate the implications of moral principles in concrete, albeit highly unusual, situations. Although others have criticized the use of trolley problems in philosophy (e.g., Hare, 1981; Pincoffs, 1986; Singer, 1999), our purpose is to point out the potential limitations of using such unrealistic scenarios in empirical behavioral science.

Trolley Problems as Experimental Stimuli in Psychology

The establishment of the first experimental psychology laboratory in 1879 by Wilhelm Wundt is typically considered the point when philosophy and psychology diverged into distinct disciplines (Boring, 1960). Since then, methodological differences have served as one boundary between these two fields whose areas of inquiry often overlap. The extent to which methods that are appropriate in one discipline can be successfully imported into the other can be limited because philosophers and psychologists often have different orientations, assumptions, and goals. Many moral philosophers seek to determine the right way to act in morally relevant situations (Quinton, 1995).[1] They primarily rely on logic and intuition to identify the rules or principles that one ought to follow, and they often use thought experiments, including trolley problems and other sacrificial dilemmas, to guide their views and bolster their arguments. By

[1] We recognize that this description is an oversimplification of the broad range of questions and methods that moral philosophers use. However, this statement accurately represents a large portion of moral philosophy and the type of scholarship from which trolley problems and other sacrificial dilemmas emerged.

treating thought experiments as analogues to complex moral problems, philosophers hope to illuminate contradictions, clarify otherwise conflicting intuitions, and demonstrate how to apply moral principles in logically consistent ways across contexts (Bloom, 2011, Horowitz & Massey, 1991).

The use of trolley problems and other sacrificial dilemmas in philosophical argumentation does not automatically legitimate their use in empirical investigations of psychology. Most moral psychologists seek to understand how people think, feel, and behave in moral situations, and they typically use empirical methods, especially experiments, to test their claims. Because most people are unlike philosophers in their ability and desire to achieve logical consistency across their beliefs (Converse, 1964; see also Bandura, 1999; Chugh, Bazerman, & Banaji, 2005; Tenbrunsel & Messick, 1999), even the most sophisticated normative accounts of moral principles may only partially explain how people actually interpret and confront moral situations. Also, tools designed to elucidate philosophers' principled arguments under the background assumption of complete rationality in discussions of normative ethics may not be well suited to test behavioral scientists' descriptive claims about the psychological processes that underlie moral judgment and behavior under the condition of bounded rationality that exists in everyday life (cf. Simon, 1957).

Nevertheless, the prospect of using sacrificial dilemmas in psychological research is attractive for at least three reasons. First, using common methods helps build an interdisciplinary body of knowledge that warrants attention and helps make the science relevant. Second, sacrificial dilemmas may appear to be a tidy way to examine moral phenomena in the laboratory because aspects of these scenarios can be easily modified, providing experimenters with the capacity to address a wide range of research questions. Third, the sacrificial dilemmas have helped to generate influential theories of moral judgment, such as Greene and colleagues' dual process theory and Mikhail and colleagues' moral grammar theory, which have spurred even more experimentation and methodological and theoretical innovations (see Greene, 2007; Greene et al., 2001; Mikhail, 2007, 2009; and for a broader discussion, Bartels et al., in press).

Threats to External Validity

Sacrificial dilemmas are convenient to use, and their visibility in scientific and popular publications has made them a prominent experimental paradigm in moral psychology. However, many experiments that use trolley problems and other sacrificial dilemmas as stimuli may have low external validity. In the sections that follow, we use trolley problems to illustrate the three ways that the artificial contexts of sacrificial dilemmas can threaten external validity. First, trolley problems are low in experimental realism because people find them to be humorous rather than serious. Second, trolley problems are low in mundane realism because it is hard to imagine how they could happen in real life (cf. Bennis, Medin, & Bartels, 2010a; Hare, 1981). Third, trolley problems are low in psychological realism because the implausibility of the scenario decouples moral reproach from judgments of immorality—a link that is fundamental to the way people experience moral situations and commonly observed in other research (e.g., Haidt, Rosenberg, & Hom, 2003; Skitka, Bauman, & Sargis, 2005; Tetlock, Kirstel, Elson, Green, & Lerner, 2000). Taken together, these limited levels of realism call into question how well studies of people's responses to trolley problems and other artificial sacrificial dilemmas generalize and help explain moral judgment in other, more common situations.

Experimental Realism: Finding Humor in the Death of Innocent People

Trolley problems and other sacrificial dilemmas were originally designed to be entertaining. Philosophers counted on the fantastic details of trolley problems and other sacrificial dilemmas to lighten an otherwise dense and heavy topic. In her original discussion of the trolley problem, for example, Foot (1969) argued that people may wish to believe that the lone victim may somehow escape his plight provided that, “the driver of the tram does not then leap off and brain him with a crowbar” (p. 9). She also discussed a story about spelunkers who became trapped in a cave because an obese member of their party got stuck. She suggests that some in the party might try to justify setting off dynamite near the man by arguing that, “We didn’t want to kill him... only to blow him into small pieces” (p. 7). In the closing of her paper, Foot even wrote, “The levity of the examples is not meant to offend” (p. 15). In stark contrast to the lightheartedness of many sacrificial dilemmas, however, people find most real-life situations involving the inevitable deaths of people in their presence to be quite sobering. In this sense, sacrificial dilemmas differ dramatically from the situations they are intended to exemplify.

There is no question that people sometimes find humor in dark events in the real world (Morreall, 2009). However, there are at least two reasons why humorous descriptions of tragic situations are problematic for behavioral scientists who wish to study the psychological mechanisms that typically underlie moral judgment. First, research on humor reveals that people see humorous situations as non-serious or removed from real life concerns, even though the situations may have negative underpinnings (Apter, 1982; Martin, 2007; McGraw, Williams, & Warren, 2014, Morreall, 2009). In particular, people find humor in benign violations, or situations that involve apparent transgressions that are actually permissible or safe for one reason or another, such as being ridiculous and impossible. If observers find a situation that involves the death of innocent people to be amusing, there is good reason to believe that they are at least partially disengaged from the moral issues at stake.

Second, humor may alter the decision making processes people normally use to evaluate moral situations. A large body of research shows how positivity is less motivating than negativity (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). For example, people donate less money when faced with pictures of happy children than pictures of sad children (Small & Verrochi, 2009). If humorous aspects of sacrificial dilemmas similarly disengage or disrupt psychological processes that prompt people to act compassionately toward victims, sacrificial dilemmas provide a distorted view of moral decision making. Moreover, other research indicates that people ignore or gloss over negative information in order to maintain a positive mood (Andrade, 2005; Isen & Simmonds, 1978). If people are enjoying the humorous features of sacrificial dilemmas, they may pay less attention to the core dilemma in the scenario to preserve their mood. In sum, the evidence suggests that humorous aspects of the situations may alter the way people approach the grave tradeoff researchers wish to study. Therefore, finding humor in sacrificial dilemmas may indicate not only low experimental realism, but also low psychological realism.

In our experiences, classroom presentations of trolley problems (the footbridge version, in particular) generate laughter (for a video clip see Sandel, 2009, 4:33). Students seem to enjoy talking about the “grisly” details of the causes and consequences of their choices in trolley problems, but they are noticeably less comfortable when discussing other morally relevant topics, such as child labor, drone strikes, waterboarding, and discrimination and harassment in the workplace. Keeping an audience entertained is a boon to ethics instructors and authors of scholarly papers alike, but researchers interested in testing descriptive theories of morality should examine people’s responses to typical rather than amusingly atypical moral situations.

Situations that ostensibly involve the death of innocent people but evoke laughter seem to miss the mark in terms of activating the processes that normally govern moral judgments.

To demonstrate that trolley problems elicit humor in experimental settings, we surveyed undergraduates and assessed their reactions to the footbridge and bystander scenarios. Although respondents considered pushing the man off the footbridge to be more wrong than flipping the switch, they found more humor in the footbridge than bystander scenario; 63% reported laughing at least a little at the footbridge version and 33% reported laughing at least a little at the bystander version (see Appendix A). The high incidence of laughter suggests that both scenarios, but especially the footbridge scenario, lack experimental realism. Therefore, one could question how well trolley problems, and by extension other sacrificial dilemmas, provide ideal tests of descriptive theories of morality.

Mundane Realism: Unlike Moral Choices People Might Actually Face

The humorous demise of innocent people is not the only way that trolley problems and other sacrificial dilemmas may lack realism. Our experiences in the classroom also suggest that people find aspects of trolley problems hard to believe. People often scoff at the notion that the fat man’s body could really stop a train, question whether there really is no place for workers on the track to go, and dispute whether anyone could really appraise all of the important aspects of the situations with certainty and in time to act. Also, people often claim that important information is missing and ask for further details. For example, they may want to know whether they know anyone on the footbridge or on the tracks (Bloom, 2011), whether anyone from the railroad is aware of the situation or in a position to help, why no other safety mechanisms are in place, and whether they can ignore the legal ramifications of their actions. Thus, trolley problems may lack mundane realism because people often reject the worlds trolley problems and other sacrificial dilemmas depict, even if they can get past the humorous elements (for more on this issue of "closed-world assumptions" see Bazerman & Greene, 2010; Bennis et al., 2010a, 2010b; Schwartz, 2010, and Tetlock & Mitchell, 2010).

Trolley problems also lack mundane realism because the catastrophes depicted in sacrificial dilemmas differ considerably from the type and scale of moral situations people typically face in real life. To illustrate this point, we measured how realistic our participants found trolley problems compared to short scenarios about contemporary social issues (viz. abortion, gay marriage; see Appendix B). As another point of comparison, we also measured the perceived realism of Kohlberg's (1981) Heinz dilemma because it played such a prominent role in theory and research on moral psychology before the emergence of trolley problems. People rated the trolley problems to be much less realistic than the short scenarios about contemporary social issues. The Heinz dilemma fell in between trolley problems and contemporary issues, but participants rated it as substantially more realistic than the trolley problems. Therefore, using trolley problems in empirical research represents a significant step *backward*, in external validity, from what used to be the prototypical moral situation. To be clear, we are not suggesting that the field go back to using Heinz or that the scenarios we created based on contemporary moral social issues are ideal stimuli. We merely use these scenarios as a reference points to demonstrate that trolley problems are much less realistic than other scenarios that one could easily create and use in research.

There are instances when researchers may be justified in selecting experimental contexts that do not mimic reality (Mook, 1983). In this particular case, however, it is unclear how or why it is better to test of theories of moral using sacrificial dilemmas than more commonly encountered situations.[2] Few participants in psychology experiments have direct experience making quick decisions that determine who will live and who will die, and few would even expect to face anything even remotely similar. Although researchers commonly assume that the psychology of trolley problems parallels real world decision making, there is little or no evidence that suggests that it is advantageous (and therefore necessary) to examine scenarios with so little in common with situations that participants more typically face. In short, mundane realism may not be absolutely necessary for every study in a research program, but it is a desirable quality for experiments to have. In the absence of explicit reasons to abandon it, researchers' default choice should be maintain mundane realism.

Before moving on, we note that a few researchers have assessed how skeptical their participants were about trolley problems. Greene, Cushman, Stewart, Lowenberg, Nystrom, & Cohen (2009), for example, asked participants to indicate a reluctance to go along with scenario, and they did not analyze data from participants who did so. About 5% of participants in one study and 12% in another circled the item: *I did not find the description from the preceding pages to be realistic, and my answers reflect my inability to take seriously the description that was given.* We believe this is a reasonable approach to data collection and better than doing nothing at all. However, the success of the approach depends on whether participants are willing and able to identify the drivers of their judgments, and so these rates of endorsement could

[2] One might argue that sacrificial dilemmas are useful as novel situations that elicit unrehearsed responses from participants, but recent data indicates that many participants are familiar with them. We presented the bystander and footbridge versions of the trolley problem to psychology undergraduates at a large university (N = 70). No researchers using the subject pool had used trolley problems previously, and course instructors had not discussed them. However, 64% recognized the switch problem and 39% recognized the footbridge problem. We found similar levels of familiarity in a marketing subject pool at the same university (N = 84; 64% switch, 52% footbridge). Thirty percent of MTurk workers are familiar with the switch problem, and familiarity increased dramatically as a function of MTurk activity; 68% of MTurk workers in the 90-98th percentile of activity and 85% of MTurk workers in the 99th percentile of activity were familiar with the switch problem (Chandler, Mueller, & Paolacci, 2013). In short, trolley problems are no longer novel to participants.

under-represent the problem. Additionally, participants may be reluctant to tell researchers that the study is absurd because demand characteristics prompt participants to behave in socially desirable ways (Orne, 1959; Weber & Cook, 1972). It therefore seems preferable to use more believable scenarios rather than depend on the accuracy of measures with known limitations.

Screening out skeptical participants in not the only way that researchers have tried to address problems associated with perceived realism. Greene et al. (2009) also asked participants to rate the likelihood that the actions taken in the scenario (e.g., pushing a man off the footbridge, flipping the switch) would actually produce the outcome described and used these data as covariates in analyses. In two studies, participants' disbelief about the events described in scenario had significant effects on choice, even though data from participants who indicated that they were unable to view the scenario realistically had already been removed from analyses. In other words, disbelief influenced choice even among participants who passed the believability check. Therefore, these studies provide empirical evidence that (i) screening out skeptical participants is insufficient to eliminate problems, and (ii) a lack of perceived realism is a legitimate threat to the validity of studies that use sacrificial dilemmas. Of course, one could try to measure and statistically control for disbelief. As with screening out skeptical participants, however, it seems more straightforward to use more believable scenarios.

Taken together, Greene et al.'s (2009) efforts to address problems associated with realism illustrate our concerns. One the one hand, researchers can reduce the influence of unrealistic stimuli on results if they screen participants and statistically control for a perceived lack of realism. On the other hand, few researchers use these practices; the vast majority of trolley studies report no such controls. Further, it is difficult to know whether these controls are sufficiently precise to completely remove the error in analyses of covariance. We believe that enhancing realism would be a more effective tool for addressing these concerns.

Psychological Realism: Immoral But Not Unwelcome

Although experimental and mundane realism are desirable qualities of many studies (Aronson et al., 1998), they may not be essential when researchers are interested in assessing whether something *can* happen rather than how frequently something *does* happen (Mook, 1983). Researchers often conduct experiments to test the capacity of a theory to predict what happens under specific (even artificial) conditions in the laboratory. In these cases, results provide evidence about whether the theory is correct, and it is the theory—not the laboratory setting—that generalizes to the real world. As a result, it is not always crucial that an experimental context resemble the real world, but even Mook explicitly endorsed the necessity of psychological realism in his landmark paper on why external invalidity may not be problematic. In other words, there is consensus that the validity of any study necessarily depends on the extent to which the research setting engages the processes of interest, irrespective of whether researchers are interested in whether something *can* or *does* happen (see also Aronson et al., 1998). Therefore, a lack of psychological realism could be the most important threat to sacrificial dilemmas.

Sacrificial dilemmas may fit definitional criteria for what a moral situation is (e.g., they involve harm), but they may not activate the same psychological processes as more realistic moral situations. That is, sacrificial dilemmas may be poor models of moral situations. Of course, scientists and engineers often use simplified models to gain traction toward understanding complex phenomena. In genetics, for example, researchers often study fruit flies rather than humans because DNA functions basically the same way in all organisms. However, simplified models only are useful if crucial elements of the model extend to the more complex phenomena that they purportedly represent. If DNA functioned differently across species, experiments conducted on fruit flies would be less able to inform researchers' understanding of human genetics. Put differently, it makes sense to use fruit flies to study genetics because fruit flies have human-like DNA, but it would not make sense to use fruit flies to study osteology because insects' exoskeletons are too different from human bones. For the same reason, the results of studies of sacrificial dilemmas may have only limited bearing on more general theories of morality. If some psychological processes differ across sacrificial dilemmas and other moral situations, even the most sophisticated accounts of how people make decisions about sacrificial dilemmas may not generalize and help to explain the way people usually make moral judgments outside of the laboratory.

One potential indication that humorous and unrealistic sacrificial dilemmas engage different psychological processes than other moral situations is that they are socially inconsequential. One characteristic feature of morality is that it inherently motivates and justifies responses to perceived violations (Bauman & Skitka, 2009; Skitka et al., 2005; cf. Hume, 1888/1739-1740). In other words, people do not sit idly and watch moral transgressions and transgressors; they are moved by them. People express outrage, report a strong sense of contempt and disgust, and fear moral contagion when they witness or even contemplate wrongdoing (Haidt, Rozin, McCauley, & Imada, 1997; Rozin, Haidt, & McCauley, 2008; Tetlock et al., 2000). They also distance themselves from morally dissimilar others; they are uncomfortable in relationships (e.g., close friends and romantic partners, but also coworkers and neighbors) with people who disagree with their moral beliefs (Haidt et al., 2003; Skitka et al., 2005; Wright, Cullum, & Schwab, 2008). In our classroom experiences with trolley problems, however, no one ostracizes or even seriously reprimands those who would push the fat man or refuse to flip the switch. If anything, our students approach the “reprobates” with curiosity, and there is a distinct absence of repulsion. Therefore, trolley problems and other sacrificial dilemmas may lack psychological realism because people act differently to people who make deviant choices in these settings than to those who break moral rules in other situations.

We conducted a demonstration to show that discomfort with morally dissimilar others disappears when people think situations are unrealistic (see Appendix C). We found that whether trolley problems had social consequences like other moral situations depended on the perceived realism of the scenario. The majority of the sample considered the scenario relatively unrealistic, and for these participants there was not a significant relationship between moral conviction and discomfort with people who disagreed with their moral judgments about the footbridge and bystander scenarios. Therefore, on average, trolley problems failed to elicit a response that has been documented by multiple other moral contexts (Haidt et al., 2003; Skitka et al., 2005; Wright et al., 2008), and researchers would not have detected an effect of moral conviction on moral distancing if prior research on moral distancing had been conducted using trolley problems. Importantly, the minority of our sample who considered the scenarios more realistic showed a significant relationship between moral conviction and discomfort with morally dissimilar others. Therefore, there does not appear to be anything about moral dilemmas per se (e.g., a difficult choice between two undesirable outcomes) that promotes tolerance of moral disagreement. Instead, tolerance emerges only when people perceive the situation to be unrealistic.

Given that trolley problems and other sacrificial dilemmas differ from other moral situations in their propensity to activate social distancing processes, they also seem likely to differ in terms of whether and how they engage other important features of moral judgment as well. Researchers who rely heavily on trolley problems and other sacrificial dilemmas may therefore miss or distort aspects of morality that normally operate under other circumstances. Moreover, it is also possible that trolley problems and other sacrificial dilemmas activate or accentuate other psychological processes that do not typically play important roles in other moral situations. Therefore, our concern about psychological realism reflects the possibility that trolley problems and other sacrificial dilemmas understate the role of some psychological processes and overstate the role of others. As a field, we cannot discern how these differences might affect theory generation without a corpus of data from a variety of stimuli, which is precisely why we encourage researchers to use experimental stimuli that are both more realistic and more diverse.

Discussion

Sacrificial dilemmas are intrinsically engaging situations that people enjoy pondering. These scenarios have generated interest in morality from scholars and the general public alike. However, we believe that trolley problems and other similar sacrificial dilemmas have low external validity. Our discussion has focused on three points: (i) trolley problems are amusing rather than sobering, (ii) trolley problems are unrealistic and unlike anything people encounter in the real world, and (iii) trolley problems do not engage the same psychological processes as other moral situations. By extension, we worry that many sacrificial dilemmas set in similarly artificial settings may exhibit similar characteristics. Therefore, we are concerned that examining people's responses to sacrificial dilemmas may provide only a partial view of how people tend to confront moral situations in their everyday lives.

Prior descriptive research that has used sacrificial dilemmas has been generative, but we believe that the field would benefit from exploring alternatives. We caution the field about continuing to develop a science of how people respond to contrived situations that may capture only some aspects of moral judgment and decision making and distort the way some psychological processes operate. Considering the dramatic rise in the number of empirical investigations that use sacrificial dilemmas, we fear that some researchers have begun to rely on them in their research because others have used them rather than because they are the best way to address their particular research question. We also worry that reviewers may now be less likely to scrutinize the methodological merit of studies that use even highly artificial sacrificial dilemmas because papers that feature them have made it through the review process in the past.

Moving Forward

We have argued that trolley problems and similarly unrealistic sacrificial dilemmas are problematic, but we have not yet provided alternatives to these scenarios. Before doing so, we wish to reiterate that we do *not* advocate abandoning moral dilemmas or scenarios entirely. There is nothing intrinsically problematic with scenarios that pose tradeoffs between active and passive harm, indirect and direct harm, or means versus ends. In fact, we think that it is important to understand how people reason, make choices, and act in situations that involve moral tradeoffs. We believe that carefully constructed scenarios can play a crucial role in helping to elucidate some of the contours of moral cognition. That said, we think that using one or a small number of hypothetical scenarios that require participants to imagine highly improbable and implausible events is problematic for many research questions. Our concern is that participants posed with such scenarios can easily get caught up in the fantastical details or reject the assumptions of these scenarios entirely, which may obscure the psychological processes the researchers intend to study. In short, there is nothing wrong with the structure of sacrificial dilemmas (i.e., the theoretical “meat” of the scenarios), but the semantics of these dilemmas can undermine their utility as tools for testing descriptive psychological theories.

To mitigate our concerns, researchers could create new scenarios involve the same kinds of tradeoffs as sacrificial dilemmas, but present them in ways that are more consistent with how people might face those tradeoffs in the real world. For example, engineers and designers routinely evaluate product safety for potential risk and likely benefits, managers sometimes have to decide which employees to retain and which to layoff, medical professionals and administrators must make decisions about the allocation of scarce medical resources, and people charged with managing natural resources sometimes have to distribute harm to one or more animal or plant populations to mitigate a risk to another species. Many examples like these can be found in the literature (e.g., Cushman, Knobe, & Sinnott-Armstrong, 2008; Reynolds, 2006; Ritov & Baron, 1999; Tetlock et al., 2000; Bartels & Medin, 2007). In short, it should not be difficult to use scenarios that have better external validity than sacrificial dilemmas.

We recognize that adding in real-world context could cause participants to react to a particular incidental detail of a specific context, but this risk is no greater in studies of moral judgment than any other topic. Also, minimalism does not necessarily eliminate all potential problems associated with idiosyncrasies in stimuli. “Bare bones” scenarios that provide very limited contextual information can fail to elicit the same responses as more contextually rich scenarios (FeldmanHall, Mobbs, Evans, Hiscox, Navrady & Dalgleish, 2012). Presenting participants with more than one stimulus—and different types of stimuli (i.e., not just a battery of sacrificial dilemmas)—is the best way to test the robustness and generalizability of a response tendency.

Sampling Different Contexts

Relying on a single set of stimuli limits external validity. If the majority of studies on a particular question all use stimuli that share some characteristics, then it can be difficult to determine how well common features of the stimuli are affecting the results (Campbell & Fiske, 1959; Wells & Windschitl, 1999). Put differently, we cannot know how well the results of studies conducted on one particular class of stimuli generalize until we have data from studies that sample a wide range of stimuli (see McGuire, Langdon, Coltheart, & Mackenzie, 2009; Ugazio, Lamm, & Singer, 2012). Of course, the literature on moral judgment and decision making now includes evidence from multiple sacrificial dilemmas. If each has low external validity, however, gathering data from several sacrificial dilemmas cannot address generalizability as well as examining at least some realistic dilemmas and testing for differences across stimuli (after ensuring that there is sufficient statistical power to detect them).

Over-reliance on a particular set of stimuli is not unprecedented, and progress can be inhibited by the very same stimuli that initially captured people's attention and fueled advances in research. In social psychology, for example, heavy use of the attitude attribution paradigm (i.e., pro-/anti-Castro essays; Jones & Harris, 1967) in the 1970's led to many incremental publications but only limited progress toward understanding the causes or breadth of correspondence bias (Gilbert & Malone, 1995). In decision-making research, studies of simple, monetary gambles have been the norm for decades. However, recent research shows that these studies have limited generalizability because people reason differently about gambles that feature monetary and non-monetary outcomes (McGraw, Shafir, & Todorov, 2010; Rottenstreich & Hsee, 2001). Similarly, in early research on memory, Ebbinghaus (1885/1913) used tasks where people memorized nonsense syllables (e.g., "fip", "jid") to examine, among other things, how memory decays over time. It is now understood, however, that the use of meaningless syllables prevented research from understanding the importance of context and content on memory processes (Baddeley, Eysenck, & Anderson, 2009). In sum, using stimuli that represent only one out of many possible contexts can produce blind-spots in theories; the weights people attach to features of situations and the choice strategies they use are heavily content- and context-dependent (Goldstein & Weber, 1995, Rettinger & Hastie, 2001). Therefore, moral psychology should test theories in a wide range of moral contexts and avoid overrepresentation of any one scenario.

Given the complexity of moral cognition and behavior in the real world, we suggest that the best research solutions will rely on a combination of scenarios, behavioral laboratory studies, and work conducted outside of the lab. Although currently underused, field work, such as Alan Fiske's (1991) observations of relational differences between Americans and Africans or Ginges, Atran, and colleagues' studies on sacred values in suicide/martyrdom attacks and peace negotiations (Ginges & Atran, 2009; Ginges, Atran, Sachedeva, & Medin, 2011) complement vignette-based and laboratory work and provide alternative means for testing theories (for a discussion of further benefits of moving beyond the laboratory, see Bartels, Bauman, Cushman, Pizarro, & McGraw, in press)

A Final Thought about the Impact of Methods on Theory

This is not the first time that one set of methods has reached such prominence in moral psychology. Kohlberg's (1981) Heinz dilemma was a mainstay of empirical investigations of morality for almost two decades. Kohlberg's theory focused on moral reasoning, and he designed the Heinz dilemma so that he could examine people's explanations for their judgments. As psychologists began to question rationalism and embrace intuitionism, they became more skeptical of whether reasons drive judgment or whether there are bi-directional effects, whereby reasons can be post-hoc justifications of intuitive judgments (e.g., Cushman, Young, & Greene, 2010; Haidt, 2001; Paxton, Ungar, & Greene, 2012). Correspondingly, studies began to focus more on moral judgments and less on the reasons people provide for them. When designing experiments that would help identify the psychological processes that underlie moral judgment, researchers shifted away from studying relatively “complex” and “messy” situations like the Heinz dilemma to studying "simpler" and "cleaner" (but less realistic) situations depicted in sacrificial dilemmas (cf., Monin, Pizarro, & Beer, 2005). Much as the Heinz dilemma caused researchers to miss aspects of moral judgment that did not stem from deliberate reasoning, sacrificial dilemmas may cause researchers to miss aspects of moral judgment that function differently when people perceive situations.

A science of morality that over-relies on any one paradigmatic set of experimental materials stands a chance of misunderstanding the fundamental processes that operate in everyday morality. This issue, of course, is not endemic to morality—it is a major problem to any area of inquiry. In the end, the confidence we have in our outcomes rests on the rigor of our process. If we are to build strong empirical science in psychology, we must always be willing to self-consciously reassess our methods.

References

Andrade, E.B. (2005). Behavioral consequences of affect: Combining evaluative and regulatory mechanisms, *Journal of Consumer Research, 32*, 355-362.

Apter, M. J. (1982). *The experience of motivation: The theory of psychological reversals.* London, UK: Academic Press.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In G. Lindsay & E. Aronson (Eds.), *The handbook of social psychology* (4th ed., Vol.1, pp. 99-142). Boston, MA: McGraw-Hill.

Aquinas, T. (1918). Of killing. In *Summa Theologica*, Part II-II, Q. 64, A. 7. (Fathers of the English Dominican Province Trans.). New York, NY: Benziger Brothers (Original work published 13th century).

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York, NY: Psychology Press.

Bandura, A. (1999). Moral disengagement in the perpetuation of humanities. *Personality and Social Psychology Review, 3*, 193–209.

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (in press). Moral judgment and decision making. In G. Keren & G. Wu (Eds.) *Blackwell Reader of Judgment and Decision Making*. Malden, MA: Blackwell.

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices?. *Psychological Science*, *18*, 24-28.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*, 154-161.

Bauman, C. W., & Skitka, L. J. (2009). In the mind of the perceiver: Psychological implications of moral conviction. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Vol. Eds.), *Psychology of Learning and Motivation, Vol. 50. Moral Judgment and Decision Making* (pp. 339-362). Burlington, MA: Academic Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad Is Stronger Than Good, *Review of General Psychology*, 5, 323-70.

Bazerman, M.H. & Greene, J.D. (2010). In favor of clear thinking: Incorporating moral rules into a wise cost-benefit analysis—Commentary on Bennis, Medin, & Bartels (2010). *Perspectives on Psychological Science*, *5*, 209-212.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010a). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*, 187-202.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010b). Perspectives on the ecology of decision modes: Reply to comments. *Perspectives on Psychological Science, 5*, 213-215.

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review, 99*, 26-43.

Boring, E. G., (1960). *A History of Experimental Psychology* (2nd ed). Englewood-Cliffs: Prentice Hall.

Brower, B. W. (1993). Dispositional ethical realism. *Ethics*, *103*, 221-249.

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods*, 1-19. doi: 10.3758/s13428-013-0365-7

Chugh, D., Bazerman, M. H., & Banaji, M. R. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. In D. A. Moore, D. M. Cain, G. F. Loewenstein, & M. H. Bazerman (Eds.), *Conflicts of interest: Problems and solutions from law, medicine and organizational settings*. London, UK: Cambridge University Press.

Converse, P. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206-261). New York, NY: Free Press.

Cooper, R. (2005). Thought experiments. *Metaphilosophy, 36*, 328-347.

Cushman, F.A., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, *108*, 281-289.

Cushman, F.A., L. Young & J. Greene (2010) “Our multi-system moral psychology: Towards a consensus view” in *The Oxford Handbook of Moral Psychology*, ed. John Doris et al, Oxford University Press

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Ebbinghaus, H. (1919). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Teachers College Columbia University (Original work published 1885).

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navardy, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*, 434-441.

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. New York: Free Press.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect in virtues and vices. *Oxford Review, 5,* 5-15.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117,* 21-38.

Ginges, J. & Atran, S. (2009). Instrumental reasoning over sacred values: An Indonesian case study. In D.M. Bartels, C.W. Bauman, L.J. Skitka, & D.L. Medin, Eds. (2009). *Moral Judgment and Decision Making: The Psychology of Learning and Motivation, Vol. 50* (pp. 193-206). San Diego: Elsevier.

Ginges, J., Atran, S., Sachdeva, S., & Medin, D. (2011). Psychology out of the laboratory: The challenge of violent extremism. *American Psychologist*, *66*(6), 507.

Goldstein, W. M., & Weber, E. U. (1995). Content and discontent: Indications and implications of domain specificity in preferential decision making. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *The Psychology of Learning and Motivation, Vol. 32. Decision Making from a Cognitive Perspective* (pp. 83-136). San Diego, CA: Academic Press.

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.). Moral psychology, Vol. 3: The neuroscience of morality: Emotion, disease, and development (pp. 35-117). Cambridge, MA: MIT Press.

Greene, J. D., Cushman, F. A., Stewart. L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*, 364-371.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814-834.

Haidt, J., Rosenberg, E., & Hom, H. (2003). Differentiating diversities: Moral diversity is not like other kinds. *Journal of Applied Social Psychology, 33*, 1-36.

Haidt, J., Rozin, P., McCauley, C., & Imada, S . (1997). Body, psyche, and culture: The relationship of disgust to morality. *Psychology and Developing Societies, 9*, 107-131.

Hare, R. M. (1981). *Moral thinking*. Oxford, UK: Oxford University Press.

Horowitz, T., & Massey, G. (Eds.). (1991). *Thought experiments in science and philosophy*. Savage, MD: Rowman and Littlefield.

Hume, D. (1888). *A treatise on human nature*. (D. F. Norton and M. J. Norton, Trans.). Oxford, UK: Oxford University Press (Original work published 1739-1740).

Isen, A. M., & Simmonds S. F. (1978). The effect of feeling good on a helping task that is incompatible with good mood, *Social Psychology, 41*, 346–349.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3,* 1-24.

Kohlberg, L. (1981). *Essays on moral development, Voume. I: The philosophy of moral development*. San Francisco, CA: Harper & Row.

Kuhn, T. (1964). A function for thought experiments. Reprinted in I. Hacking (Ed.) *Scientific Revolutions* (1981), pp. 6-27. Oxford, UK: Oxford University Press.

Mach, E. (1976). *Knowledge and error: Sketches on the psychology of enquiry* (T. J. McCormack & P. Foulkes, Trans.). Dordrecht, Netherlands: Reidel (Original work published 1897).

Martin, R. A. (2007). *The psychology of humor: An integrative approach*. Burlington, MA: Elsevier Academic Press.

McGraw, A. P., Shafir, E., & Todorov, A. (2010). Valuing money and things: Why a $20 item can be worth more and less than $20. *Management Science, 56*, 816-830.

McGraw, A.P., Williams, L.T., & Warren, C. (2014). The rise and fall of humor: Psychological distance modulates humorous responses to tragedy. *Social Psychology and Personality Science, 5,* 566-572.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology, 45,* 577-580.

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

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. In D.M. Bartels, C.W. Bauman, L.J. Skitka, & D.L. Medin (Eds.) "*Moral Judgment and Decision Making: The Psychology of Learning and Motivation, Vol. 50*," San Diego: Elsevier

Monin, B., Pizarro, D. A., & Beer, J. A. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology, 11,* 99-111.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-388.

Moretto, G., Ladavas, E., Mattioli, F., & di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience, 22*, 1888-1899.

Morreall, J. (2009). *Comic relief: A comprehensive philosophy of humor.* Malden, MA: Wiley-Blackwell.

Myers, D. G. (2010). *Social Psychology* (10th ed.). New York, NY: McGraw-Hill.

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem." *Emotion, 12*, 364-370.

Orne, M. T. (1959). The nature of hypnosis: Artifact and essence. *Journal of Abnormal and Social Psychology, 58*, 277-299.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163-177.

Pincoffs, E. L. (1986). *Quandaries and virtues: Against reductivism in ethics*. Lawrence, Kansas: University of Kansas.

Pinker, S. (2008, January 13). The moral instinct. *New York Times*. Retrieved from http://www.nytimes.com/2008/01/13/magazine/13Psychology-t.html.

Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy & Public Affairs, 18*, 334–351.

Quinton, A. (1995) Definition of philosophy. In T. Honderich (Ed.). *The Oxford companion to philosophy* (p. 666). New York, NY: Oxford University Press.

Rawls, J. (1999). *A Theory of Justice* (Revised edition). Cambridge, MA: Harvard University Press (Original work published 1971).

Rettinger, D. A., & Hastie, R. (2001). Content effects on decision making. *Organizational behavior and human decision processes, 85*, 336-359.

Reynolds, S. J. (2006). Moral awareness and ethical predispositions: Investigating the role of individual differences in the recognition of moral issues. *Journal of Applied Psychology, 91*, 223-243.

Ritov, L. & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes, 79*, 79-94.

Rottenstreich, Y. & Hsee, C. K. (2001). Money, kisses, and electric shocks: An affective psychology of risk. *Psychological Science, 12*, 185-190.

Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotions, 3rd ed.* (pp. 757-776). New York, NY: Guilford Press.

Rozin, P., & Royzman, E. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review, 5*, 296-320.

Sandel, M. (2009). *Justice: What's the right thing to do? Episode 01. The moral side of murder*. Retrieved from http://youtu.be/kBdfcR-8hEY.

Schacter, D. S., Gilbert, D. T., & Wegner, D. M. (2011). *Psychology* (2nd ed.). New York, NY: Worth.

Schwartz, B. (2010). The limits of cost-benefit calculation: Commentary on Bennis, Medin, & Bartels (2010). *Perspectives on Psychological Science, 5*, 203-205.

Simon, H. A. (1957). *Models of Man*. New York: Wiley.

Singer, P. (1999). Living high and letting die. *Philosophy and Phenomenological Research, 59*, 183-187.

Skitka, L. J., Bauman, C. W., & Sargis, E. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*, 895-917.

Small, D. A., & Verrochi, N. M. (2009). The face of need: Facial emotion expression on charity advertisements. *Journal of Marketing Research, 46*, 777-787.

Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly, 44*, 684–707.

Tetlock, P. E., Kirstel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*, 853-870.

Tetlock, P.E., & Mitchell, G. (2010). Situated social identities constrain morally defensible choices: Commentary on Bennis, Medin, & Bartels (2010). *Perspectives on Psychological Science, 5*, 206-208.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *Monist, 59*, 204-217.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion, 12*, 579-590.

Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. New York, NY: Oxford University Press.

Wade, N. (2007, September 18). Is ‘do unto others’ written into our genes? *New York Times*. Retrieved from http://www.nytimes.com/2007/09/18/science/18mora.html.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364-389). New York, NY: Oxford University Press.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77*, 273-295.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115-1125.

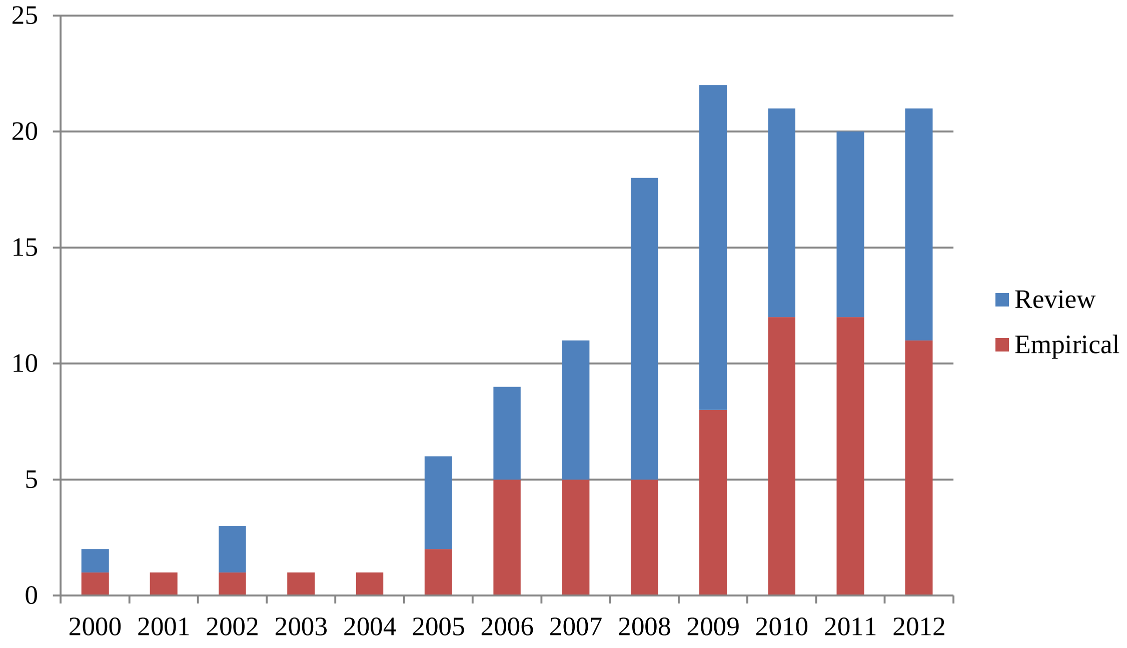
Wright, J. C., Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: Implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality and Social Psychology Bulletin, 34*, 1461–1476.

Author Notes

We thank Geoff Goodwin, David Tannenbaum, Shaun Nichols, and the Moral Research Lab (MoRL) for thoughtful comments and suggestions, and we thank Elizabeth Follmer, Bridget Leonard, Christina Kan, and the Emotion, Decision, Judgment, and Intuition (EDJI) Lab for helpful research assistance.

Figure 1

Number of published papers that explicitly discuss trolley problems



Note: When reviewing the literature, we focused mainly on books and journals within psychology. We included publications from related disciplines only if they heavily cited psychology theories and research and addressed questions that fit within the purview of psychology. We excluded papers focused on normative ethics.

Appendix A

Eighty-four undergraduate psychology students at a large, public university read either the bystander or footbridge version of the trolley problem. We used the versions Greene et al. (2001) presented in the text of their paper (not the stimuli they used in their studies) because they are fairly standard descriptions of the scenarios and similar to what appears in many empirical studies of trolley problems:

> A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Should you turn the trolley in order to save five people at the expense of one?
>
> A trolley threatens to kill five people. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. The only way to save the five people is to push this stranger off the bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Should you save the five others by pushing this stranger to his death?

We asked participants, “Is it wrong to flip the switch [push the fat man]?” and three questions that assessed humor: “Is this funny?” “Is this amusing?” and “Did this make you laugh?.” Participants could either indicate “0: No” or answer affirmatively on a five-point scale with point labels that ranged from “1: A little” and “5: A lot.” We averaged responses to the three humor items ($\alpha = .84$). Consistent with prior research, students considered pushing the man off the footbridge to be more wrong ($M = 3.3$, $SD = 1.5$) than flipping the switch ($M = 1.4$, $SD = 1.5$), $F(1, 82) = 34.72$, $p < .001$. However, participants also perceived more humor in the footbridge ($M = 2.1$, $SD = 1.4$) than the bystander scenario ($M = 1.0$, SD = 1.2), $F(1, 82) = 7.13$, p < .01. Also, 63% of participants reported that the footbridge was at least a little humorous, and 33% reported finding at least a little humor in the bystander version, χ^2 (1, 84) = 8.13, $p < .01$.

Appendix B

Two hundred twenty three people from Amazon's MTurk website read and responded to five scenarios presented in a counterbalanced order. After reading each scenario, participants indicated whether the actor in the scenario should perform the action described using a four point scale with point labels of *definitely no, probably no*, *probably yes,* and *definitely yes*. Participants then rated how much they agreed or disagreed with four statements designed to assess mundane realism: "This scenario is realistic," "This scenario is similar to choices people make in real life," "It's easy to imagine being in a situation like this," and "This would never happen in real life" (reverse scored). Participants responded using 7-point scales with point labels that ranged from *strongly support* to *strongly oppose*. We averaged scores for analyses (α = .77 - .79 across scenarios)

	Scenario	*M*	*SD*
Footbridge	A runaway trolley is hurtling down the tracks toward five people who are unable to escape. You are standing on a footbridge over the tracks next to a large man. You can push the man onto the tracks to stop the trolley and save the lives of the five people, but the man will be killed. Should you push the man?	3.5^{a}	1.6
Bystander	A runaway trolley is hurtling down the tracks toward five people who are unable to escape. You can flip a switch to divert the trolley onto a side track and save the lives of the five people, but one person on the side track will be killed. Should you flip the switch?	3.9^{b}	1.6
Heinz	A woman was near death from a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town had recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to make. He paid $200 for the radium and charged $2,000 for a small dose of the drug. The sick woman's husband went to everyone he knew to borrow the money, but he could only get together about $ 1,000 which is half of what it cost. He told the druggist that his wife was dying and asked him to sell it cheaper or let him pay later. But the druggist said: "No, I discovered the drug and I'm going to make money from it." Should the husband break into the man's store to steal the drug?	4.8^{c}	1.5
Gay Marriage	Two gay men have been in a romantic relationship for a several years. They are committed to each other and plan to be together for the rest of their lives. Should they be able to be legally married, if that is what they want?	6.4^{d}	1.1
Abortion	A 19 year old woman is in the first trimester of pregnancy and is uncertain about what to do. She is unmarried and cannot support herself financially. Should she be able to terminate the pregnancy by having an abortion, if that is her decision?	6.5^{d}	1.0

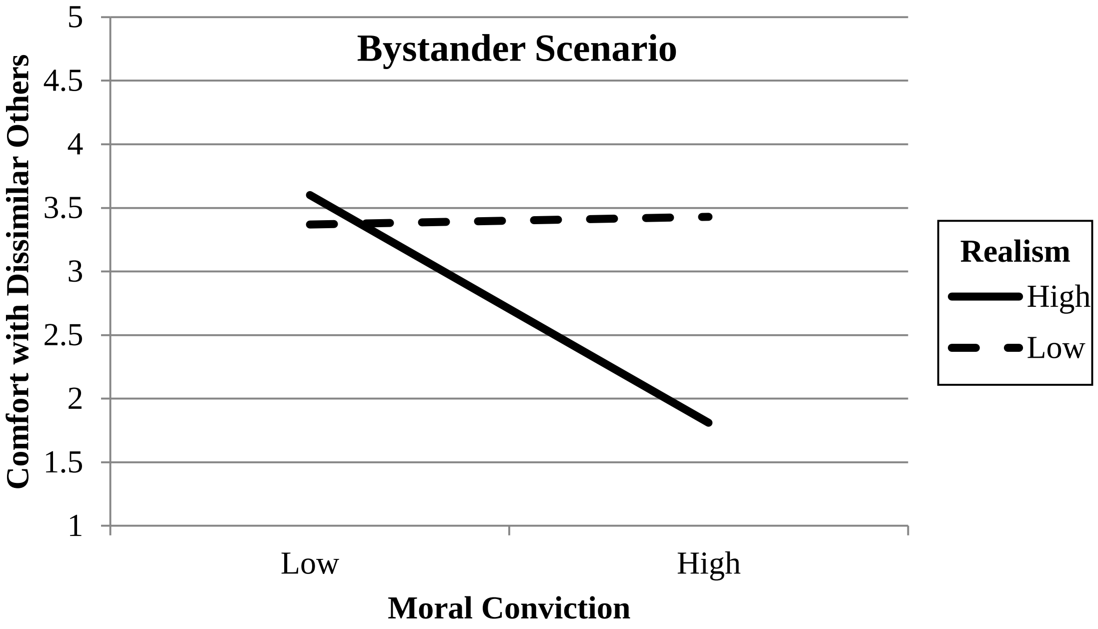
Note: Superscripts denote means that differ at $p < .05$.

Appendix C

Forty-eight people from Amazon's MTurk website read the bystander and footbridge scenarios in a random order (see Appendix B for the exact wording of the scenarios). After reading each scenario, participants indicated whether the actor in the scenario should perform the action described using a 7-point scale with point labels that ranged from *strongly support* to *strongly oppose*. We then asked participants, "How much does your choice about whether to flip the switch [push the man] relate to your personal moral convictions and core moral values?" Participants responded on 5-point scales with point labels that ranged from *not at all* to *very much*. Then, participants completed a 10 item social distance scale (Skitka et al., 2005). Specifically, we asked, "How comfortable or uncomfortable you would be with having the following relationships with someone who has the opposite opinion about this situation?," and we presented the following relationships in a random order: President of the U.S., Governor of your state, coworker, roommate, marry into your family, someone you would personally date, your personal physician, a close personal friend, teacher of your children, and your spiritual advisor (α = .94 for the bystander; α = .95 for the footbridge). Participants responded on 5-point scales with endpoint labels of *uncomfortable* to *comfortable*. Finally, participants rated how much they agreed or disagreed with five statements designed to assess realism (α = .74 for the bystander; α = .70 for the footbridge): "This scenario is funny," "This scenario is humorous," "This scenario is realistic," "This scenario is similar to choices people make in real life," and "It's easy to imagine being in a situation like this". All measures were centered before use in analysis (Aiken & West, 1991).

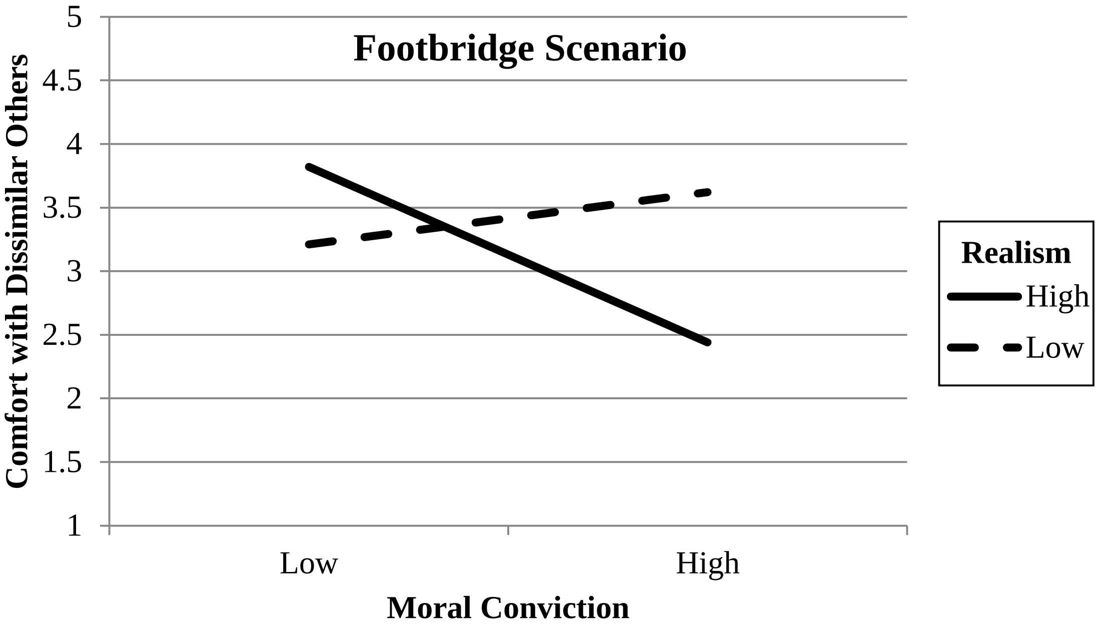
Bystander

OLS regression analysis tested the effect of moral conviction and realism on social distance for the bystander scenario. Neither the effect of moral conviction, β = -.16, *t*(45) = -1.24, *p* = .220 nor the effect of realism were significant, β = 1.01, *t*(45) = 1.64, *p* = .108. However, the interaction of moral conviction and realism was significant, β = -.32, *t*(45) = -2.02, *p* = .049. Analyses of simple effects explored the effect of moral conviction on comfort at high (+1 SD), mean, and low (-1 SD) levels of realism. When realism was high, the simple slope of moral conviction was significant, β = -.42, *t*(45) = -2.20, *p* = .033. As moral conviction increased, comfort with dissimilar others decreased. However, the simple slope of moral conviction was not significant at mean levels of realism, β = -.16, *t*(45) = -1.24, *p* = .221, or when realism was low, β = .09, *t*(45) = 0.50, *p* = .620.



Footbridge

OLS regression analysis tested the effect of moral conviction and realism on social distance for the footbridge scenario. The effect of moral conviction was not significant, β = -.18, *t*(45) = -1.61, *p* = .113, but the effect of realism was significant, β = 1.29, *t*(45) = 2.54, *p* = .014. Additionally, the interaction of moral conviction and realism was significant, β = -.42, *t*(45) = -3.16, *p* = .003. Analyses of simple effects explored the effect of moral conviction on comfort at high (+1 SD), mean, and low (-1 SD) levels of realism. When realism was high, the simple slope of moral conviction was significant, β = -.53, *t*(45) = -3.37, *p* = .002. As moral conviction increased, comfort with dissimilar others decreased. However, the simple slope of moral conviction was not significant at mean levels of realism, β = -.18, *t*(45) = -1.61, *p* = .114, or when realism was low, β = .17, *t*(45) = 1.05, *p* = .298.



Taken together, the results show that perceived realism moderates the effect of moral conviction on social distance. At mean levels of perceived realism and aggregated across levels of perceived realism, the association between moral conviction and moral distancing is not significant. Therefore, trolley problems appear to lack psychological realism.