UNIVERSITY OF CALIFORNIA SAN DIEGO

Harming, Helping, and Blaming: An Inquiry Into Mechanisms Of Moral Cognition

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Experimental Psychology

by

Arseny Alexey Ryazanov

Committee in charge:

Professor Piotr Winkielman, Chair
Professor Karen Dobkins
Professor Craig McKenzie
Professor Dana Nelkin
Professor Samuel Rickless

2019

The Dissertation of Arseny Alexey Ryazanov is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Professors Nicholas Christenfeld, Dana Nelkin, Samuel Rickless, Piotr Winkielman, and Craig McKenzie for the countless hours they spent discussing and guiding a wide range of research ideas that developed into a cohesive research program on moral cognition. I would also like to acknowledge Tinghao Wang, Carsten Erner, and Jonathan Knutzen for their contributions to these projects. I owe my growth as a thinker and researcher to the variety of perspectives that working with such a talented group gave me.

Chapter 1, in full, is a reprint of the material as it appears in Cognitive Science 2018. Ryazanov, Arseny; Knutzen, Jonathan; Rickless, Samuel; Christenfeld, Nicholas, Nelkin, Dana. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 2 has been submitted for publication of the material as it may appear in Cognitive Science, 2019. Ryazanov, Arseny; Wang, Tinghao; Nelkin, Dana; Christenfeld, Nicholas; Rickless, Samuel. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is currently being prepared for submission for publication of the material. Ryazanov, Arseny; Wang, Tinghao; Nelkin, Dana; McKenzie, Craig; Rickless, Samuel. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is currently being prepared for submission for publication of the material. Ryazanov, Arseny; Erner, Carsten; Winkielman, Piotr. The dissertation/thesis author was the primary investigator and author of this material.

Chapter 5 has been submitted for publication of the material as it may appear in the Journal of Experimental Social Psychology, 2019. Ryazanov, Arseny; Nelkin, Dana; Rickless, Samuel; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this material.

Chapter 6, in full, is a reprint of the material as it appears in Nonprofit & Voluntary Sector Quarterly, 2018. Ryazanov, Arseny; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 7, in full, is a reprint of the material as it appears in Social and Personality Psychology Compass, 2018. Ryazanov, Arseny; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 8, in full, is a reprint of the material as it appears in the Journal of Experimental Social Psychology, 2018. Ryazanov, Arseny; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this paper.

VITA

2013 Honours in Arts, McGill University

2013-2015 Teaching Assistant, University of California San Diego

2015 Master of Arts, University of California San Diego

2015-2019 Teaching Assistant, University of California San Diego

2019 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Ryazanov, A. A., Knutzen, J., Nelkin, D., Christenfeld, N. J. S. & Rickless, S. (2018). Intuitive Probabilities and the Limitation of Moral Imagination. *Cognitive Science*, 10.1111/cogs.12598

Ryazanov, A. A., & Christenfeld, N. J. S. (2018). Incremental mindsets and the reduced forgiveness of chronic failures. *Journal of Experimental Social Psychology*, *76*, 33-41.

Ryazanov, A. A., & Christenfeld, N. J. (2018). On the Limited Role of Efficiency in Charitable Giving. *Nonprofit and Voluntary Sector Quarterly*, 0899764018773899.

Ryazanov, A. A., & Christenfeld, N. J. S. (2017). The strategic value of essentialism. *Social and Personality Psychology Compass*, e12370

Larsen, B. A., Ryazanov, A. A., Gravano, J. T., & Christenfeld, N. J. (2015). Competition Breeds Desire. *Basic and Applied Social Psychology*, *37*(1), 81-86.

Leavitt, J. D., Ryazanov, A. A., & Christenfeld, N. J. (2014). Amazing but true: A preference for fiction-like non-fiction. *Scientific Study of Literature*, *4*(2), 196-210.

FIELDS OF STUDY

Major Field: Experimental Psychology

Studies in Social and Cognitive Psychology
Professors Nicholas Christenfeld, Piotr Winkielman, Craig McKenzie

ABSTRACT OF THE DISSERTATION

Harming, Helping, and Blaming: An Inquiry Into Mechanisms Of Moral Cognition

by

Arseny Alexey Ryazanov

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2019

Professor Piotr Winkielman, Chair

Eight chapters examine cognitive processes underlying three moral judgments: How much harm is acceptable for greater good, how much to help others, and how responsible to hold them for their successes and failures. Chapters 1-3 examine how judgments of actions that cause harm to achieve a greater good are sensitive to expected value (the ratio of good done to harm done), outcome likelihoods, and where shifts in outcome likelihoods occur. Findings contradict dominant dual-process theories of moral cognition, which posit that people either react to the harm caused by the action or to the net benefit resulting from it, irrespective of the specific ratio of harm done to good done.

We demonstrate that moral judgments are remarkably sensitive to this ratio, in ways partially consistent with Prospect Theory. Chapter 4 provides further evidence for the interaction of affective and deliberative processes by demonstrating how incidental affect can shift moral risk preferences.

Chapter 5 explores the mental representation of good deeds. The proposed Moral Accounting Model illustrates how moral credit from prior beneficence excuses further beneficence. Effort, effect, domain generalizability, temporal generalizability, and temporal diffusion are identified as features of moral credit. Chapter 6 identifies the extent to which people care about the effectiveness of their beneficence: Though donors prefer to give to more efficient charities of the options they are presented with, whether the options explicitly fail to meet or exceed efficiency standards does not affect donor behavior.

Chapter 7 examines responsibility attribution, challenging a prevalent view in lay theory research that thinking of people as changeable is universally adaptive. It provides a theoretical argument for how viewing people as changeable may result in holding others increasingly personally responsible for their circumstances. Chapter 8 provides empirical evidence for this process: The same mindset inductions used to demonstrate the benefits of changeability are shown to increase blame of others for continual failures.

Implications for real-world decision-making, from how to program autonomous vehicles to avoid collisions, to how to encourage donation to charity, to how to address structural barriers to achievement are discussed.

Introduction

Is doing harm to some for the greater good acceptable? How much should we help others? How responsible should we hold them for their successes and failures? Philosophers have debated what the answers to such questions ought to be for millennia. More recently, psychologists have measured the responses people give to such questions, and how situational factors previously thought to be irrelevant to moral judgment can affect their responses. This dissertation examines the cognitive processes underlying three kinds of moral judgment: expected value calculation in decisions involving causing harm for the greater good, how mental accounting of prior good deeds excuses people from further beneficence, and how attributing responsibility to others reflects changeability beliefs. The real-world implications of these processes are discussed, contributing to the debate on what the answers to such ethical questions ought to be.

Probability and excepted value in moral judgment

Chapters 1-4 examine how moral judgments involving harm to some for the benefit of others incorporate outcome probabilities and expected value calculation. Moral judgment has typically been studied by asking participants whether to engage in variations of actions that are certain to kill one person but will also certainly save five people. Most people endorse diverting a trolley onto a track with one person on it, but not pushing a man into the path of the trolley, to save the five ahead. Such inconsistencies have been taken as evidence of a dual-system model of moral cognition: an irrational, affective, action-focused system is thought to dominate in the "push" case, and a deliberative, outcome-focused system is thought to dominate in the "divert" case.

But what if the “push” case feels less likely to actually save the five, and more likely to kill the one, than the “divert” case, despite the stated certainty of outcomes? Across seven studies, Chapter 1 examines how intuitively-felt outcome likelihoods can override stated outcomes, and that these intuitive likelihoods influence moral judgment. Findings also suggests that some moral principles, such as the distinction between intended harm and foreseen harm, operate, at least in part, through the varying intuitive likelihood of harm each conveys.

Varied outcome likelihoods could impact moral judgments by changing the action’s expected value, or the ratio of good done to harm done. The role of expected value in moral dilemmas has been overlooked because dominant dual-system models of moral cognition propose that people respond either affectively to the harm caused by the action, or deliberatively to the net benefit of the outcome, and as such should be insensitive to the exact ratio of good done to harm done, or the expected value of the action. In Chapter 2, seven studies establish evidence for sensitivity to expected value ratio and use Prospect Theory as a framework for exploring whether moral preferences systematically deviate from expected value when outcomes are probabilistic rather than certain. Moral judgments appear to not be the result of a competition between emotional action-based and rational outcome-based systems, as proposed by dual-system models, but rather the output of a single system that integrates expected value with other considerations. Furthermore, we observe divergent patterns of moral judgment under single evaluation, where a single plan is presented, and under joint evaluation, where a participant chooses between plans. Though judgments of plans involving probabilistic harm track expected value under single evaluation, under joint evaluation participants are willing to endorse plans with lower expected value ratios that carry a lower likelihood of harm over plans with higher expected value ratios that result in certain harm. Chapter 2 thus raises the issue of whether joint evaluation preferences or single evaluation preferences ought to inform questions of practical ethics.

Chapter 3 explores whether people are sensitive to where shifts in probability occur. For example, is raising the risk of harm from 0% to 25% treated differently from raising the risk of harm from 75% to 100%? Findings indicate sensitivity to where the shift in harm occurs, and provide additional evidence for moral preferences under joint evaluation deviating further from expected value than under single evaluation.

Chapter 4 further examines the integration of affective and deliberative paths in moral cognition by examining how incidental affect impacts expected value calculation in moral decisions. We uncover that negative incidental affect, induced by photographs of disgusting food, can shift moral risk preferences in an experimental paradigm based on the classic Asian disease problem, where participants choose between the certain deaths of a smaller number of individuals or risking a chance of a larger number of individuals dying. The comparison of the impact of incidental affect on moral risk preferences and on monetary risk preferences indicates that incidental affect's influence on moral decision-making is the result of a general process, rather than a domain-specific process.

Current basic and applied moral cognition research is conducted with little regard for outcome probabilities or expected value, despite the important role Chapters 1-4 show ratio, outcome likelihoods, and shifts in outcome likelihoods to play. Chapters 1-4 thus bring the study of moral dilemmas closer to real-world decision-making, where moral dilemmas can involve imposing risks of harm and varied numbers of individuals. Such research is relevant to a wide range of practical ethical issues, including whether to pursue novel technologies that promise benefits but also carry risks, how autonomous vehicles should be programmed to make tradeoffs between passenger and pedestrian safety, and public health decisions.

Beneficence & Moral Accounting

The second section of the dissertation examines the cognitive representation of doing good deeds. While the influence of situational factors on whether an individual chooses to help have been studied in great detail, the role of an individual's own prior behavior in deciding whether to help has received less attention outside of research on how good deeds can license people to subsequently behave unethically. One proposed explanation for this moral licensing effect is that good deeds grant moral credit. However, little is known about the mental representation of moral credit. Across seven studies, Chapter 5 proposes the Moral Accounting Model, which outlines how moral credit from prior good deeds excuses further beneficence: Moral credit is found to be luck-dependent, effort-sensitive, time-sensitive, and domain-concentrated.

One implication of the Moral Accounting Model is that people may have fairly limited concern for the impact of their beneficent actions, in that the effectiveness of their actions is just one of many ways by which beneficence grants moral credit. Chapter 6 examines the extent to which donors are sensitive to charity efficiency, an increasingly popular metric of how much of the money raised by a nonprofit is spent on administrative overhead, rather than the actual cause. Four studies find that people have low levels of concern about how charities spend the money donated to them, suggesting that the increased focus on charity performance metrics may not increase overall donation rates, but may guide donations to charities that are presented as relatively efficient, regardless of whether the charities fail to meet or exceed efficiency standards.

Changeability & Blame

The third section of the dissertation turns from judgments of actions to judgments of people in examining how the attribution of responsibility to others reflects changeability beliefs. Two related literatures examine the representation of changeability: essentialism, or the belief that people and social categories have underlying immutable essences, and lay theories of ability, or beliefs regarding whether people's attributes are either fixed or changeable through persistent effort. Though both essentialism and fixed lay theories are generally regarded as harmful cognitive processes that result in maladaptive outcomes, such as stereotyping and underachievement, Chapter 7 identifies an increasing number of findings that are inconsistent with the perspective that changeable lay theories are universally adaptive.

Chapter 7 provides a theoretical foundation for how thinking of people and groups as unchangeable can provide the benefit of absolving them of personal responsibility for having low status, which could be particularly important in situations where structural barriers prevent achievement. In addition to proposing that the observed benefits of changeable mindsets may result from perceiving a positive underlying core essence, rather than the belief in changeability itself, Chapter 7 predicts that a changeable mindset could cause increased blame attribution for continuous failures by ascribing increased personal control for the failures.

Chapter 8 provides empirical evidence for this process: Across five studies, after a standard induction of either an entity (fixed) or incremental (changeable) view of various traits, participants induced to view a trait as changeable blamed a person showing consistently maladaptive levels of the trait more than entity-induced participants did because they perceived the individual as having more control over her failures. Such findings suggest that increasingly popular mindset intervention programs, which induce changeable views of ability and intelligence to help address group and individual achievement gaps, may result in increased blame of those unable to overcome barriers to their achievement, in addition to failing to address the underlying causes of the inequity.

Eight chapters thus uncover the cognitive structures underlying decisions regarding how much harm is acceptable for greater good, how much to help others, and whether to hold them responsible for their successes and failures. Understanding the cognitive processes underlying moral cognition will help uncover general moral principles, which can help guide the resolution of complex, real-world ethical questions, from how autonomous vehicles should be programmed to avoid accidents, to how to encourage more beneficence, to how to combat societal inequities.

Chapter 1: Intuitive Probabilities and the Limitation of Moral Imagination

Arseny A. Ryazanov[1], Jonathan Knutzen[2], Samuel C. Rickless[2],

Nicholas J. S. Christenfeld[1], Dana K. Nelkin[2]

[1]Department of Psychology, [2]Department of Philosophy

University of California San Diego

Author Note:

The order of the senior authors is arbitrary. We would like to thank UC San Diego's Frontiers of Innovation Scholars Program and the Dean of Arts and Humanities, Cristina Della Coletta, for supporting our research.

Corresponding Author: Arseny A. Ryazanov (aryazano@ucsd.edu)

Dept. of Psychology, UC San Diego

9500 Gilman Drive, La Jolla, CA 92093-0109

Abstract

There is a vast literature that seeks to uncover features underlying moral judgment by eliciting reactions to hypothetical scenarios such as trolley problems. These thought experiments assume that participants accept the outcomes stipulated in the scenarios. Across seven studies (N = 968), we demonstrate that intuition overrides stipulated outcomes even when participants are explicitly told that an action will result in a particular outcome. Participants instead substitute their own estimates of the probability of outcomes for stipulated outcomes, and these probability estimates in turn influence moral judgments. Our findings demonstrate that intuitive likelihoods are one critical factor in moral judgment, one that is not suspended even in moral dilemmas that explicitly stipulate outcomes. Features thought to underlie moral reasoning, such as intention, may operate, in part, by affecting the intuitive likelihood of outcomes, and, problematically, moral differences between scenarios may be confounded with non-moral intuitive probabilities.

Keywords: Moral judgment, Morality, Probability, Intuition, Trolley problem

Intuitive Probabilities and the Limitation of Moral Imagination

There is a large and important literature in moral theory centered around eliciting participants' reactions to scenarios of various kinds and drawing conclusions about their judgments. For example, systematic psychological investigation focused on the widely studied Trolley Problems has uncovered a variety of features of scenarios that may account for participants' moral judgments, including whether the agent is described as causing harm to another through physical contact (e.g., Cushman, Young, & Hauser, 2006), whether information about motives is provided (e.g., Nichols & Knobe, 2007), whether harm is described as being intentional (e.g., Cushman et al., 2006; Hauser, Cushman, Young, Kanj-Xing Jin, & Mikhail, 2007; Mikhail, 2000; Cushman, Young, & Hauser, 2006; Moore, Clark, & Kane, 2008; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006), and many other factors.

In arriving at these conclusions, it is typically simply assumed that participants have accepted the claims made about features of the scenarios, and in particular claims about what outcomes will occur if agents act in certain ways (though there are some notable exceptions such as Royzman and Baron's (2002) and Greene et al.'s (2009) studies, discussed below). In a series of studies, we set out to test this assumption. Perhaps when participants are told that, in a given scenario, if an agent acts in a particular way a certain outcome will ensue, participants do not accept this claim. Instead, perhaps they only accept that a certain outcome *might* ensue. If this hypothesis is correct, then participants' judgments could be affected by a factor that is not built into the scenario at all, and is in fact at odds with the scenario as described. Further, if participants are in fact substituting their own judgments about the probabilities of outcomes for scenario stipulations, this should lead us to rethink the way in which participants' reactions to scenarios are tested and what conclusions can be drawn. Our experiments focused mainly on scenarios eliciting moral judgments, but we believe that the relevance of these findings potentially extends to the use of scenarios intended to elicit all sorts of judgments, including moral, prudential, and linguistic, among others.

To see what is at stake, consider Philippa Foot's (1967) classic Trolley Problem, which has been the centerpiece of an entire literature in moral theory. A host of variations of the case has been used as a test of intuitions, and the results have been thought by many to provide strong support for a particular kind of moral theory over its main rival. One variant, developed by Judith Jarvis Thomson (1976), consists of two scenarios. In one, call it "Side Track", five people are tied to a trolley track and will be killed unless a bystander pulls a lever that switches the trolley onto a side track. One person is tied to the side track and will be killed if the bystander pulls the lever. In another, call it "Footbridge," five people are tied to a trolley track and will be killed unless a bystander pushes a large man off a bridge above the track, in which case his body will stop the trolley. If the bystander pushes the large man off the bridge, he will be killed, but the five will be saved. A large majority of participants judges that it is morally permissible to turn the trolley in Side Track, but morally impermissible to push the large man in Footbridge (see, e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). But in each case, if the bystander turns the trolley or pushes the large man, five will be saved and one will die. Thus, if the only morally relevant factor in determining permissibility were consequences—as Consequentialist moral theories have it—then it should be permissible *both* to turn the trolley and to push the large man. Are we simply inconsistent and mistaken in our judgments? Or is there another morally relevant factor (or factors) that can explain in a principled way why turning the trolley is permissible and pushing the large man is not? Some moral theorists have offered the following explanation: in Footbridge, unlike in Side Track, the one is *used* without his consent, if the bystander saves the five (see, e.g., McIntyre, 2004). It is plausible that we have a right not to be used against our will, even when the consequences would be better overall if we were. This fact, it is claimed, explains the moral difference between the two cases, removing the need to attribute inconsistency in most participants' reactions.

The trolley cases continue to be used by philosophers as "intuition pumps," and in classrooms as well as in journals intuitive reactions to the cases are employed in important arguments for non-consequentialism. Intuitions about cases are typically used as part of a methodology known as "reflective equilibrium" (see Rawls, 1971). On this approach, one tries to reach a kind of equilibrium between plausible general principles on the one hand, and intuitions about particular cases on the other. This might require rejecting some intuitions in favor of others, if intuitions are ultimately inconsistent with each other or with plausible principles. While the approach is subtle and does not rely only on intuitions about cases, it is nevertheless true that intuitions play a powerful role in reasoning to a particular moral theory.

As these cases have caught the attention of experimental psychologists in recent years, the burgeoning research program that tests participants' reactions to the cases in systematic and creative ways has yielded intriguing results. One such result is that in cases in which philosophers had initially thought that only a single factor (such as whether or not the victim was used against his will in the plan to save others) explains our differential judgments, the cases also vary on other dimensions that appear to play an explanatory role, such as whether or not there is physical contact in the case (e.g., Cushman et al., 2006). If it turns out that participants' moral judgments are affected by what many philosophers have taken to be morally irrelevant features of cases, then this is very important information for moral theorists to have. For example, if on reflection participants would disavow the idea that directness of physical contact makes a moral difference, then moral theorists cannot claim to have successfully isolated a single morally salient factor that is both explaining and justifying participants' initial moral judgments. Further, this is also important information for psychologists to have in working to identify the mechanisms operative in moral reasoning, and for anyone interested in human behavior in a wide variety of contexts, from decisions in wartime to emergency rooms to legislatures.

We set out to study various versions of moral principles featuring in ordinary moral thought, isolating morally salient features of situations in turn. But some of us, in teaching ethics, had been struck that even after setting out hypothetical cases clearly and repeatedly, students often explicitly commented that they just were not going to accept the stipulated features in the case. This experience led us to try to test whether participants in experiments also resist accepting stipulated features of the cases and, if so, whether this affects their judgments of permissibility.

We are not the first to raise this set of issues. Some researchers have expressed concern about resistance to stipulated features (Bennis, Medin, & Bartels, 2010; Christensen & Gomila, 2012; Green et al., 2009). Some have tried to control for participants' substitution of their own intuitive probability judgments. For example, Royzman and Baron (2002), studying whether participants make different moral judgments when agents harm victims ‘directly’ and ‘indirectly’, eliminated from consideration those participants who they found had judged differentially on probability of outcomes across ‘direct’ and ‘indirect’ scenarios. But they did so along with several other factors and did not test what contribution, if any, differential probability judgments make in generating differential moral judgments.

Greene et al. (2009, p. 365) also tried to control in two ways for what they call “unconscious realism,” that is, “a tendency to unconsciously replace a moral dilemma’s unrealistic assumptions with more realistic ones” (p. 365). First, they told participants the scenarios were “unrealistic” and asked them to suspend disbelief about them, eliminating responses from participants who reported being unable to do so. Second, they asked participants who had previously evaluated the moral acceptability of protagonists performing actions in the various scenarios how probable it was that each protagonist’s action would be (a) as described in the dilemma (e.g., five lives saved at the cost of one), (b) worse than this, or (c) better than this. They found that ratings of higher probability that the outcome would be worse than as described were correlated with lower moral acceptability ratings. As Greene et al. (2009) pointed out, this raises the question of whether participants’ substitution of their own probability estimates of outcomes affects their moral judgments. But, as they also noted, participants’ probability estimates might have been offered as post-hoc rationalizations for their prior moral judgments, and the evidence was at most for a correlation and not causation. So, while they assumed the existence of unconscious realism in an attempt to control for it in investigating other aspects of the scenarios that affect moral judgments, and while their results were suggestive of the hypothesis that the effects of unconscious realism “may be real,” they did not themselves offer studies that test the effect of probability estimates alone on moral judgments. Rather, they urged others to think about controlling for this possibly real effect in future studies.

Finally, Kortenkamp and Moore (2014) focused directly on assessing probability as a factor in moral judgment, but they did not explore whether participants substitute their own probability judgments when responding to scenarios in which outcomes were stipulated as certain. They asked for participants’ probability judgments only in connection with scenarios in which outcomes of protagonists’ actions were already described as *uncertain* to happen (e.g., “x might happen” and “you don’t know whether” (p. 380)). They did, however, find that participants’ responses concerning moral rightness or wrongness differed between scenarios in which outcomes were stipulated as certain and scenarios in which outcomes were stipulated as uncertain. Yet, interestingly, when they probed for probabilities in the uncertain scenarios, they did not find a significant correlation between probability judgments and moral judgments (although they did find a small effect when it came to expected value judgments and moral judgments). Nevertheless, taken as a whole, these studies, while not showing that people substitute probabilities for stipulated certainty, do show that moral judgments are impacted by people thinking about outcomes as probabilistic.

In the studies described below, we directly put to the test the hypothesis that participants substitute their own estimates of the probability of outcomes for those stipulated in the scenarios, and that their doing so affects their moral judgments about cases. Unlike in Kortenkamp and Moore’s (2014) study, we presented scenarios in which the outcomes are stipulated and in which the protagonists know that particular outcomes will occur, to test whether participants are in fact substituting their own judgment for the stipulated features.

And unlike Kortenkamp and Moore (2014), Royzman and Baron (2002), and Greene et al. (2009), we presented pairs of scenarios that varied only in outcomes that differed in their antecedent probabilities (such as whether someone would die from a train running over their foot or their neck). Our approach also differs in that we distinguished among a variety of specific probabilities, including both the probability that the one will die and that the five will live if the agent acts and also both the probability that the one will live and that the five will die if the agent does *not* act. We tested whether participants substitute their own estimates of the probability of any one of these outcomes, and whether any one of these probability estimates could affect participants' moral judgments.

Study 1 has four parts. Although they were initially part of the same experiment, for ease of exposition we present the first two as Studies 1a and 1b. In Study 1a, we presented participants with a Trolley Problem that was designed to vary on whether or not the harm to the single person was a means to save the five, or simply a foreseeable result of saving them. We asked participants to report not only how permissible they thought the action was in the two cases, but also how likely they thought the action was to actually save the five and kill the one. Study 1b also presented participants with pairs of cases designed to isolate changes in the perceived likelihood of outcomes by keeping fixed additional factors such as whether or not the harm was intended as a means. In this study, we limited the changes to antecedent probabilities that the one would die if the protagonist took action (e.g., turning the trolley). Studies 1c and 1d provide evidence that the participants' lack of acceptance of certain outcomes was not a result of the pragmatics of asking them to respond to the likelihood of the event, nor a particular result of participants' use of a slider.

To control for the possibility that participants' responses concerning the likelihood of outcomes might influence their moral assessments, in Study 2 we presented the same scenarios as in Studies 1a and 1b and asked participants only for their moral assessments of the scenarios, without first asking about the likelihood of outcomes.

Studies 3 and 4 parallel the probability-isolating scenarios from Study 1b. Instead of varying the antecedent probability of the one dying if the protagonist took action, we instead varied the antecedent probability of the five being saved if the protagonist took action.

For the first four studies, we employed the same sort of moral judgment scale used by Cushman, et al. (2006), namely, a Likert scale, which runs from 1 (impermissible) to 7 (permissible). Because many philosophers assume that the concept of moral permissibility is binary rather than scalar, in Study 5 we presented participants with one of the scenarios from Study 4, using only a Yes/No question concerning whether the protagonist's action would be morally permissible.

In Study 6, we extended the range of probability estimates assessed. People may bring intuitive probabilities to bear not only on the likelihood of the action resulting in the death of the single person and the saving of the five, but also on the risk to each group in the case of inaction. To investigate how different probability estimates concerning each of four different outcomes might affect moral judgment, we presented participants with four scenario sets using a within-subjects design and asked them for four different probability judgments (two concerning how probable outcomes are if the protagonist acts and two concerning how probable outcomes are if the protagonist does not act), as well as eliciting their moral judgments on a Likert scale.

Finally, in Study 7 we investigated whether scenarios designed to capture moral distinctions in quite abstract terms are nevertheless perceived as varying in perceived probabilities.

Study 1a

Procedure

One hundred and twenty one participants located within the United States were recruited as participants via Amazon's Mechanical Turk (mean age = 35.6, SD = 13.2; 62.8% female). We used a pair of Trolley Cases, adapted from Thomson (1976), written to capture the intended/foreseen distinction, forms of which often appear in the moral judgment literature (e.g., Cushman et al., 2006). In one member of the pair, the death of one individual is a foreseen consequence of diverting a trolley that will otherwise run over five individuals. In the other, the individual's body is instrumental to saving the five: pushing him in front of the trolley stops it from hitting the five. In both scenarios, the death of the one and saving of the five are stipulated as outcomes of the protagonist's action. However, pushing a person in front of a trolley may seem more likely to kill the one, as well as less likely to save the five, than diverting the trolley. (See supplementary Materials for scenarios.)

Participants were also presented with three other scenario sets that will be discussed subsequently as Study 1b. Scenarios were presented in a randomized order. After reading each scenario, participants were asked to respond to how likely they thought the death of the single individual would be if the protagonist decided to perform the action, using a percentage scale (e.g., *If Sam decides to divert the trolley in order to save the five, how likely is it that the lone individual will die?; 0–100%)*. They were also asked to estimate a second likelihood: the likelihood that the five would be saved should the protagonist decide to perform the action (e.g., *If Sam decides to divert the trolley in order to save the five, how likely is it that the five will be saved?; 0-100%).* Next, participants were asked to rate the permissibility of the action on a 7-point scale (*How permissible would it be for Sam to divert the trolley in order to save the five?;* 1 = Impermissible to 7 = Permissible*)*. Participants also provided demographic information.

Results

The pair of scenarios was analyzed for within-subject differences in perceived likelihoods and permissibility using paired-samples t-tests. Participants reported that the one was significantly more likely to die when his death would result from being pushed onto the tracks (mean = 92.9%, *SD* = 15.8), compared to when it would result from the trolley being diverted into him (mean = 89.2%, *SD* = 20.3), *t*(120) = 2.83, *p* = .005, *d* = .20 (see Table 1). Participants also reported perceiving differences in the likelihood of the five being saved, *t*(120) = 3.60, *p* < .001, *d* = .36, with diverting the trolley being more likely to save them than pushing a man in its way (mean push = 83.1%, *SD* = 24.7; mean divert = 91.2%, *SD* = 19.5). Differences in these perceived likelihoods were paralleled by differences in permissibility: action in the scenario perceived as more likely to result in the death of the one individual and less likely to save the five was also rated as significantly less permissible, *t*(120) = 7.20, *p* < .001, *d* = .75 (mean push = 2.96, *SD* = 1.95; mean divert = 4.46, *SD* = 2.07). In neither case did participants take the death of the one, stipulated as the outcome of the protagonist's action, to be 100% likely, nor did participants think that the protagonist's action was 100% likely to save the five, an outcome that was also explicitly stipulated.

A bootstrap mediation analysis with 10,000 resamples using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014) revealed that differences in the perceived likelihood of the one dying did not mediate permissibility differences between the two scenarios, (indirect effect *ab* = .04, *p* = .13, 95% CI [-.01, .12]; direct effect *c'* = 1.45, *p* < .001, CI [1.0, 1.9]; 3% mediated), nor did differences in the perceived likelihood of the five being saved mediate permissibility differences between the two scenarios, (indirect effect *ab* = .05, *p* = .31, 95% CI [-.05, .16]; direct effect *c'* = 1.45, *p* < .001, 95% CI [.32, .65]; 3% mediated), failing to clearly demonstrate a relationship between permissibility and outcome likelihood, despite both significantly varying between the two scenarios.

Table 1.1. Rated likelihoods and permissibility of scenario set intended to vary along the foreseen/intended distinction (Study 1a).

Version	Less likely (divert trolley)	More likely (push man)	sig diff.
Rated likelihood of the one dying (%)	89.2	92.9	**
Rated likelihood of the five being saved (%)	91.2	83.1	***
Permissibility	4.46	2.96	***

***p < .01* ****p < .001*

Discussion

The two scenarios differ along the intended/foreseen distinction, but participants also appear to register a difference in the perceived likelihood of the one dying and the five being saved. Why this difference in perception? One possibility is that the difference between the diversion of the trolley and the use of the bulk of the man as an obstacle to block the trolley's path is itself associated with a difference in the perceived likelihood of the outcome occurring, independently of the protagonist's intentions. Another is that people judge an intended harm as generally more likely to occur, since they think the protagonist would likely take extra steps to try to bring it about, whereas the protagonist might take steps to reduce the likelihood of its occurrence in the case of merely foreseen harm. Given that in one of the scenarios the death of the one person was more likely to occur and that it was less likely for that death to be accompanied by the saving of five, it is reasonable, especially given other differences, that the agent's conduct was judged less morally permissible. We do not, however, find evidence for the perceived likelihoods mediating the permissibility judgments. It may be that the scales, especially an unfamiliar one on moral permissibility, may be too crude to capture participants' intuitions when so many factors are varying between scenarios.

In any case, two things are clear. First, people do not accept outcomes as they are stipulated in the moral scenarios. Second, scenarios written to vary on only one dimension can easily and inadvertently vary also in the perceived likelihood of the specified harm and benefit occurring. Since it is morally relevant how likely harms and benefits are to occur, it would not be surprising if judgments of this sort have some effect on participants' moral verdicts. For the next study, we explored whether participants' perceived likelihoods affect their permissibility judgments in sets of scenarios that are designed not to differ on any other potentially morally relevant dimension.

Study 1b

Procedure

As noted above, the participants of Study 1 were presented with three additional scenario-pairs in which a protagonist is faced with an action that will save five individuals but kill one other. The members of each additional scenario-pair were parallel except for the way in which the single individual would die if the protagonist performed the action. The death of the one individual and the survival of the five were explicitly stipulated as the outcomes of action in every case. In one case the mode of death, for example, the trolley severing his neck, had a plausibly higher perceived likelihood of being fatal than the mode of death in the other, for example, the trolley severing his foot. The other two scenario sets involved (a) pushing a man off a ledge, causing him to fall either 10 feet or 150 feet to his death in order to reach five to save them, and (b) pushing either small rocks or a large boulder off a ledge onto a person below in order to reach five in time to save them. (See Supplementary Materials for scenarios.) If participants substitute their perceived likelihoods of the one individual's death for the scenarios' clear assertion that the individual will die in each case if the agent acts, then they will rate the odds of his death as different between the two cases. This in turn might affect their assessment of the permissibility of the respective actions.

After reading each scenario, participants responded to how likely they thought the death of the single individual would be if the protagonist decided to perform the action, as well as to how likely they thought the saving of the five would be if the action was performed. They also rated the permissibility of the action on a 7-point Likert scale (1 = Impermissible to 7 = Permissible).

Results

We explored the contribution of altering the perceived likelihood of outcomes on permissibility using a repeated measures anova, with scenario set and condition specified as fixed factors. First, we examined whether the predicted permissibility differences emerged from having manipulated the perceived likelihood that the one would die, despite stipulating that the death was certain. The anova yielded a significant permissibility difference, $F(1, 120) = 21.3, p < .001$, partial $\eta^2 = .15$ (mean less likely = 4.68, *SD* = 1.97; mean more likely = 4.27, *SD* = 2.00; see Table 2). There was also a main effect difference between scenario sets on permissibility, $F(2, 240) = 8.48, p < .001$, partial $\eta^2 = .07$ (mean set 1 = 4.56, *SD* = 2.08; mean set 2 = 4.23, *SD* = 1.97; mean set 3 = 4.64, *SD* = 1.92), but there was no significant interaction between scenario set and condition, *F*(2, 240) = 2.03, *p* = .13, partial η^2 = .017, suggesting that all three scenario sets varied equally in permissibility.

Next we examined whether perceived likelihoods of the one dying varied within each scenario pair. A repeated measures anova revealed that they did, *F*(1, 120) = 47.1, *p* < .001, partial η^2 = .28 (mean low = 79.0, *SD* = 27.7, mean high = 90.9, *SD* = 18.2). They also varied between scenario sets, *F*(2, 240) = 4.18, *p* = .016, partial η^2 = .068, (mean set 1 = 87.2, *SD* = 22.5; mean set 2 = 84.4, *SD* = 25.2; mean set 3 = 83.3, *SD* = 24.6). There was no interaction between scenario set and perceived likelihood of the one dying, *F*(2, 240) = 1.98, *p* = .14, partial η^2 = .02, suggesting that perceived likelihoods varied similarly across sets of scenarios.

We also explored whether our scenarios varied the perceived likelihood of the five being saved by the action. Our manipulation did not influence the perceived likelihood that the five would be saved, *F*(1, 120) = .823, *p* = .37, partial η^2 = .007 (mean low = 87.0, *SD* = 21.8; mean high = 87.6, *SD* = 21.0), though there was a significantly different perceived likelihood that the five would be saved between sets of scenarios, *F*(2, 240) = 12.5, *p* < .001, partial η^2 = .09 (mean set 1 = 91.7, *SD* = 18.9, mean set 2 = 85.3, *SD* = 21.9; mean set 3 = 84.9, *SD* = 22.5). There was no interaction between condition and scenario, *F*(2, 240) = .98, *p* = .38, η^2 = .008.

Next, a bootstrap mediation analysis with 10,000 resamples (Tingley, et al., 2014) revealed that differences in the perceived likelihood of the one dying partially mediated the permissibility differences between pairs of scenarios, (mediated effect *ab'* = .22, *p* < .001, 95% CI [.15, .30]; direct effect *c'* = .19, *p* = .02, 95% CI [.03, .36]; 53% mediated). These mediation models specified random intercepts and slopes for subjects.

Table 1.2. Rated likelihoods and permissibility across three scenarios sets that vary along perceived likelihood of the one dying despite death being stipulated (Study 1b).

	Scenario		
Version	Less likely	More likely	sig diff.
Rated likelihood of the one dying (%)	79.0	90.9	***
Rated likelihood of the five being saved (%)	87.0	87.6	ns
Permissibility	4.68	4.27	***

***p < .001

Discussion

Despite outcomes being explicitly stipulated, participants reported a less than 100% likelihood of their occurrence. They also reported divergent likelihoods of the one dying within each of the three pairs of scenarios. Furthermore, an increased perceived likelihood of the one dying as a result of the actions corresponded, quite reasonably, with a judgment of a lower degree of permissibility for the action relative to the other member of its pair: Judgments of the likelihood of the action resulting in the one person's death mediated, partially, judgments of the permissibility of the action. Given that the scenarios were designed to differ morally only in the intuitive likelihood of the action actually resulting in harm to the one, the finding that the mediation is only partial most likely reflects the intrinsic noisiness of participants' estimations of probabilities and permissibility.

Next, we explored two factors that might be thought to cause participants to report uncertainty (or lower than 100% probability) about the outcomes even if they really accepted the certain stipulation. One is the pragmatics of asking likelihood questions about a stipulated outcome, which might suggest that we want an answer less than 100%. The other is that our scale featured a slider that ranged from 0% to 100%, and so any error in reporting certainty would have to be on the low side, which would produce a mean of less than 100%. It is worth noting that such effects cannot readily account for the observed differences between our paired scenarios. If participants are just picking some likelihood arbitrarily less than 100% to make the question seem reasonable, or are just reporting 100% with some (bounded-at-100) error, then these effects should be equal in the two versions of the scenarios, and clearly that is not the case; participants are responding lawfully to perceived differences in outcomes. Nonetheless, we explored whether these two factors contribute to the less-than-certain reports in a scenario with a high perceived likelihood of the stipulated outcome actually coming to pass.

Study 1c & 1d

We used the scenario that had been rated as having an outcome closest to 100% from Study 1b (the trolley severs the neck of the one if the switch is thrown, rated as 94.5% likely to result in the death of the one). If the difference from 100% is just due to random error, then when participants enter their response to the likelihood question freely, rather than using a slider capped at 100%, the error ought to be roughly symmetrical around 100%, and the average should be certainty. This is the approach of Study 1c. Study 1d examined the pragmatics issue by making the ratings not just of the one certain thing, but of a variety of factors whose perceived likelihoods reasonably varied, and included one whose likelihood was 100%. The addition of these other questions before the critical question about the stipulated certainty of the one person dying if the action is taken should reduce any suggestion that we want them to respond with a probability less than 100%.

Procedure 1c

Participants read the sever-the-head scenario, and were asked, *If Rachel decides to divert the trolley in order to save the five, how likely is it that the lone individual will die (e.g., the likelihood of a flipped coin coming up heads is 50%)*? Participants entered any numeric value they wanted into a text box that appeared below the question.

Results 1c

Sixty-two participants (mean age = 34.6, *SD* = 10.3, 74.2% female) rated the action as having an 89.5% (*SD* = 24.7) likelihood of killing the lone individual, which significantly differed from 100%, $t(62) = 3.32$, $p = .002$, $d = .43$. Thus even when entering a percentage value, with no slider-induced norm, participants still did not accept that the stated outcome would certainly happen.

Procedure 1d

Participants read the same scenario and answered four other likelihood questions before the critical one. These questions were chosen to elicit responses that varied, justifying the inquiry about likelihood, and also to include a question to which the participants would answer 100%, to normalize participants reporting 100% on the critical question, if that is what they felt the answer really was. Participants were asked how likely a fair six-sided die was to come up less than four, and how likely the die was to come up less than seven. They also were asked how likely it is that Rachel knows one of the five individuals on the track ahead, and how likely it is that she knows the lone person on the track. Finally, participants were asked the critical question: *If Rachel decides to divert the trolley in order to save the five, how likely is it that the lone individual will die?*

Results 1d

Sixty-four people (mean age: 24.5 *SD* = 12.3; 60.9% female) participated in the study. We included what we thought would be a simple math problem about the likelihood of a regular six-sided die coming up less than seven, so that participants would see that we had at least one question where 100% was a reasonable answer. However, our math problem turned out to be insufficiently easy, and 16 (of 64) people did not respond with 100% on the likelihood of the die showing less than seven. To be conservative, we examined the ratings of the likelihood of the diverted trolley killing the lone individual for the participants who had reported 100% for the die roll (though the data are no different for all subjects). These participants on average reported that likelihood as 92.9% (*SD* = 21.6), again significantly different from 100%, *t*(47) = 2.26, *p* = .029, *d* = .38.

Discussion

Even when not using a slider that caps responses at 100%, and in the context of questions where asking about likelihood was quite ordinary and 100% was a reasonable answer, participants nonetheless did not accept that the action would certainly result in the stipulated death of the one. There was no indication of either the use of a slider or a potentially surprising presentation of questions concerning likelihood contributing to the original effect. Having conservatively selected the scenario rated as most likely to result in death for this test, it is likely that this pattern of findings generalizes to the other scenarios tested. This suggests that rated likelihoods are not the result of a norm dictated by the use of a slider or by the question being asked in isolation, but that participants are indeed not accepting stipulated outcomes, not only in the less likely versions of our scenarios, but in the more likely ones as well.

It is, however, possible that the difference in estimates of the permissibility of the action across each pair depended on participants being sensitized to the perceived likelihood of the act being fatal. That is, being asked to rate the likelihood of the action's fatality could have directed participants' attention to perceived likelihoods and thereby influenced permissibility judgments. The next experiment tested whether the permissibility difference we observed would persist if participants were not asked to rate the likelihood of action outcomes, but just read various scenarios that included stipulations of outcomes (as before). Would participants, in the absence of questions regarding likelihood, nevertheless find action in the scenarios previously found to be more likely to kill the one less permissible?

Study 2

Procedure

One hundred and nineteen participants located in the United States were recruited via Mechanical Turk (mean age = 34.4, *SD* = 11.1; 60.5% female). We repeated the procedure from Study 1, although with both questions regarding likelihoods removed, leaving a single permissibility question for each scenario.

Results

Even without questions regarding likelihood, action in the scenarios in which harm to the one had been previously judged to be more likely was rated as less permissible. In the scenario set from Study 1a, causing the death of the one by throwing a switch was judged more permissible (*M* = 4.29, *SD* = 2.12) than pushing him into the way of the trolley (*M* = 2.96, *SD* = 1.90), *t*(118) = 7.48, *p* < .001, *d* = .66. For the scenario sets used in Study 1b, a repeated measures anova revealed a significant permissibility difference as well, *F*(1, 118) = 15.6, *p* < .001, partial η^2 = .12, (mean low = 4.21, *SD* = 1.97; mean high = 4.01, *SD* = 2.01). The main effect of scenario set was significant *F*(2, 236) = 7.20, *p* = .001, partial η^2 = .058, (mean set 1b = 4.26, *SD* = 1.99; mean set 2b = 3.87, *SD* = 1.99; mean set 3b = 4.20, *SD* = 1.98), as was the interaction of scenario set and condition *F*(2, 236) = 4.50, *p* = .012, partial η^2 = .037 (mean set 2 low = 4.74, *SD* = 2.01, mean set 2 high = 4.39, *SD* = 2.14; mean set 3 low = 4.38, *SD* = 2.01; mean set 3 high = 4.09, *SD* = 1.93; mean set 4 low = 4.95, *SD* = 1.87, mean set 4 high = 4.34, *SD* = 1.93), suggesting that some scenarios were rated as having bigger permissibility differences than others (see Table 3.)

Table 1.3. Rated permissibility between scenarios sets that vary along perceived likelihood of the one dying despite the death being stipulated when not asked probability questions (Study 2).

Scenario Set	Study 1a Set			Study 1b Sets		
Version	Less likely (divert trolley)	More likely (push man)	sig diff.	Less likely (e.g. sever foot)	More likely (e.g. sever neck)	sig diff.
Permissibility	4.29	2.96	***	4.21	4.01	***

****p<.001*

Discussion

The data support the notion that participants use their perceived likelihoods in forming their moral judgments, even when attention is not explicitly called to those likelihoods. It is possible that questions about likelihood may still have somewhat increased participants' attention to their perceived likelihoods in the previous studies, but this study suggests that perceived likelihoods can influence moral judgment even in the absence of such attentional focusing.

In the first four studies, we focused on the perceived likelihood of the protagonist's action being fatal to the one individual in each scenario. As we indicated earlier, in the moral dilemmas used, we also stipulated that the action causing the death of the one individual would save a larger group of five people. This raises the question of whether participants see the protagonist's action as 100% likely to save the five, for a divergence in the perceived likelihood of this outcome could have implications for participants' moral judgments as well. In the next study, we explored whether participants substitute another perceived likelihood for a stipulated outcome that is generally present in trolley problems—the five being saved.

Study 3

Procedure

One hundred and twenty-one participants located within the United States, recruited via Amazon's Mechanical Turk, completed the study (mean age = 33.1, SD = 9.59; 60.3% female). Three pairs of scenarios were created to vary along the perceived likelihood that the five would be saved despite explicit stipulation that the five would indeed be saved. In Set 1, a protagonist could drop either a bicycle or a granite block onto a track to stop a trolley from running over five; in Set 2, a protagonist could use either fishing line or climbing rope to rescue a group of five climbers; and in Set 3, a protagonist intent on keeping a bear from mauling five climbers could scare it off by either throwing pebbles at it or shooting it with a tranquilizer dart. (See Supplementary Materials for scenarios.) In both cases, it was explicitly stated that, should the protagonist decide to act, the action would save the five climbers and kill one individual. Study 3 used the same procedure as Study 1b, asking about the likelihood of the one dying as a result of the protagonist's actions, the likelihood of the five being saved, and the permissibility of the action for each scenario.

Results

We explored whether varying the perceived likelihood of the five being saved by an action would result in permissibility differences between the scenarios. A repeated measures anova found that varying the perceived likelihood of the five being saved by an action resulted in permissibility differences within scenario pairs, *F*(1, 120) = 30.8, *p* < .001, partial η^2 = .20, (mean low = 3.75, *SD* = 1.97; mean high = 4.20, *SD* = 1.92). There was also a main effect of set, such that scenario pairs varied in permissibility from one another, *F*(2, 240) = 10.7, *p* < .001, partial η^2 = .08, (mean set 1 = 3.22, *SD* = 1.86; mean set 2 = 3.86, *SD* = 1.90; mean set 3 = 3.85, *SD* = 1.95). A significant interaction between scenario set and condition, *F*(2,240) = 4.67, *p* = .010, partial η^2 = .04 (mean set 1 low = 4.06, *SD* = 1.95, mean set 1 high = 4.38, *SD* = 1.76; mean set 2 low = 3.69, *SD* = 1.89; mean set 2 high = 4.02, *SD* = 1.90; mean set 3 low = 3.51, *SD* = 2.05, mean set 3 high = 4.19, *SD* = 1.80), suggested that the permissibility difference was much more pronounced in set 3 than in sets 1 or 2.

We next verified that the scenarios had indeed varied the perceived likelihood of the five being saved by the action. Our manipulation did influence the perceived likelihood that the five would be saved, *F*(1, 120) = 24.8, *p* < .001, partial η^2 = .17, (mean low = 70.7, *SD* = 30.8; mean high = 79.7, *SD* = 24.2). There was a main effect of perceived likelihood of the five being saved between sets of scenarios, *F*(2, 240) = 19.1, *p* < .001, partial η^2 = .14 (mean set 1 = 80.8, *SD* = 25.8; mean set 2 = 73.7, *SD* = 27.4; mean set 3 = 71.1, *SD* = 29.9) and again there was an interaction between set and condition, *F*(2, 240) = 7.73, *p* = .001, partial η^2 = .06 (mean set 1 low = 78.6, *SD* = 27.7, mean set 1 high = 82.9, *SD* = 23.8; mean set 2 low = 70.3, *SD* = 29.9; mean set 2 high = 77.2, *SD* = 24.3; mean set 3 low = 63.2, *SD* = 32.8, mean set 3 high = 79.1, *SD* = 24.2). As with the permissibility ratings, the largest effect was observed for the third set.

Next we examined whether perceived likelihoods of the one dying varied within each scenario pair. The repeated measures anova revealed that they did not, *F*(1, 120) =

.88, p = .35, partial η^2 = .007 (mean low = 89.0, *SD* = 16.6; mean high = 89.5, *SD* = 16.5). The perceived likelihood of the one dying as a result of the action did, however, vary between scenario sets, $F(2, 240)$ = 3.86, p =.023, partial η^2 = .03, (mean set 1 = 91.0, *SD* = 14.9; mean set 2 = 89.1, *SD* = 16.0; mean set 3 = 87.6, *SD* = 18.4). There was no interaction between scenario set and perceived likelihood of the one dying, $F(2, 240)$ = 1.19, p = .31, partial η^2 = .010.

Next, a bootstrap mediation analysis with 10,000 resamples revealed that differences in the perceived likelihood of the five dying partially mediated the permissibility differences between pairs of scenarios, (mediated effect *ab* = .21, p < .001, 95% CI [.12, .36]; direct effect *c'* = .24, p < .001, 95% CI [.14, .28]; 47% mediated). These mediation models specified random intercepts and slopes for subjects.

Because we had observed an interaction, each pair of scenarios was analyzed for differences in the two perceived likelihoods and permissibility with paired-samples t-tests. In Set 1, participants did not report dropping the granite block as more likely to save the five (*M* = 82.9%, *SD* = 23.8) than dropping the bicycle (*M* = 78.6%, *SD* = 27.7), $t(120)$ = 1.77, p = .080, d = .17, though there was an observed permissibility difference: Action in the scenario written to be intuitively more likely to save the five was seen as more permissible, $t(120)$ = 2.89, p = .005 d = .17, (mean block = 4.38, *SD* = 1.76; mean bicycle = 4.06, *SD* = 1.95). The perceived likelihood of the one dying marginally differed between the two scenarios, $t(120)$ = 1.79, p = .075 d = .12, (mean block = 90.1%, *SD* = 15.2; mean bicycle = 91.9%, *SD* = 14.7). In both scenarios, participants did not take the death of the one to be 100% likely to occur, nor did they assign a likelihood of 100% to the five being saved.

In Set 2, participants reported that securing the five climbers with climbing rope was significantly more likely to save them (*M* = 77.2%, *SD* = 24.3) than securing them with fishing line (*M* = 70.3%, *SD* = 29.9), *t*(120) = 2.94, *p* = .004 *d* = .25. A difference in permissibility paralleled the difference in the perceived likelihood of the five being saved: action in the scenario perceived as more likely to save the five was seen as more permissible, *t*(120) = 3.12, *p* = .002 *d* = .17 (mean climbing rope = 4.02, *SD* = 1.90; mean fishing line = 3.69, *SD* = 1.89). The perceived likelihood of the one dying did not differ between the two scenarios, *t*(120) = .175, *p* = .86, *d* = .05, (mean climbing rope = 89.0%, *SD* = 15.5; mean fishing line = 89.2%, *SD* = 16.6). Neither scenario produced a judgment of 100% likelihood of the one dying or a judgment of 100% likelihood of the five being saved.

In Set 3, participants reported that shooting the bear with a tranquilizer gun was significantly more likely to save the five climbers (*M* = 79.1%, *SD* = 24.2) than throwing pebbles at it (*M* = 63.2%, *SD* = 32.8), *t*(120) = 5.70, *p* < .001, *d* = .55. A difference in permissibility paralleled the difference in the perceived likelihood of the five being saved: action in the scenario perceived as more likely to save the five was seen as more permissible, *t*(120) = 6.08, *p* < .001, *d* = .35, (mean tranquilizer = 4.19, *SD* = 1.80; mean pebbles = 3.51, *SD* = 2.05). The perceived likelihood of the one dying did not differ between the two scenarios (mean tranquilizer = 87.7%, *SD* = 18.8; mean pebbles = 87.6%, *SD* = 18.1), *t*(120) = .190, *p* = .850, *d* = .01. Participants did not take the stipulated outcome of the one dying to be 100% likely, nor did they assign a 100% chance to the stipulated outcome that the five would be saved.

Discussion

Across the three sets of scenarios, despite the saving of the five being explicitly stipulated, participants reported not only that the likelihood of the outcome was well below 100%, but also that there was a difference in likelihood between pairs of scenarios. Furthermore, an increased perceived likelihood of the five being saved corresponded to an increased degree of permissibility of the action (just as an increased perceived likelihood of the one dying corresponded to a decreased degree of permissibility in Study 1b).

The next study examined whether, as with the pair of studies exploring the perceived likelihood of the one individual's death, the differences in permissibility would persist without focus being directed to the likelihoods. That is, in the absence of questions regarding likelihoods, would participants nevertheless find action in the scenario versions previously found to be more likely to result in the deaths of the five to be less permissible? Because Set 1 was not rated as having a significant difference in perceived likelihood of the five being saved, and Set 2 had modest effect size, relative to both Set 3 and the scenarios sets in Study 1b, Set 3 may be the most informative test case going forward.

Study 4

Procedure

One hundred and nineteen participants located in the United States were recruited via Mechanical Turk (mean age = 31.4, *SD* = 8.59; 55.4% female). Study 4 repeated the procedure from Study 3, although with questions relating to likelihood removed, leaving just the single permissibility question for each scenario.

Results

Permissibility differences between scenario sets that vary in perceived likelihood, but do not probe likelihood, were analyzed using a repeated measures anova. The pairs of scenarios, written to vary in perceived likelihood of the five being saved only, did not show an overall difference when not probing for likelihood, *F*(1, 118) = .74, *p* = .39, partial η^2 = .006, (mean low = 3.64, *SD* = 1.75; mean high = 3.67, *SD* = 1.77). The main effect of scenario set was significant *F*(2, 236) = 5.60, *p* =.005, partial η^2 = .045, (mean set 1 = 3.80, *SD* = 1.79; mean Set 2 = 3.60, *SD* = 1.75; mean set 3 = 3.57, *SD* = 1.74), and there was a significant interaction of scenario set and condition *F*(2, 236) = 3.01, *p* = .05, partial η^2 = .025. This suggested that some scenarios were rated as having bigger permissibility differences than others.

Because we again found an interaction, we examined the sets individually. Permissibility differences were observed in one of three scenario sets. In Set 3, where the protagonist could save climbers from a bear by using either a tranquilizer gun or pebbles, the difference in permissibility remained significant, *t*(118) = 2.36, *p* = .02, *d* = .10, (mean tranquilizer = 3.66, *SD* = 1.72; mean pebbles = 3.48, *SD* = 1.76). Permissibility did not, however, differ in Set 1, where either a bicycle (*M* = 3.83, *SD* = 1.79) or granite block (*M* = 3.76, *SD* = 1.80) would be dropped, *t*(118) = .97, *p* = .33, *d* = .04, or in Set 2, where climbers would either be secured using climbing rope (*M* = 3.59, *SD* = 1.78) or fishing line (*M* = 3.61, *SD* = 1.72), *t*(118) = .35, *p* = .72, *d* = .01.

Discussion

The findings show that, even when participants were not explicitly asked about likelihood, the perceived likelihood of an action's saving five can override stipulated outcomes and influence moral judgments. These results suggest that having participants think specifically about the likelihood of the stipulated outcome of action did sensitize them to those differences, and likely enhanced the rated permissibility differences in the prior study. Nonetheless, a difference can emerge even without attention being explicitly directed to the likelihoods.

What explains the fact that we found significant differences with respect to moral permissibility judgments in Set 3, but not Set 1 and Set 2? It is notable that when participants were presented with Set 1 in Study 3, we found an insignificant difference in participants' perceived likelihoods of outcomes. Perhaps surprisingly, participants there rated the likelihood of the five being saved to be very similar whether a bicycle or a 100-ton granite block was released onto the track. Participants did offer significantly different likelihood judgments in Set 2, but notably, these differences were not as great as those in Set 3. It is possible that in cases like Set 1, where participants are not inclined to offer very different likelihood judgments even when asked about likelihood of outcomes, simply being asked can enhance the salience of even minimal perceived likelihood differences in answering questions about moral permissibility. It is also possible that when asked about the likelihood of saving the five, other perceived likelihoods are brought to salience. (We explore the possibility in Study 6 that participants are responding to a number of different perceived probabilities related to each scenario.) In any case, it is perhaps not surprising that the set which revealed the greatest difference in perceived likelihoods in Study 3, namely Set 3, generated the greatest difference in moral judgments when participants were not asked about likelihood at all.

Until now, we have used continuous ratings of permissibility. Participants use such a scale without complaint, and the responses do appear to covary with their perception of various morally relevant likelihoods. However, the meaning of such a continuum is not entirely clear. What, for example, does it mean for an action to be rated halfway between permissible and impermissible? In many contexts, people treat the words “permissible” and “impermissible” in a way parallel to the words “legal” and “illegal,” and yet it is not at all clear what it would mean to be halfway between “legal” and “illegal.” Moreover, most conceptual analyses of permissibility and impermissibility treat these concepts as binary. For example, “impermissible” is most often treated as synonymous with “forbidden” or “something one ought not to do”, which express non-scalar notions. And the very few defenses of scalar conceptions of rightness or permissibility, such as Lockhart (2000) and Peterson (2013), recognize that the scalar view is non-standard. Thus, there is at least an important concept of (im)permissibility that is non-scalar.

Now it does not follow from the fact that there is a non-scalar concept of (im)permissibility that people are unable to use the words “permissible” and “impermissible” to express scalar notions. We believe that this is what is going on when participants are presented with Likert permissibility scales. For example, they may mean by “less permissible” something like “involving a more serious moral infraction” or “its being *worse* (or causing *more harm*, or infringing a *more stringent* right) to do one thing rather than another” (such as to kill rather than steal a piece of gum). At the same time, the standard conception of moral permissibility in moral theory corresponds to the standard (and, so far as we are aware, the only) treatment of legality in legal theory. All felonies are illegal, and none is more (or less) illegal than any other. But clearly some felonies are worse than others, cause more harm than others, or infringe more stringent rights than others.

If there is an important concept captured by "permissible" that is binary, and judgments of permissibility are affected by participants' substituted perceived likelihoods of outcomes, then we should expect to see this reflected when using a binary measure of permissibility. For the next study we adopted a binary (yes/no) permissibility evaluation and explored whether differences in the perceived likelihoods in scenarios can influence binary permissibility judgments. For this study, we also tested whether the permissibility judgment differences would be apparent when each participant rated only one of the pair of scenarios, preventing their making any comparative likelihood or permissibility rating based on the small alterations within each pair. We also did not ask for explicit likelihood judgments, to avoid the possibility of priming driving any difference in permissibility judgments.

Study 5

Procedure

One hundred and twenty-two participants located in the United States were recruited via Mechanical Turk (mean age = 35.5, *SD* = 11.7; 63.9% female). Study 5 used Set 3 from Studies 3/4 to investigate binary permissibility judgments. Each participant saw only one member of the pair of scenarios. Rather than respond on a 7-

point scale, participants made a binary yes/no judgment regarding the permissibility of the protagonist's action.

Results

Of the 61 participants who saw the version of the scenario in which the protagonist can save the five climbers from the bear by shooting it with a tranquilizer gun, 32 (52.5%) thought it would be permissible to perform the action. Of the 61 participants who saw the version in which the protagonist can save the five climbers from the bear by throwing pebbles at it, only 21 (34.4%) thought it would be permissible to act. The difference in the distributions of judgments was significant, χ^2(N = 122) = 4.06, *p* = .044, OR = 2.10, 95% CI [1.01, 4.36], suggesting that participants were more likely to think it permissible to kill the one to save the five when the bear would be shot with the tranquilizer gun than when pebbles would be thrown at the bear, despite both versions explicitly stating that the action would save the five climbers.

Discussion

In this study, perceived likelihoods were found to override explicitly stated outcomes and to affect binary permissibility judgments, just as in previous studies they had affected scalar permissibility judgments. It was also apparent that the effect emerges in a between-subjects design just as it had in within-subjects designs. And, as before, the effect emerges even when there is no explicit question about the likelihood of outcomes.

Study 6

So far we have examined the way that people bring their own perceived likelihoods to bear on moral decisions with respect to pairs of written scenarios where the chances of the protagonist's action leading to the death of one individual, or to the saving of five individuals, are plausibly different. However, this does not exhaust (potentially morally relevant) likelihoods. In particular, one can distinguish between the likelihoods of outcomes consequent upon inaction as well as upon action. The likelihoods in the previous studies reflect the likelihood of certain events occurring should a particular action be performed. What if one varied the likelihood of events occurring should the action *not* be performed? Would participants judge it to be less permissible to kill one to save five if the five had some chance of surviving anyway? Or would they judge it to be more permissible to act if the single person was already at some risk of death? This study explores these questions.

Procedure

One hundred and twenty-seven participants located within the United States were recruited via Mechanical Turk (mean age = 35.0, *SD* = 10.4; 63.0% female). One set from Studies 1/2 was used, where our intent was to manipulate perceived likelihoods of the one dying, as was one set from Studies 3/4, where the intent was to manipulate perceived likelihoods of the five being saved. Two new scenario pairs were created to vary along perceived likelihoods relating to inaction. The first set stipulated that the one would survive if the action was not taken, but varied along the perceived likelihood that this would occur. The second was intended to let intuitions vary about the likelihood that the five would die if the action was not taken. Thus, in this study, we inquired about the perceived likelihoods of all four of these outcomes: (a) how likely the one was to live if the protagonist did not act, (b) how likely the one was to die if the protagonist acted, (c) how likely the five were to die if the protagonist did not act, and (d) how likely the five were to live if the protagonist acted. We also asked about permissibility. This allowed us to explore whether all of these perceived likelihoods have ramifications for moral judgments, as well as the extent to which they can be altered independently, even in scenarios where the outcomes are stipulated.

We used Set 1 from Studies 1b and 2, in which the action would either kill the one by severing his neck or his foot, which varied the perceived likelihood that the one would die as a result of the action. The next set varied the perceived likelihood that the one would die if the action were not taken. In one version an individual miner has his foot stuck in the rubble, and in the other he is trapped and severely injured. Although we asserted that the one would live in both cases, it seems intuitively that the one whose foot is stuck is more likely to do so. As a result, it might seem that blasting the hole in the rubble he is trapped in to rescue five other miners is more permissible in the latter case.

We used Set 3 from Studies 3 and 4, in which the five could be saved from a bear by throwing pebbles at the bear or shooting it with a tranquilizer dart, thus varying the perceived likelihood that the five would be saved with action. The final set varied the perceived likelihood that the five would survive despite inaction: Although in both members of the pair the death of the climbers is stipulated, it might be thought less likely that the climbers will die when they fall 10 feet onto a flat rock surface, compared to when they plummet 1,000 feet onto jagged rocks. This may make killing one person to save five seem less permissible.

Participants saw each version of each scenario, randomly ordered. After reading each scenario, they answered four questions regarding likelihood and one regarding permissibility. Two of the likelihood questions concerned the likelihood of the one dying and the five being saved if the protagonist decides to act—these were the same questions we asked in Studies 1 and 3. Two more likelihood questions, relating to inaction, were also asked: one about the likelihood of death of the one individual if the protagonist decides not to act (e.g., *If Jason decides not to rapidly traverse the narrow ledge, how likely is it that the man on the ledge will* die? 0-100%), the other about the likelihood of the five surviving if the protagonist decides not to act (e.g., *If Jason decides not to rapidly traverse the narrow ledge, how likely is it that five will survive*? 0-100%).

Results

Set 1. This set was intended to vary the perceived likelihood of the one dying if the protagonist acts. Indeed, replicating the finding from Study 1b, severing the neck was rated more likely to lead to death (*M* = 92.5%, *SD* = 19.4) than severing the foot (*M* = 77.8%, *SD* = 27.3), *t*(126) = 5.61, *p* < .001, *d* = .62 (see Table 4). Fifty out of the 127 participants thought severing the foot had a greater than 95% chance of actually killing the one, whereas 99 out of the 127 thought severing the head did. A difference in permissibility paralleled this difference in perceived likelihood: Action in the version of the scenario rated more likely to lead to the one individual's death was rated less permissible (mean neck = 4.63, *SD* = 2.02; mean foot = 4.94, *SD* = 1.91), *t*(126) = 2.61, *p* = .010, *d* = .16.

Other perceived likelihoods did not differ between the two scenarios of Set 1. As in Study 1b, the perceived likelihood of the five being saved if action was taken did not differ, *t*(126) = 1.28, *p* = .20, *d* = .11(mean neck = 89.9%, *SD* = 21.2; mean foot = 92.0%, *SD* = 17.1). The perceived likelihood of the one dying if action was not taken also did not differ, *t*(126) = .872, *p* = .39, *d* = .08 (mean neck = 19.3%, *SD* = 30.1; mean foot =

17.0%, *SD* = 27.4). The perceived likelihood of the five surviving if action was not taken did not differ either, *t*(126) = 1.24, *p* = .22, *d* = .11 (mean neck = 21.1%, *SD* = 30.9, mean foot = 17.9%, *SD* = 27.3). Outcomes that were stipulated to occur were assessed as less than 100% likely, and outcomes that the scenario suggested would not occur (e.g., the single person dying on the tracks even if no action is taken) were rated as having some non-zero chance of occurring.

Set 2. This set was intended to vary the perceived likelihood of the five being saved if the protagonist acts. As in Study 4, shooting the bear with a tranquilizer gun was again thought to be more likely to save the five (*M* = 83.2%, *SD* = 21.0) than throwing pebbles at it (*M* = 72.1%, *SD* = 28.6), *t*(126) = 5.14, *p* < .001, *d* = .44. Fourty-one participants thought it more than 95% likely that throwing pebbles at the bear would prevent it from attacking, while 49 participants thought the same about shooting the bear with the tranquilizer gun. A difference in permissibility paralleled this difference in perceived likelihood: The action judged more likely to save the five was deemed more permissible, *t*(126) = 2.36, *p* = .020, *d* = .12 (mean tranquilizer = 4.06, SD = 1.97; mean pebbles = 3.83, SD = 1.94).

Other perceived likelihoods did not differ between the scenarios of Set 2. Again, the perceived likelihood of the one dying if action was taken did not differ, *t*(126) = .661, *p* = .51, *d* = .04 (mean tranquilizer = 89.3%, *SD* = 18.6; mean pebbles = 88.5%, *SD* = 20.8). The perceived likelihood of the one dying if action was not taken did not differ either, *t*(126) = .350, *p* = .73, *d* = .02 (mean tranquilizer = 11.8%, *SD* = 21.3; mean pebbles = 12.3%, *SD* = 22.6), nor did the perceived likelihood of the five surviving if action was not taken, *t*(126) = .718, *p* = .47, *d* = .07 (mean tranquilizer = 28.1%, *SD* =

29.2; mean pebbles = 26.1%, *SD* = 28.0). Outcomes that were stipulated to occur were assessed as less than 100% likely, and outcomes that were stipulated as not occurring were assessed as more than 0% likely.

Set 3. This set was intended to vary the perceived likelihood of the death of the one, if the protagonist does not act to save the five. In the mining scenario participants thought the one was more likely to die when trapped and severely injured (*M* = 39.6%, *SD* = 30.4) than when only his foot is stuck (*M* = 29.1%, *SD* = 33.9), *t*(126) = 3.40, *p* < .001, *d* = .33. A difference in permissibility paralleled this difference in perceived likelihood: Action leading to the one's death in the scenario in which the one was perceived as less likely to survive in the absence of action was judged more permissible, *t*(126) = 4.36, *p* < .001, *d* = .24 (mean severe = 4.89, *SD* = 1.91, mean foot = 4.42, *SD* = 1.97).

Other perceived likelihoods did not differ between the pairs of Set 3. The perceived likelihood of the one dying if action was taken did not differ, *t*(126) = .587, *p* = .56, *d* = .07 (mean severe = 90.3%, *SD* = 20.1; mean foot = 88.9%, *SD* = 22.6). The perceived likelihood of the five being saved if action was taken did not differ, *t*(126) = .873, *p* = .38, *d* = .08 (mean severe = 82.9%, *SD* = 23.5; mean foot = 84.6%, *SD* = 20.8), nor did the perceived likelihood of the five surviving if action was not taken, *t*(126) = .280, *p* = .78, *d* = .02 (mean severe = 20.1%, *SD* = 26.3; mean foot = 20.7%, *SD* = 27.7). Stipulated outcomes were not judged as either definitely occurring or definitely not occurring.

Set 4. The final set was intended to vary the perceived likelihood of the five surviving without the protagonist acting to secure their bridge. Despite the stipulation that falling would lead to the climbers' deaths in either scenario, participants reported thinking that if the climbers fell 10 feet onto a flat rock surface they were more likely to survive (*M* = 37.8%, *SD* = 36.2), compared to the scenario where they would plummet 1,000 feet onto jagged rocks (*M* = 19.4%, *SD* = 27.6), *t*(126) = 5.16, *p* < .001, *d* = .57. Judgments of permissibility significantly differed, inversely to the difference in perceived likelihood of the five surviving, *t*(126) = 5.20, *p* < .001, *d* = .34 (mean short fall = 3.64, *SD* = 2.01; mean long fall = 4.31, *SD* = 1.93).

Perceived likelihoods other than the likelihood of the five surviving even without action did not differ between scenarios. Participants did not think that the pair of scenarios varied along likelihood of the one dying if action was taken, *t*(126) = .594, *p* = .55, *d* = .05 (mean short fall = 90.1%, *SD* = 16.2; mean long fall = 89.3%, *SD* = 17.7). Nor did they take the scenarios to differ in likelihood of the one dying if action was not taken, *t*(126) = .395, *p* = .69, *d* = .04 (mean short fall = 13.8%, *SD* = 25.3; mean long fall = 14.7%, *SD* = 25.1). However, there was a marginal difference in the perceived likelihood of the five being saved, *t*(126) = 1.90, *p* = .060, *d* = .16 (mean short fall = 84.1%, *SD* = 19.9; mean long fall = 80.6%, *SD* = 23.7), suggesting that the perceived likelihood of the five surviving due to inaction contributed to the perceived likelihood of the five being saved. Outcomes that were stipulated to occur were assessed as less than 100% likely, and those that participants might be expected to understand as not occurring were judged as more than 0% likely.

Table 1.4. Rated likelihoods and permissibility between scenario sets that vary along four intuitive probabilities, despite the outcome being stipulated for each (Study 7).

Scenario Set	Set 1			Set 2			Set 3			Set 4		
Version	Less likely (sever foot)	More likely (sever neck)	sig diff.	Less likely (pebbles)	More likely (tranquilizer gun)	sig diff.	Less likely (foot trapped)	More Likely (body trapped)	sig. diff.	Less likely (fall 10 feet)	More likely (fall 1000 feet)	sig diff.
Rated likelihood of the one dying if action taken(%)	**77.8**	**92.5**	*******	88.5	89.3	***	88.9	90.3	ns	80.1	89.3	ns
Rated likelihood of the five being saved if action taken (%)	92.0	89.9	ns	**72.1**	**83.2**	*******	84.6	82.9	ns	84.1	80.6	†
Rated likelihood of the one dying w/ out action (%)	17.0	19.3	ns	12.3	11.8	ns	**29.1**	**39.6**	*******	13.8	14.7	ns
Rated likelihood of the five surviving w/out action (%)	17.9	21.1	ns	26.1	28.1	ns	20.7	20.1	ns	**37.8**	**19.4**	*******
Permissibility	4.94	4.63	*	3.83	4.06	*	4.42	4.89	***	3.64	4.31	***

Bolded likelihoods reflect the probability intended to intuitively vary in each scenario
*†p < .1   * p < .05   **p < .01   ***p < .001*

In this study, where we find differences in the perceived likelihoods of the one dying and the five surviving without action, and the one surviving and the five dying with action, and associated differences in the permissibility of action, we can explore the relationship between perceived likelihoods and permissibility. That is, we can test whether the alterations in perceived likelihood are mediating the differences in permissibility. Bootstrap mediational analysis (Tingley et al., 2014) was used to investigate this for each of the four sets, with random intercepts entered for each subject.

There was a significant indirect effect of Set 1 (which was designed to manipulate perceived likelihood of the action actually killing the single person) on permissibility through probability, *ab* = .15, *p* < .001, 95% CI [.04, .28] (10,000 resamples). The direct effect was no longer significant, *c'* = .16, p = .21, 95% CI [-.09, .40], suggesting that the difference in perceived likelihood of the one actually dying fully mediated permissibility differences. While the effect sizes reported throughout these studies are small, their magnitude may be misleading due to the large variance in perceived likelihoods—despite change in perceived likelihoods being small relative to this range, their movement entirely mediates the permissibility difference observed here.

For Set 2 (which manipulated perceived likelihood that the action would indeed save the five), there was a significant indirect effect on permissibility through probability, *ab* = .21, *p* < .001, 95% CI [.12, .33] (10,000 resamples). The direct effect was no longer significant, *c'* = .01, p = .89, 95% CI [-.18, .20], suggesting that the difference in perceived likelihood of the five actually being saved by the action fully mediated permissibility differences between the two scenarios.

There was a significant indirect effect of Set 3 (which manipulated the perceived likelihood that the single person might die even without action) on permissibility through probability, *ab* = .12, *p* < .001, 95% CI [.04, .22] (10,000 resamples). The direct effect remained significant, *c'* = .35, *p* < .001, 95% CI [.14, .56], suggesting that the difference in perceived likelihood of the one dying regardless of whether action was taken partially mediated permissibility differences. This may suggest that some counterfactual probabilities are hard to explicitly convey, despite their effect on permissibility.

Finally, there was a significant indirect effect of Set 4 (which manipulated the perceived likelihood of the five surviving even without action) on permissibility through probability, *ab* = .29, *p* < .001, 95% CI [.16, .45] (10,000 resamples). The direct effect remained significant, *c'* = .39, p < .001, 95% CI [.14, .63], suggesting that the difference in perceived likelihood of the five surviving regardless of whether action was taken partially mediated permissibility differences. This provides further evidence that the rated likelihood may not necessarily perfectly capture the intuitive likelihood's impact on permissibility judgments.

Discussion

Study 6 found that each scenario we had written to specifically vary one of four perceived likelihoods was associated with corresponding differences in participants' judgments about likelihood as well as permissibility. This result supports the hypothesis that intuitions regarding a variety of kinds of likelihoods can be substituted for scenario stipulations. Stipulated outcomes were never taken to have a 100% probability of occurring. The action that was described as saving five people was not seen as perfectly likely to have that effect, nor was the action described as omitting to save them seen as leaving them to their certain death. Similarly, the action that was described as resulting in the death of the single individual was not seen as 100% likely to have that effect, nor was the corresponding omission seen as any kind of guarantee that the individual would survive. Moreover, the fact that differences in each type of perceived likelihood were associated with differences in permissibility judgments supports the hypothesis that intuitions about each type of likelihood can have an effect on moral judgment. In all cases, we found that the perceived likelihoods did mediate, either partially or fully, the impact on permissibility. This reinforces the notion that intuitions about actual likelihoods, even when the relevant outcomes are stipulated to occur, can alter moral judgments.

Our studies so far suggest that perceived likelihoods are at least partial drivers of permissibility judgments. We can now revisit the hypothesis we entertained in our discussion of Study 1a, viz. that perceived intention may influence probability judgments. Our final study explored whether such probability differences are relevant not only in consideration of the details of scenarios, but even occur as a result of the underlying moral principles the scenarios are thought to convey. The question is whether, regardless of the details, describing a harm as intended can make it seem more likely to occur than describing it as merely foreseen.

Study 7

Procedure

One hundred and seventeen adults located in the United States were recruited via Mechanical Turk (mean age = 36.7, *SD* = 12.0; 60.7% female). Study 7 presented participants with two scenarios, written to differ along the foreseen/intended distinction abstractly, to avoid any potential effect of differing details on perceived likelihoods. Participants read both scenarios, and then two questions.

Foreseen:
Plan X
Imagine you are told that Jason has a plan to save a group of children from a terrorist attack, and that a foreseen, but not intended, consequence of Jason carrying his plan out is that Thomas, a bystander, is killed.

Intended:
Plan Y
Imagine you are told that Harry has a plan to save a group of children from a terrorist attack, and that Harry intends to kill a bystander, Robert, because for Harry's plan to work it is necessary that Robert is killed.

Participants responded to "*Which plan is more likely to result in the death of the bystander?; -7 = Plan X is more likely than Plan Y to result in the bystander's death to 7 = Plan Y is more likely than Plan X to result in the bystander's death*" and "*Which of the two plans is morally worse?; -7 = Plan X is much worse than Plan Y to 7 = Plan Y is much worse than Plan X*".

Results

Responses to the scenario set were analyzed for differences in relative perceived likelihoods and relative moral acceptability with one sample *t* tests comparing responses to 0 (equal). Participants reported thinking that the bystander was significantly more likely to die in the intended plan than in the foreseen plan scenario $t(116) = 6.54$, $p < .001$, $d = .60$ ($M = 1.93$; $SD = 3.20$). A difference in relative moral acceptability paralleled this difference in perceived likelihood: the plan describing the intended death of a bystander was judged to be significantly less morally acceptable than the plan describing a foreseen death, where the bystander's death had been perceived to be less likely, $t(116) = 8.92$, $p < .001$, $d = .82$ ($M = 2.99$; $SD = 3.63$). Relative perceived likelihood ratings correlated with relative moral acceptability ratings, $F(1, 115) = 23.4$, $p < .001$, $r = .41$.

Discussion

We began our investigation in Study 1a with a scenario pair modeled on the classic trolley dilemma. Responses to this dilemma and others like it are sometimes taken to support the claim that something like the Doctrine of Double Effect is tacitly operative in moral judgment. Our final study shows that the foreseen/intended distinction, presented in the abstract and stripped of additional likelihood-relevant details, still conveys likelihood-relevant information that appears to contribute to differences in moral judgments. This is an important result. Together with the results from Studies 1-6, it suggests that judgment patterns consistent with the Doctrine of Double Effect cannot be readily interpreted as showing that something like the DDE is tacitly operative in moral thought, since in many cases such a principle is confounded with differences in perceived likelihoods. Since the details of the scenarios—both those related to and those unrelated to intention—convey differential likelihood information relevant to and potentially impacting moral judgment, moral dilemma tasks designed to test for the DDE face the serious challenge of isolating a pure comparison on the dimension of foreseeability and intention, and not implicitly conveying differential likelihoods of outcomes. This is not to say that the challenge cannot be met, or that something like the DDE is in fact tacitly operative. But more work needs to be done to demonstrate a clear test of the operation of the DDE.

Meta-Analytic Results

A meta-analytic approach was taken to combine results across the studies that were designed to differ only in perceived likelihoods, not in morally relevant factors (namely, Studies 1b and 2-6). We calculated average effect sizes for permissibility differences between pairs of scenarios across studies when participants had been asked to rate likelihoods of the outcomes, and also when they had not been so asked, controlling for the sample size of each study. Studies that asked about likelihoods alongside permissibility had a standardized effect size of .24, *p* = .001, when averaging the overall effect sizes of Study 1b, Study 3, and Study 6, taking into account the number of participants in each study. A parallel examination of studies that did not ask about likelihood showed a standardized effect size of .17, *p* = .022 across Study 2, Study 4, and Study 5. The last study required using Chinn's (2000) approach to converting the odds ratio to an effect size.

General Discussion

In seven studies, we explored the extent to which people accept stipulated outcomes in hypothetical moral scenarios, how this can differ between scenarios, and how such differences in intuitive likelihoods can result in different judgments of permissibility. Study 1a found that a scenario set representative of those used to study the Doctrine of Double Effect differed not only in terms of whether the harm was intended or foreseen, but also in the intuitive likelihood of death of the one individual, as well as in the intuitive likelihood of the five being saved, despite both of these outcomes being stipulated. Study 1b sought to test whether scenario pairs which held everything else constant and differed only with respect to the intuitive probability that the individual would die if the protagonist acted would produce differences in probability and permissibility ratings, and found, across multiple scenarios, that they do. Studies 1c and 1d supported the idea that participants reporting less-than-certain likelihoods was not an artifact of our experimental method.

Study 2 verified that differences in permissibility ratings could arise in the absence of questions regarding the intuitive likelihood of the death of the one individual, though these questions may have sensitized participants to intuitive differences in likelihood. Study 3 examined scenarios written to isolate the difference in intuitive likelihood of the five being saved, rather than of the one dying, and showed that this difference alone could also result in a difference in permissibility ratings between scenarios. Study 4 demonstrated that even in the absence of questions regarding the intuitive likelihood of the five being saved, differences along this dimension could result in divergent permissibility ratings. Study 5 replicated this result while presenting each participant with only one member of a pair of scenarios and with a binary yes/no permissibility judgment, suggesting that our findings hold even when participants are not exposed to both versions of a scenario, without any sensitization due to asking about intuitive likelihoods, and with a response measure that better reflects most philosophers' view of the non-scalarity of permissibility. Study 6 explored how two other probabilities present in many moral dilemmas—the likelihood of the one dying and of the five surviving without the protagonist's intervention—could likewise affect permissibility ratings. Study 7 extended our findings about the connection between intuitive likelihood and moral judgment, suggesting that the foreseen/intended distinction similarly implicates differential judgments of likelihood, and that this can have a corresponding effect on moral judgment.

These findings have implications that are both methodological and conceptual. It seems clear that people do not understand the scenarios in precisely the way they are intended. We have shown that interventions that we assert will save five people may not be seen as necessarily doing so. Nor will the steps taken to, perhaps, save those five necessarily harm the single individual. In fact, the gap between how scenarios are conceived and how they are understood may go beyond this. When we assert that five people are helpless on one of the tracks the trolley might travel, perhaps we are wrong about there being precisely five of them. And perhaps those five (or so) people, ones we have imagined to be just random innocents, are actually seen as complicit in their own fate—who, after all, hangs out on tracks down which trolleys can readily travel? Such a view is actually familiar to those who have posed moral scenarios to people only to have them quibble with the premises. "Why not yell ahead and alert the people on the tracks?" or "Perhaps the conductor should just put the switch in the middle position and safely derail the trolley." Such views are not irrational, though they do pose challenges to researchers. People are, generally, wise to take their prior probabilities into account in decision making, and, in other contexts, are often criticized for doing so insufficiently (Tversky & Kahneman, 1974).

When we ask participants whether it is worth sacrificing one to save five in detailed scenarios, it appears that many understand the questions in terms of relative risk. Differences between individuals in moral judgments may be due in some part to the extent to which they accept the assertions within the scenario rather than substitute their own intuitive probabilities, how they form the estimate of those probabilities, and then, to some extent, what intuitive moral theory they hold and apply. That last factor—the one of most direct interest to those investigating people's moral reasoning—is more difficult to isolate than one might hope. Simply asserting that everything else is fixed appears insufficient.

As an alternative explanation of our data, it might be suggested not that the intuitive probabilities that participants substitute for stipulations affect their moral judgments, but rather that participants' moral judgments affect the intuitive probabilities that they substitute for stipulations (see Liu & Ditto, 2013). Although some of our studies do not rule out this possibility, others speak against it. Consider, for example, Set 1 in Study 2, which consists of a pair of cases that differ only with respect to which body part of the one individual, neck or foot, will be severed by a trolley that has been diverted in his direction. Although participants are told in both cases that the one individual will die, they judge the likelihood of his dying as falling significantly short of 100% when what is severed is his foot (*M* = 79.8%), but as falling nearer to 100% when what is severed is his neck (*M* = 94.5%). If the difference in participants' judgments of degree of moral permissibility of diverting the trolley were driving the difference in their judgments of the likelihood of the one's death, then something other than the latter difference would need to be driving the former. But what could that be? Apart from the foot/neck difference, the two cases are exactly the same. Were there some *further* factual difference between the cases that might account for the difference in moral judgment (e.g., that in the one case the five are all children while in the other they are all adults, or that in the one case the protagonist made a promise to the five while in the other the protagonist did not), it might be possible to explain the difference in participants' moral judgments in a way that coheres with the assumption that this difference is itself driving the difference in participants' intuitive probabilities. But there is no such *further* factual difference in Set 1 of Study 2. We are, therefore, left with what seems in any event to be the most plausible explanation of participants' moral judgments in these cases: Participants find it easier to justify diverting the trolley at least in part *because* they believe that it is less likely that the one who will experience the severing of a body part will actually die.

Our studies have important implications for the fields of moral psychology and experimental philosophy. Until we can ensure that participants are not substituting their own intuitive probability estimates for experimenters' stipulations, we are not permitted to draw any unqualified conclusions from studies of moral judgments that are based on participants' evaluations of protagonist behavior in vignettes that are simply presented to them on paper or online. Our studies also suggest that we need to be very careful before drawing any conclusions about folk moral intuitions on the basis of moral conversations with people who may well refuse to accept stipulations about outcomes in hypothetical cases. As previously noted, some people (e.g., in classrooms) push back against the stipulations, and we often exclude their reactions from the data for this reason; but our results establish that many participants who do not *explicitly* resist the stipulations in fact resist them anyway (consciously or not). Future studies need to correct for the fact that many participants judge whether a protagonist's behavior is morally better or worse, (more or less) morally permissible or impermissible, on the basis of their own sense of how likely it is that certain events will occur or won't occur depending on whether the protagonist acts or does not act. Future work needs to address the best way of making this correction: One solution might be to write scenarios in which intuitive probabilities are matched, and another might be to find a way to present scenarios so that participants accept all the embedded stipulations.

If, as we suspect, participants routinely refuse to accept stipulations that conflict with their antecedent probability expectations, then some doubt might be cast on a wide variety of past vignette-driven experiments, designed to elicit not merely *moral* judgments but other philosophical judgments (e.g., about the compatibility of free will with determinism), causal judgments, linguistic judgments, mental state judgments, and more. The proper design of studies in vignette-driven psychology and experimental philosophy needs to be rethought.

There are at least two more questions raised by our studies that deserve further consideration. The first question is whether further intuitive probabilities not discussed in our studies play some role in participants' moral judgments about the actions of protagonists in hypothetical scenarios. Participants, for example, might think that there is a non-negligible likelihood of a runaway trolley posing a danger to more people than the six who are usually mentioned in a typical trolley case. They could also include their intuitive likelihood that the various protagonists are on the track in the path of a trolley through their own negligence. It might even be that participants' moral judgments are affected by their estimate of the *ex ante* likelihood of the hypothetical scenario itself. Shtulman and Tong (2013), for example, found that, as an individual difference variable, regarding extraordinary events as possible (i.e. having a non-zero likelihood) was associated with finding different actions as possibly being permissible. It would certainly be valuable to know just how many, and how extensively, antecedent probability estimates of different types affect moral judgment.

The second question concerns participants' judgments about the morality of risk. To discover which features of the world participants take to be morally significant (and how morally significant they take them to be), researchers ask them questions based on vignettes in which it is stipulated that such-and-such an action or omission *will* (in no uncertain terms) lead to so-and-so outcome (e.g., “If Sam throws a switch, the trolley *will* be diverted”). What, then, do participants really believe about what morality forbids or permits in the presence of greater or lesser risk? If, for example, participants are faced with a scenario in which the protagonist must, in order to save five, choose between imposing a near-100% risk of death on one or imposing a 50% risk of death on two (or, perhaps, a scenario in which the protagonist must choose between a course of action that has a near-100% chance of saving one and a course of action that has a 50% chance of saving two), what will they recommend? Scenarios typically stipulate fixed outcomes, but in the real world there is always uncertainty about the results of acting, or failing to act. An understanding of people’s moral reasoning should reflect this.

Chapter 1, in full, is a reprint of the material as it appears in Cognitive Science 2018. Ryazanov, Arseny; Knutzen, Jonathan; Rickless, Samuel; Christenfeld, Nicholas, Nelkin, Dana. The dissertation/thesis author was the primary investigator and author of this paper.

References

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*(2), 187-202.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in medicine*, *19*(22), 3127-3131.

Christensen, J., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Behavioral Reviews*. *36*(4), 1249-1264.

Cushman, F., Young, L., & Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgments: Testing Three Principles of Harm. *Psychological Science*, *17*(12), 1082–1089.

Foot, P (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293*(5537), 2105–2108.

Greene, J., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L., & Cohen, J.(2009). Pushing Moral Buttons: The Interaction Between Personal Force and Intention in Moral Judgment. *Cognition, 111*(3), 364-371.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & language*, *22*(1), 1-21.

Kortenkamp, K. V., & Moore C. F. (2014). Ethics under uncertainty: the morality and appropriateness of utilitarianism when outcomes are uncertain. *American Journal of Psychology*, *127*(3), 367-382.

Liu, B. S., & Ditto, P. H. (2013). What dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science*, *4*(3), 316-323.

Lockhart, T. (2000). *Moral Uncertainty And Its Consequences*, New York: Oxford University Press.

McIntyre, A.(2004). Doctrine of Double Effect, The Stanford Encyclopedia of Philosophy (Winter 2014 Edition), Edward N. Zalta (ed.). Retrieved from http://plato.stanford.edu/archives/win2014/entries/double-effect/

Mikhail, J. (2000). Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice. Cornell University PhD dissertation.

Moore, A., Clark, B., & Kane, M. (2008). Who shalt not kill?: Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557.

Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous* 41(4), 663-685.

Peterson, M. (2013). *The Dimensions of Consequentialism: Ethics, Equality and Risk*, Cambridge: Cambridge University Press.

Rawls, J. (1971). *A Theory of Justice*, Cambridge, MA: Harvard University Press.

Royzman, E. B., & Baron, J. (2002). "The Preference for Indirect Harm." *Social Justice Research* 15(2), 165-184.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Shtulman, A., & Tong, T. (2013). Cognitive parallels between moral judgment and modal judgment. *Psychonomic Bulletin & Review*, 20, 1327-1335.

Thomson, J. J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist 59*(2): 204-217.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. Journal of Statistical Software , *59*(5).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Chapter 2: Beyond Killing One to Save Five: Sensitivity to Ratio and Probability in Moral Judgment

Arseny A. Ryazanov, Tinghao Wang, Dana Kay Nelkin, Nicholas J. S. Christenfeld,

Samuel C. Rickless

University of California San Diego

Correspondence concerning this article should be addressed to Arseny Ryazanov,

Department of Psychology, University of California, San Diego, La Jolla, CA 92093

Contact: aryazano@ucsd.edu

Abstract

Some would support killing one innocent person to save five, while others would argue that such acts are wrong, regardless of their maximizing utility. The research focus on moral dilemmas concerning tradeoffs between one and five lives may obscure more nuanced commitments involved in moral decision-making. Six studies (N = 2177) examine the role of the actual numbers of lives involved, and the impact of making the outcomes probabilistic rather than certain. Study 1 examines the extent to which people are sensitive to the ratio of lives saved to lives ended by a particular action. Study 2 verifies that the ratio rather than the difference between the two values is operative. Study 3 examines whether participants treat probabilistic harm to some as equivalent to certainly harming fewer, when expected values are held constant, exploring whether prospect theory can account for responses to moral dilemmas. Study 4 compares sensitivity to probabilistic outcomes and expected value under single and joint evaluation. Study 5 investigates whether participants are willing to pay an expected value premium to distribute risk of harm among more individuals rather than certainly harm fewer under joint evaluation. Studies 6 and 7 explore an analogous issue regarding the sensitivity of probabilistic saving. Participants are remarkably sensitive to expected value for probabilistic harms under single evaluation, but deviate from it under joint evaluation. Participants deviate from expected value for probabilistic saving even under single evaluation. Collectively, the studies are consistent with the view that people's moral judgments are based on the principle of threshold deontology, and raise the question of whether joint or single evaluations are normatively correct.

Keywords: morality; probability; risk; decision-making; moral dilemma

Beyond Killing One to Save Five: Sensitivity to Ratio and Probability in Moral Judgment

A much-studied moral dilemma arises when an individual has a choice to save a relatively large number of people by killing a relatively small number of people. Is it permissible, for example, to kill one to save five? When moral philosophers first introduced such cases (Foot, 1967), it was with the aim of supporting non-consequentialist moral theory. Foot made an empirical prediction that people would reject the idea that it is permissible to kill one to save five, in service of the normative claim that considerations *in addition* to mere consequences are essential to answering moral questions.

While this was certainly not the end of the story, dilemmas of the same general form as Foot's original example have had tremendous staying power. In particular, with some notable exceptions we discuss below, the 1:5 ratio of the classic cases has been a mainstay in both moral theorizing and experimental moral psychology, and dilemmas are presented as though the stipulated outcomes of possible actions are certain to occur, unlike in many real-world situations. Further, dilemmas are typically presented in single, rather than joint form (e.g., asking participants whether, in order to save two, one ought to impose a 100% risk of one dying, rather than asking participants whether, in order to save two, it is better to impose a 50% risk of two dying than a 100% risk of one dying). In this article, we depart from each of these three formal features of the original dilemma in order to better understand just what moral considerations people bring to bear in a whole range of moral dilemmas, including real-world ones.

Ratio and Moral Judgment

Systematic study of participants' responses to classic moral dilemmas called into question the original prediction that most people reject the permissibility of killing one to save five, at least once confounding factors are removed. Recent studies on dilemmas of this kind have suggested that people's moral judgments can systematically vary according to a variety of factors, such as the existence of physical contact (Cushman, Young, & Hauser, 2006), the intentional structure of the action (Hauser, Cushman, Young, Kang-Xing, & Mikhail, 2007; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006), individual differences on working-memory-capacity tasks (Moore, Clark, & Kane, 2008), and personality traits (Arvan, 2013).

Empirical researchers have proposed several dual-process theories of moral cognition, according to which judgments in moral dilemmas are based on two competing processes: an outcome-based (or model-based) process responsible for "consequentialist" judgments (which rest on the idea that only consequences matter to deciding questions of moral permissibility), and an action-based (or model-free) process responsible for "non-consequentialist" judgments (which rest on the idea that considerations in addition to consequences matter (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene & Haidt, 2002; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Green et al., 2009; Haidt, 2001; Cushman, et al., 2006; Cushman, Young, & Greene, 2010; Cushman & Greene, 2011; Paxton, Ungar, & Greene, 2012; Cushman, 2013; Crockett, 2013). Though outcome-based and action-based processes have typically been construed as "winner take all" processes (e.g., Cushman, 2013), recent work suggests that the distinct brain regions responsible for each of these processes (e.g., Shenhav & Greene, 2010; 2014) are jointly integrated in moral judgments (Hutcherson, Montaser-Kouhsari, Woodward, & Rangel, 2015; Cohen & Ahn, 2016).

Many studies, however, assume that dual processes must reflect two distinct, competing moral principles: absolutist consequentialism (according to which only the consequences of action and inaction matter; see, e.g., Sinnott-Armstrong, 2015) and absolutist non-consequentialism (according to which certain non-consequentialist constraints, e.g., constraints against killing or against intending death, are firm and exceptionless; see, e.g., Quinn, 1989; Rickless, 1997). This oversimplification can also be seen in researchers' tendency to categorize the moral judgments into being either simply "consequentialist" or "non-consequentialist." (See Kahane, Everett, Earp, Farias, & Savulescu, 2015, who have challenged this tendency, but based on reasons different from the ones we provide, as well as Conway, Goldstein-Greenwood, Polacek, & Green, 2018; Conway & Gawronski, 2013).

What this dichotomy obscures is that those opposed to killing at a particular ratio may in a principled way shift to endorsing killing at a higher ratio of good done and harm done. For example, of those who reject killing one to save five, it may be that some endorse killing one to save ten. And of those who accept killing one to save five, some may reject it when the number killed grows to four. Such shifts would support the view that people in fact reason according to a more nuanced principle of morality than absolutist consequentialism or absolutist non-consequentialism, and would be consistent with the integration of outcome-based and action-based processes observed by Hutcherson et al. (2015).

A philosophical framework that accounts for such shifts is threshold deontology (see Moore, 1997). On this view, there is a threshold of good that needs to be achieved beyond which it is morally permissible to infringe a non-consequentialist moral constraint. For many threshold deontologists, the threshold is not a fixed amount of good, but rather varies according to, among other things, the ratio of good achieved (or expected) to harm done (or expected). That is, there is a general moral constraint against killing, but this constraint is overridden when the consequences of inaction are bad enough. Someone who finds it morally unacceptable to kill four to save five (or to kill 99 to save 100), but who finds it morally acceptable to kill one to save five (or ten, or twenty), exhibits judgments that are consistent with threshold deontology, but not with absolutist consequentialism or absolutist deontology (see Alexander & Moore, 2016).

Cohen and Ahn (2016) have proposed an alternative candidate to explain people's moral judgments, namely, subjective utilitarian theory, as a single process underlying people's judgments in moral dilemmas. The theory states that people choose the option that brings the maximal amount of personal value, with personal value purposefully left underspecified (Cohen & Ahn, 2016). It is possible that even if threshold deontology is the correct moral theory, it is not what is operative in actual moral decision-making. While a full comparative evaluation between subjective utilitarian theory and threshold deontology on explanatory grounds is not possible here, we believe, on the basis of the studies described below, that threshold deontology can better explain a persisting asymmetry in participant responses regarding doing and merely allowing harm. At the same time, we take this to be consistent with surprising results in how individuals value outcomes as Cohen and Ahn's data suggest.

Threshold deontologists often focus on the permissibility of overriding moral constraints in extreme cases, where the only available alternative is catastrophic (Nozick 1974; Fried, 1978; Nagel, 1979; Moore, 1997; though see Thomson, 1990 and Brennan, 1995). An example would be the killing of one person in order to avoid the destruction of a large city or an entire nation. This deviation from the paradigmatic 1:5 ratio also appears in psychological research (Nichols & Mallon, 2006; Bartels, 2008). The underlying assumption seems to be that the expected value of the ratio doesn't matter unless the result of inaction amounts to a catastrophe. In the studies described below, we explore the power of threshold deontology to explain participants' responses to a variety of dilemmas. To do this, we present dilemmas with a variety of different ratios, from very high to very low.

Expected value, ratio, and probability

Most studies on moral dilemmas have focused on actions whose outcomes are described as certain to happen (or at least not described in terms of probabilities of anything less than 100%). In real life, however, we rarely know with certainty what will happen if we act one way rather than another, and often work with probabilities somewhere between 0 and 1. Moreover, in psychological research, even when participants are told that outcomes are certain, there is evidence that they often substitute their own probability estimates of less than 100% for outcomes that are described as certain (Ryazanov, Knutzen, Rickless, Christenfeld, & Nelkin, 2018; Shou & Song, 2017). So, in addition to departing from the 1:5 ratio and from the focus on catastrophic alternatives, we also depart from the presumption of certainty to test whether participants treat equivalent expected values similarly when probabilities of harm or rescue differ.

Recent research suggests that moral judgments are sensitive to expected value. Tassy, Oullier, Mancini, and Wicker (2013), and Shenhav and Greene (2010) found that participants were sensitive to the number of people who could be saved in dilemmas where participants could choose one of two groups to save. However, neither of these tasks constituted sacrificial moral dilemmas, in that they involved choosing to benefit one party at the expense of another (picking between positive outcomes), rather than imposing harm on one party for the benefit of another. Whether participants exhibit sensitivity to expected value (outcome) in actions where harm is caused remains unclear.

There are still further questions regarding the relationship between harm and benefit in moral dilemmas: How is the ratio of good done to harm done to be defined? And does the difference between the amount of harm and good done, or ratio of good done to harm done, matter to moral judgment? Shenhav and Greene's (2010) studies showed a relationship between expected value and nonsacrificial moral dilemmas, but whether the operative factor was the difference in the number of lives saved or the ratio of lives saved to lives lost was not tested. Mikhail (2011) hypothesizes that participants' moral grammar includes a "moral calculus of risk", which treats the justifiability of imposing a risk of unintentional harm as a means of knowingly producing a good effect as a matter of whether the probability of producing the harm multiplied by the value of the harm is morally worse than the product of three factors: (i) the probability of knowingly producing the good effect, (ii) the probability that the good effect would not be achieved without risk of harm, and (iii) the value of the good effect. But Mikhail (2011) does not test whether the moral calculus of risk governs participants' judgments. In our studies, we test whether the expected difference between, or the expected ratio of, good and bad plays a role in moral judgment. We define the ratio here in a way that incorporates the number of people who might be saved and the probability that they will be saved, as well as the number of people at risk of being killed and the probability that they will be killed:

$$R = EV_{\text{lives saved}} / EV_{\text{lives ended}} = N_{\text{lives saved}} \times P_{\text{lives saved}} / N_{\text{lives ended}} \times P_{\text{lives ended}}$$

Researchers have also started to study the role of probability in moral reasoning. Fleischhut, Meder, and Gigerenzer (2017) found that moral judgments when outcomes are certain to occur differ from when those outcomes are uncertain, though without specifying any probability for the outcomes' occurrence. In addition to varying the number of lives that could be saved, Shenhav and Greene (2010) varied the probability that the latter group of people do not actually need saving, (e.g. the probability that a group of people blocked in an office building will successfully escape anyway). There are two different probabilities at stake as well, which they didn't study: the probability that the plan of saving them will be successful, and the probability that the plan will kill a number of people. Do these probabilities matter? How will they interact with the role of expected value? These questions remain unanswered.

Probability in moral reasoning may have a similar structure to the more general phenomenon of risk-seeking and risk-aversion. In many domains people systematically deviate from expected value, as elaborated in prospect theory (Kahneman & Tversky, 2013; Tversky & Kahneman, 1992). Among other things, prospect theory predicts a "certainty effect" that "contributes to risk aversion in choices involving sure gains and to risk seeking in choices involving sure losses" (Kahneman & Tversky, 2013, p. 263). Are people sensitive to the certainty effect, showing risk-averse or risk-seeking tendencies when it comes to lives in moral dilemmas? One way to explore this issue is to keep the expected values fixed, while varying the probability of killing and the number of people killed at the same time. Alternatively, one might keep the expected values fixed while varying the probability and number of people saved. Prospect theory suggests that people will be sensitive to probability while expected values are held constant. That is, if people are risk seeking for losses, they might be willing to accept a greater expected number of people being killed when the harm is probabilistic rather than certain. And if they are risk averse for gains, they might need a greater expected number of people being saved in order to be willing to harm the same number of people when the saving of lives is probabilistic rather than certain. Some indication comes from work on variants of the "asian disease problem", in which participants must decide between certain losses of life and probabilistic losses of life, as well as between certainly saving a group of individuals and probabilistically saving a group of individuals (Tversky & Kahneman, 1981). While typically used to explore the framing effect of losses and gains, these non-sacrificial dilemmas suggest that probabilities and expected value may matter in sacrificial moral dilemmas, and that, furthermore, probabilistic harm and saving may be treated differently (Diederich, Wyszynski, & Ritov, 2018). It remains to be seen whether sacrificial moral dilemmas treat harm as a loss, and the benefit as a gain, and, if so, whether participants are risk seeking for harm and risk averse for probabilistic saving.

Thus, this paper has two aims: (1) to test the hypothesis that participants exhibit judgments consistent with threshold deontology, rather than absolutist consequentialism or absolutist deontology; (2) to systematically examine the role of expected value and probability in moral judgment, which further involves (i) testing how the expected number of people being saved and the expected number of people being killed interact with each other, and (ii) testing how varying the probability of harming and successfully saving affects participant responses.

Single Versus Joint Evaluation

People's responses to moral uncertainty can be explored in single evaluation scenarios. Such cases would involve asking people whether someone should act when one of the outcomes is specified as uncertain, e.g., whether one should impose a 50% risk on two people in order to save five. They can also be studied in joint evaluation. This would involve asking people which of two options is preferred when they vary in the level of uncertainty, e.g., whether, in order to save five, it is better to impose a 100% risk on one person or a 50% risk on two people. According to general evaluability theory, evaluations made between simultaneously available alternatives can differ from evaluations of each alternative made separately (Hsee, 1996; Hsee & Zhang, 2010; Hsee, Loewenstein, Blount, & Bazerman, 1999). For example, participants may express equal support for preventing the extinction of Australian mammals and preventing skin cancer among farm workers when asked about just one or the other, despite strongly prioritizing the latter when choosing which to support (Kahneman & Ritov, 1994). Recently, Barak-Corren, Tsay, Cushman, and Bazerman (2018) found that participants' preference between two versions of the trolley problems—one in which participants decided whether to drop one person onto a trolley track to save others ahead or divert the trolley onto a track with one person on it to save others ahead—is reversed when the evaluations are made jointly, rather than individually. We thus explore the role of expected value, and deviations from it when harm and saving are probabilistic rather than certain, under both single and joint evaluation.

The Studies

Study 1 examined the extent to which people are sensitive to specific numbers involved on the harming side and helping side. Study 2 explored whether the expected value ratio, and not, for example, the difference between the number of lives saved and the number of lives lost, is operative in participant sensitivity to expected value, consistent with the hypothesis of threshold deontology. Study 3 examined the extent to which people are sensitive to ratio and the probability of the harm ($P_{\text{lives ended}}$). Study 4 further explored the sensitivity to the probability of harm under both single and joint evaluation. Study 5 explored whether participants are willing to increase the expected amount of harm done in order to distribute risk of harm among more individuals. Finally, Studies 6 and 7 explored analogous issues regarding the sensitivity to the probabilities on the saving side ($P_{\text{lives saved}}$).

Study 1

Study 1 Materials and Participants

Participants read one of four scenarios, in which a group of people could be saved from a terrorist attack by harming a bystander. As an example, one of the scenarios goes as follows:

Harry has a plan to save two people from a terrorist attack. For Harry's plan to work it is necessary that Robert, a bystander, is killed. If Harry carries out this plan, he will certainly save the two people, but in doing so will certainly kill the bystander.

Participants saw a version of the scenario where one would be killed to save two (ratio (R) = 2), one would be killed to save one hundred (R = 100), one would have a 50% chance of being killed to save 50 (R = 100), or 99 would be killed to save 100 (R ≈ 1.01). Participants responded to a single question regarding their confidence that the action should be carried out, adapted to each scenario, along an eleven-point scale. For example, participants were asked, *Should Harry carry out a plan that he knows will certainly save two people, but will certainly kill a bystander?* (-5: *very confident he should not carry out the plan*, to 5: *very confident he should carry out the plan*). We chose this dependent variable, rather than a continuous forbidden-obligatory measure, because according to many consequentialists and non-consequentialists, actions cannot be *more or less* morally right or wrong (see Ryazanov et al., 2018), and questions that don't explicitly ask about confidence (e.g., agree-disagree) implicitly do by asking for degree of agreement/disagreement. We used a single scale measure, rather than process dissociation as endorsed by Conway & Gawronski (2012), because we are interested in whether judgments change at different expected values, rather than in the strength of deontological and utilitarian "inclinations" that could contribute to moral judgments. One hundred and twenty-two participants were recruited via Amazon's Mechanical Turk (110 passed an attention check and were retained for analysis; 56.4% female; mean age = 32.9, SD = 8.30).

Study 1 Results

We began by examining how participant responses to whether the action should be carried out correspond to the action's expected value ratio. An ANOVA, with expected value ratio entered as a continuous factor, yielded a significant effect of expected value on moral judgment, $F(1,108) = 35.4$, $p < .001$, $r^2 = .25$ (mean R 1.01 = -2.23, SD = 2.67; mean R 2 = .11, SD = 2.45; mean R 100 = 2.12, SD = 2.79 collapsing across both versions of the ratio; see Figure 1). Thus, participants were sensitive to the expected value ratios presented to them. Next we examined whether different forms of the same expected value ratio showed different preferences, since we had two scenarios with R = 100. A planned comparison of a 50% chance of killing one to save 50 and a 100% chance of killing one to save 100 did not reveal a significant difference: $t(55) = .218$, $p = .83$, $d = .06$ (mean 1v100 = 2.20, SD = 2.93; mean 50%1v50 = 2.04, SD = 2.70).

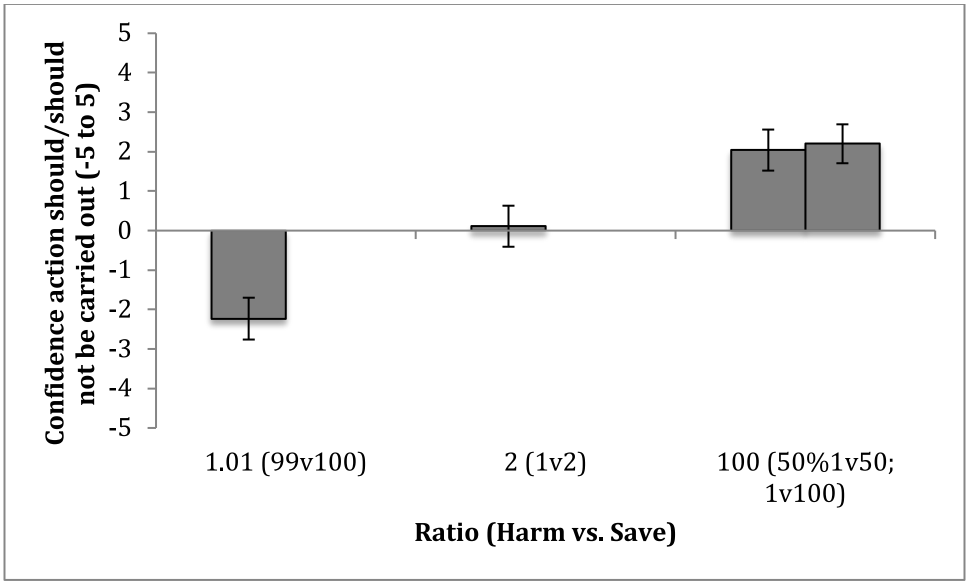


Figure 2.1. Sensitivity to expected value ratio of lives saved to lives lost. Error bars represent one standard error.

Categorizing responses into "would not act" (responses < 0) and "would act" (responses > 0), revealed that while 80% would act in the 1v100 dilemma, only 48% would do so in the 1v2 dilemma, $\chi^2(N = 51) = 5.40$, $p = .020$, *Cramer's V* = .33.

Study 1 Discussion

Study 1 found evidence for sensitivity to expected value ratio. Participants more often endorsed an action that harmed one to save others when the ratio regarding the expected value of lives saved to the expected value of lives lost was larger. This sensitivity suggests that people are neither absolutist deontologists nor consequentialists, instead making decisions based on the principle of threshold deontology.

Study 1 also includes preliminary evidence regarding sensitivity to probabilistic forms of the same expected value ratio. People treated the two scenarios whose ratios were identical no differently, despite one of them involving a probabilistic harm and one a certain harm. In this instance at least, moral uncertainty doesn't have any impact independent of the expected value ratio on people's application of their ethical principles. Contrary to prospect theory's prediction that participants would prefer a chance of harm to the one over the certain death of one, to save a matched number of people (50 and 100, respectively), we did not see evidence for this variation in single evaluation. It could be that such a preference emerges in joint evaluation, however, when two such plans are compared side by side. We tested this hypothesis in Study 4.

But first, we note that the results of the first study suggest that it is ratio that matters when it comes to harming some to save others. It is possible that such decisions could be made not on the ratio, but instead on the number of lives gained. That is, sacrificing 99 to save 100 involves a net gain of one life, and in this way is similar to sacrificing one to save two. However, our data indicate that the latter option is regarded much more favorably than the former. In Study 2, we set out to test more directly the hypothesis that it is in fact ratio that is operative.

Study 2a

Study 2a Materials and Participants

Study 2a sought to verify that the ratio of lives saved to lives lost, rather than the difference between the two expected values, was operative. Each participant read about one of three plans, which were identical to those of Study 1 except for the numbers involved. The plans involved killing 1 to save 2 (difference = +1 life), killing 10 to save 20 (difference = +10 lives), or killing 100 to save 200 (difference = +100 lives). One hundred and fifty-eight participants were recruited via Amazon's Mechanical Turk (135 passed an attention check and were retained for analysis; 63.7% female; mean age = 34.2, SD = 10.8).

Study 2a Results

We examined differences in responses to each of the three scenarios. If people favor the action more when it saves more net lives, there should be a sharp increase in supporting the action as the net number goes from +1 to +10 to +100 lives. If, instead, people are sensitive to the ratio, then the three scenarios should be regarded as effectively identical, with the action in each one saving twice as many as are sacrificed. An ANOVA, with difference in number of lives saved / lost entered as a continuous factor, confirmed that there was no significant difference between the scenarios, $F(1, 133) = .50$, $p = .48$, $\eta^2 = .004$ (Mean 1v2 = -.16, SD = 3.10; Mean 10v20 = 0, SD = 2.63; Mean 100v200 = .27, SD = 2.60; see Figure 2a).

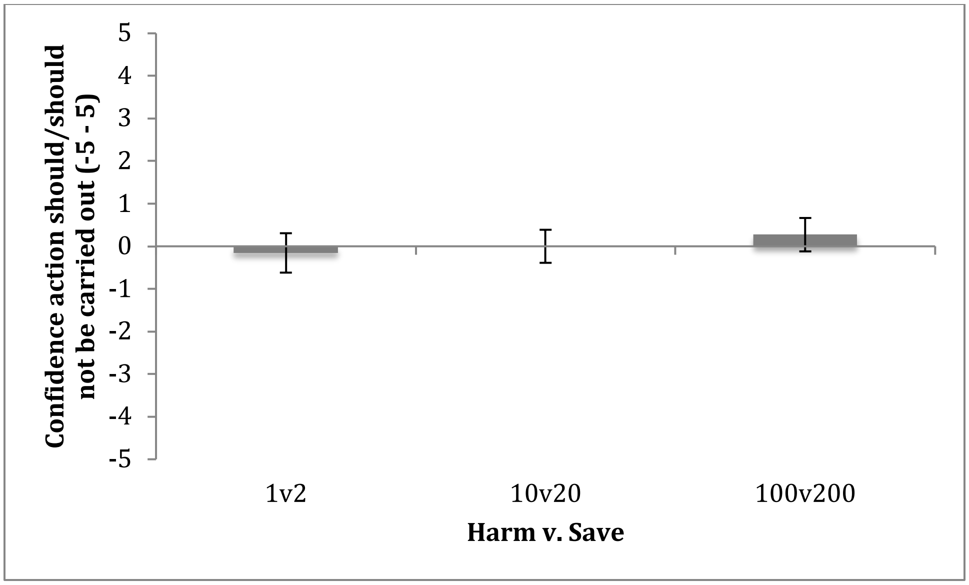


Figure 2.2.a. Sensitivity to ratio of lives saved to lives lost, rather than difference between lives lost and saved, in abstract scenarios. Error bars represent one standard error.

Study 2a Discussion

The lack of sensitivity to the difference between the number of lives lost and the number of lives saved suggests that participants in Study 1 were sensitive to ratio, instead of the net gain in lives from the action (or, even the raw number killed or saved). The scenario we used was a fairly abstract one, which, while specifying the numbers involved, did not flesh out how the fewer would die, nor how that would save the many. We sought to replicate our effects also with more detailed scenarios.

Study 2b

Study 2b Materials and Participants

Concrete scenarios were created in which the ratio of lives saved to lives lost was held constant, though the difference in numbers between the two groups varied, to examine whether it is the ratio of lives saved to lives lost that is operative, rather than the difference in the expected values. Each participant again read about one of three plans: killing 1 to save 2, killing 10 to save 20, or killing 100 to save 200, though now, instead of more abstract plans, we utilized a more detailed scenario that involved setting off an explosion to prevent a rocket from reaching a house. Subjects in the condition pitting sacrificing ten against saving twenty read the following scenario:

A missile has been accidentally fired at a house with 20 people in it. Bob is in charge of a missile defense tool that can destroy this missile by firing a rocket that can automatically detect the missile's location. The rocket will incapacitate the missile by setting off an explosion in the air near it. As Bob knows, the rocket's explosion near the missile will disable the missile, but will also kill 10 people standing in a field over which the missile will be intercepted. Firing the rocket given the timing and flight path of the missile is the only available option to prevent the missile from continuing on its path to the house with 20 people in it. Bob also knows the following facts. If Bob does not intervene, then the missile will certainly hit the house and kill all 20 people in it; if Bob intervenes, then the rocket Bob can set off will certainly destroy the missile and spare the people in the house, but will certainly kill the 10 people in the field.

Participants were asked, *Should Bob set off a rocket that he knows will kill 10 people, but that he also knows will destroy a missile that will otherwise kill 20 people?* (-5: *very confident he should not set off the explosion,* to 5: *very confident he should set off the explosion*). One hundred and forty-nine participants were recruited via Amazon's Mechanical Turk (129 passed an attention check and were retained for analysis; 54.7% female; mean age = 34.2, SD = 9.78).

Study 2b Results

As in the case of abstract scenarios, an ANOVA, with difference in number of lives saved / lost entered as a continuous variable, confirmed that there was no significant difference between any of the more concrete scenarios, in which the expected value ratio was held constant, but the difference between the number of lives saved and the number of lives lost was varied, *F*(1, 127) = .35, *p* = .55, η^2 = .003. (Mean 1v2 = 1.96, SD = 2.41; Mean 10v20 = 2.14, SD = 2.65; Mean 100v200 = 1.75, SD = 2.46;, see Figure 2b).

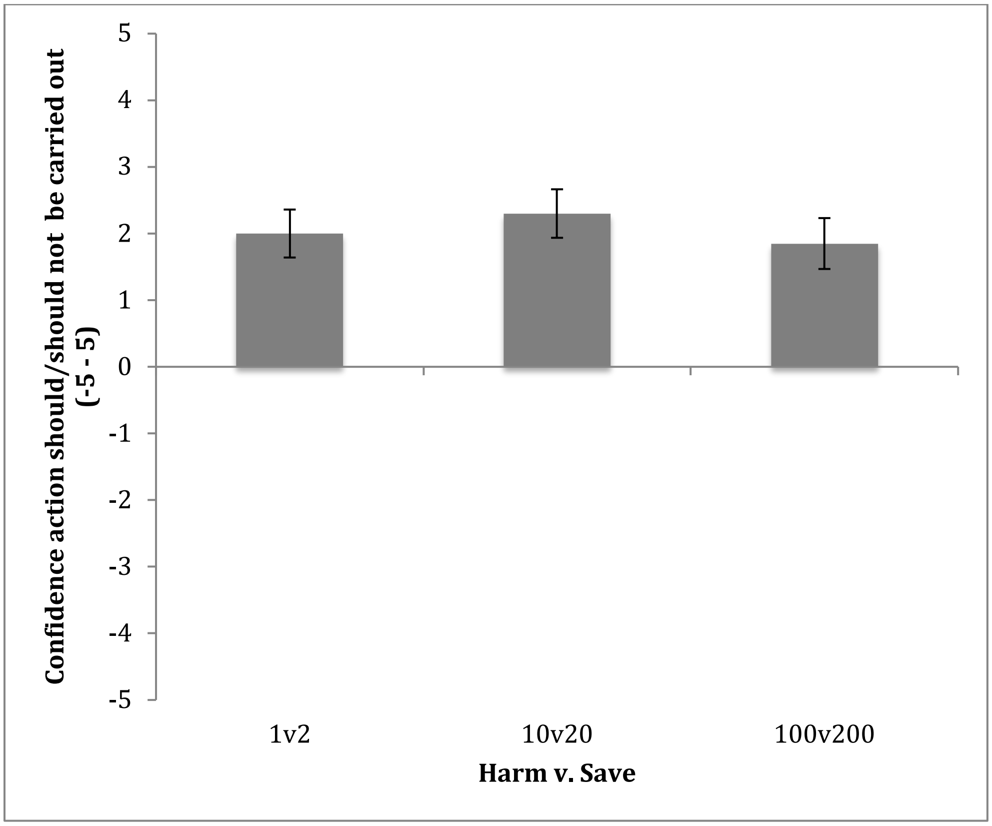


Figure 2.2.b. Sensitivity to ratio of lives saved to lives lost, rather than difference between lives lost and saved, in concrete scenarios. Error bars represent one standard error.

Study 2b Discussion

Again, participants remained insensitive to the difference between the number of lives lost and the number of lives saved, suggesting that participants are sensitive to ratio, rather than to difference in number of lives saved/lost or to some combination of the two. Consistent with studies that find participants to be more willing to act in concrete than in abstract situations (Agerström & Björklund, 2009; Amit & Greene, 2012), an exploratory analysis revealed that participants expressed greater confidence in action in the concrete scenarios of Study 2b than in the abstract scenarios of Study 2a, $t(262) = 6.30$, $p < .001$, $d = .78$. Sensitivity to ratio, rather than difference, was apparent in both kinds of scenarios.

Thus far, we have found participant responses to be consistent with threshold deontology. Next, we examine how such judgments incorporate probability, beginning with comparing probabilistic harm to certain harm.

Study 3

Study 3 Materials and Participants

Study 3 examined the effect of probability across three expected value ratios: R = 5, R = 2, and R = 1.25. Probabilities of the people being sacrificed were varied, with expected value remaining constant. Probabilities explored included 1%, 10%, 20%, 50%, 80% (only for 4v5), and 100%. We again utilized the more detailed scenarios from Study 2b, which involved setting off an explosion to prevent a rocket from reaching a house. In the R = 5 cases, for example, participants were asked, *Should Bob set off a rocket that he knows will have a X% chance of killing Y persons* [where XY=100]*, but that he also knows will destroy a missile that will otherwise kill 5 people?* (-5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*). 706 participants were recruited (616 passed an attention check and were retained for analysis; mean age = 36.5, SD = 12.0, 61.1% female). Each subject rated only one scenario, and provided brief demographic information.

Study 3 Results

As in Study 1, an ANOVA with probability and ratio entered as continuous factors revealed a significant effect of expected value ratio, regardless of the probability of harm, $F(1, 612) = 42.5$, $p < .001$, $\eta_p^2 = .065$ (mean 1v2 = 1.86, SD = 2.45; mean 1v5 = 2.97, SD = 2.13; mean 4v5 = 1.41, SD = 2.75, see Figure 5a). However, there was no significant effect of probability while keeping expected value matched, $F(1, 612) = .22$, $p = .64$, $\eta_p^2 < .001$, nor was there an interaction of probability and ratio, $F(1, 612) = .01$, $p = .94$, $\eta_p^2 < .001$, see Figure 5b. There was no sensitivity to probability for any of the individual ratios (all p's > .22). This suggests that, again, under single evaluation participants were sensitive to expected value ratio, but not to the probability of harm, even when the uncertain harm covered the range down to a 1% chance of occurrence.

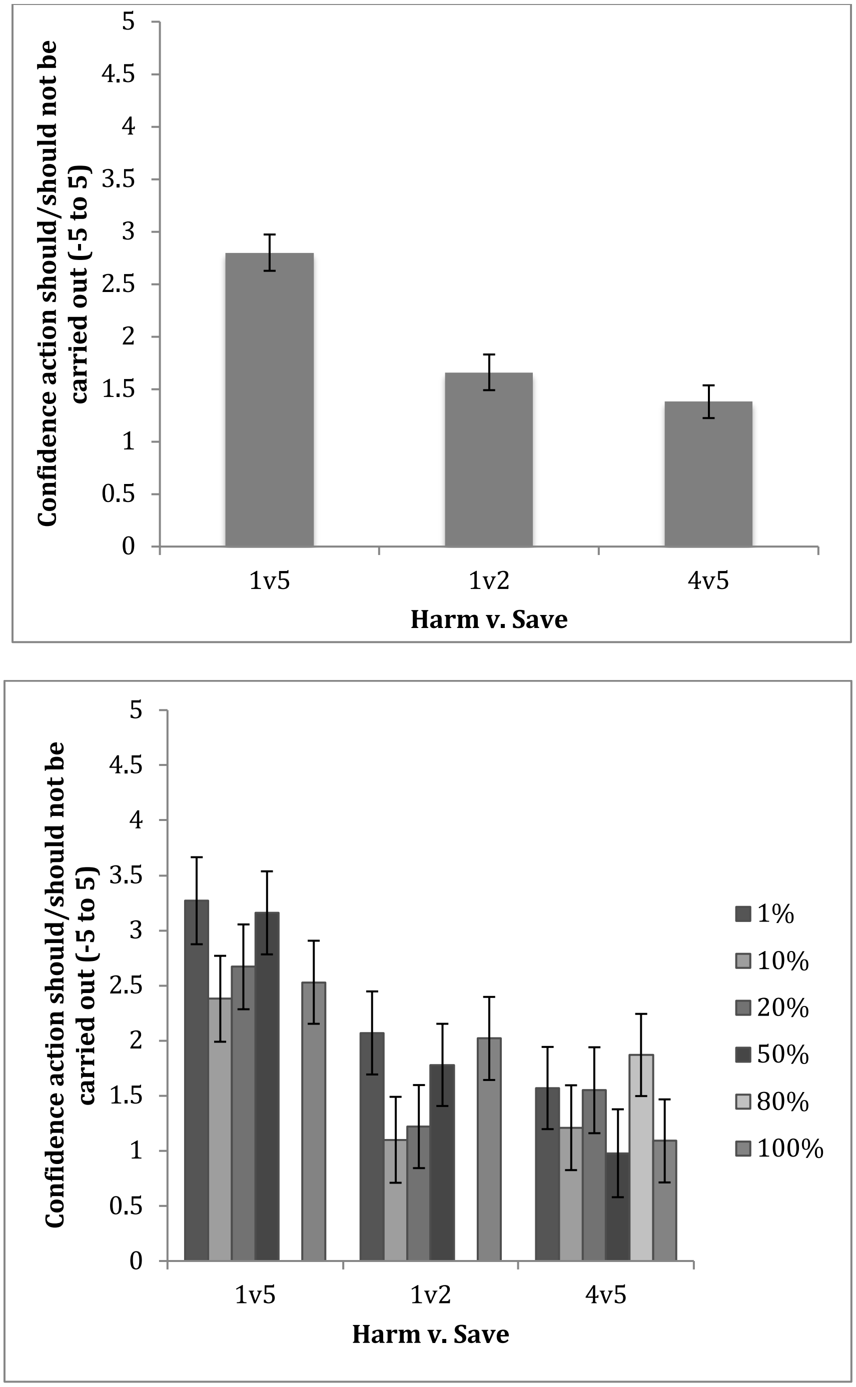


Figure 2.3. Insensitivity to various forms of the same expected value ratio of lives saved to lives lost in more concrete scenarios when probability of harm is varied (a). Sensitivity to expected value ratio (b). Error bars represent one standard error. 80% harm only tested for 4v5 because other ratios cannot achieve it with whole numbers.

Study 3 Discussion

Study 3 showed that people are sensitive to expected value, even across a fairly subtle range. In single evaluation, even with increased power from a larger number of participants, we continued to see no clear relationship between probability of harm, when expected value is fixed, and moral judgment. Regardless of how the expected value was presented to participants (e.g. 1% chance of killing 100 to save 5, or 1 certainly being killed to save 5), participants remained sensitive to the value. That participants shift judgments when the expected value of the saving changes from 2 to 5 suggests that, rather than making exceptions to deontological thinking only in catastrophic cases, people instead incorporate expected value into their moral calculation, and, as expected value increases, linearly shift from responses that disapprove of sacrifice to responses that approve of it. This is consistent with the view that our participants, in actuality, hold an ethical view that is consistent with threshold deontology.

Study 4

Study 4 expanded on the findings from Studies 1 and 3, that people treat an action that is 50% likely to harm N people as equivalent to an action that certainly harms N/2 people. Though the single evaluation of these actions produced remarkably similar moral judgments, it could be that participants nevertheless prefer probabilistic deaths over certain death when the two are jointly evaluated, as would be suggested by prospect theory, if lives harmed are perceived as potential losses.

Study 4 Materials and Procedures

Study 4 asked participants to evaluate one of two plans, as they had in Study 3. One plan was certain to kill one bystander, but would save two people from a missile that had accidentally been fired at them. The other plan involved a 25% chance of killing four bystanders to save the two from the missile. After evaluating one of the two plans, (e.g. Should Bob set off a rocket that he knows will kill 1 person, but that he also knows will destroy a missile that will otherwise kill 2 people?; -5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*), participants were exposed to both plans, labeled as Plan X and Plan Y, and were asked to compare them—*Which of the following two plans is morally better: firing Rocket X, which certainly kills 1 bystander and saves the group of 2 people, or firing Rocket Y, which has a 25% chance of killing 4 bystanders and saves the group of 2 people?* (-5 – 5; *Firing Rocket X is much better than firing Rocket Y – Firing Rocket Y is much better than firing Rocket X*). We chose to ask which plan is morally better for the joint evaluation, rather than asking participants for their level of confidence that the action should be carried out, because asking a confidence question in joint evaluation would imply that the agent must perform one of the two actions, when the agent could also choose not to act. Two hundred and five participants were recruited via Amazon's Mechanical Turk (179 retained for analysis, 63.7% female, mean age = 35.5, SD = 12.3).

Study 4 Results

Under single evaluation, there was no preference between a plan that risked a 25% chance of killing four to save two, and a plan that would certainly kill one to save two, *t*(177) = .17, *p* = .87, *d* = .025 (mean 100% = 2.02, SD = 2.42; mean 25% = 1.97, SD

= 2.10, see Figure 3a). However, under joint evaluation there was a significant preference for the plan that spread the risk across four individuals, over the plan that would certainly kill one, *t*(178) = 2.26, *p* = .025, *d* = .17 (mean = .53, SD = 3.12, see Figure 3b).

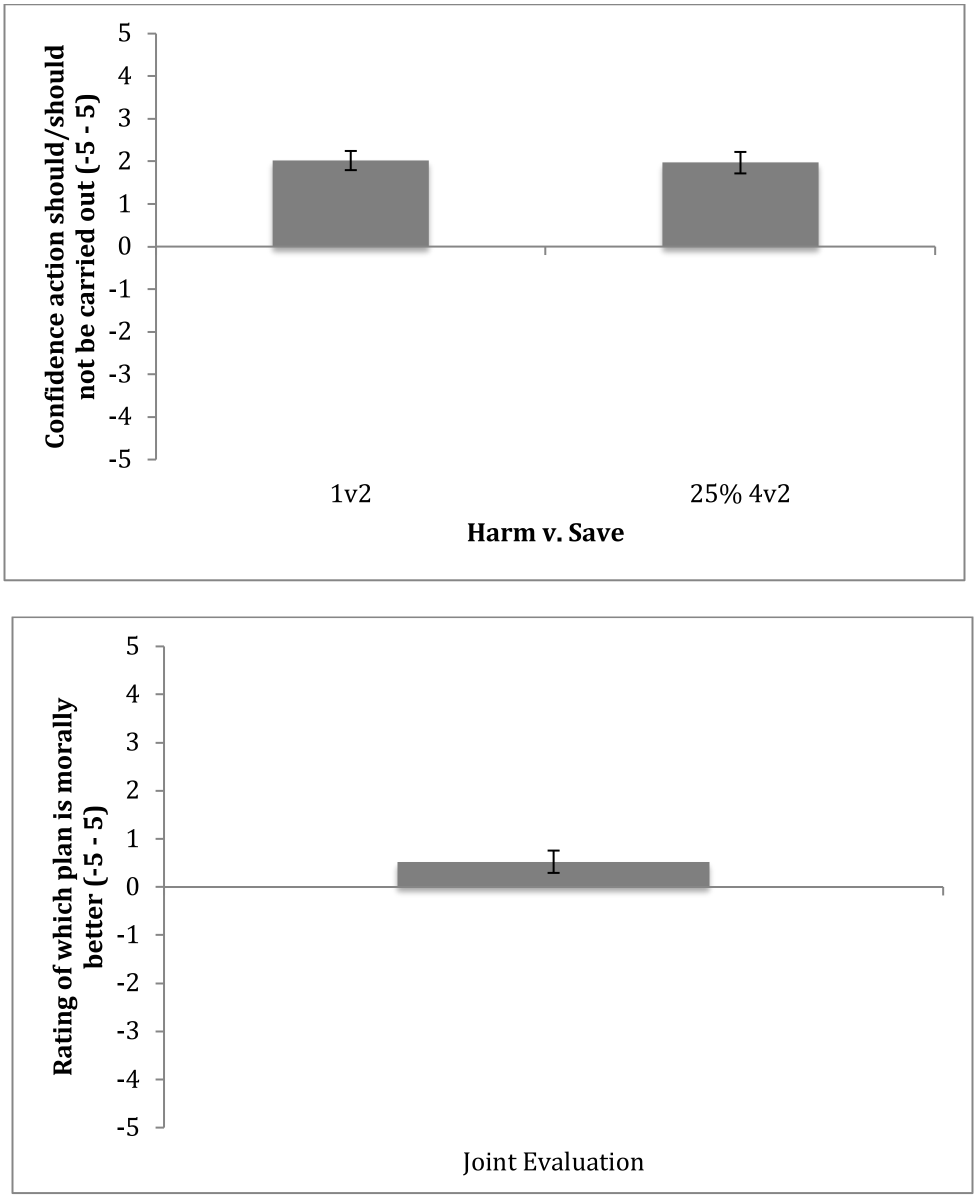


Figure 2.4. Insensitivity to whether harm is probabilistic or certain when expected value of lives saved to lives lost is fixed under single evaluation (a). Preference for probabilistic harm over certain harm when expected value is fixed under joint evaluation (positive ratings indicate a preference for probabilistic harm b). Error bars represent one standard error.

Study 4 Discussion

While participants did not favor distributing harm when making a single evaluation, in the joint evaluation they expressed moral preference for an action that saves two by risking a 25% chance of killing four, over an action that saves two by certainly killing one. Such a preference is consistent with risk seeking in the context of losses, as suggested by prospect theory.

In single evaluation, people do not exhibit risk aversion (or preference), and their judgment reflects just the mathematical expected value of the risk. However, in joint evaluation, a significant preference for risk emerges. People appear to be using different principles in the two cases. In the former, there is a weighing of the harm required against the good that will be done, and whether it is ethically justified. In the latter, when the action must be taken, the question of whether acting to save the two is morally permissible is no longer salient to the participant; instead, the most relevant question concerns whether it is morally better to sacrifice one for certain, or two with a 50% chance. Here the preference for spreading risk emerges.

People's use of different intuitive rules to decide the single and joint evaluation cases suggests that it should be possible not just to find two cases tied in the single evaluation with a preference in the joint evaluation, as we did in the current study, but also to construct cases tied in the joint evaluation with a preference in the single evaluation. We explored this in the next study. If people prefer risky harm, when the expected values are equal, they should be roughly indifferent if the expected value of the risky harm is suitably larger than the expected value of the certain harm. In the single evaluation case, where risk preferences did not emerge, they should find this action less tolerable.

Study 5

Study 5 Materials

Study 5 had the same design as Study 4, though with the expected value of the two plans not equated. Half of the participants were exposed to a scenario where the plan would kill one to save two (R = 2), and half were exposed to a scenario where the plan would risk a 25% chance of harming 8 to save 2 (R = 1). After evaluating one of the two plans, participants saw both plans, and were again asked which one was morally better. Two hundred and nine participants were recruited via Amazon's Mechanical Turk for participation in Study 4 (180 were retained for analysis after excluding participants who failed an attention check, mean age = 34.4, SD = 10.5, 61.7% female).

Study 5 Results

Participants indicated that the plan with certain harm was more acceptable than the plan with probabilistic (and twice as great) harm, *t*(178) = 2.12, *p* = .035, *d* = .31 (mean certain harm of 1 = 1.70, SD = 2.47; mean 25% chance of 8 harmed = .90, SD = 2.61, see Figure 4a). However, when both plans were compared directly, there was no preference between them, *t*(179) = .11, *p* = .91, *d* < .01(mean =-.028, SD = 3.28), despite the certain harm action having an expected value ratio of 2, and the probabilistic harm action having an expected value ratio of 1 (see Figure 4b). That is, the certain harm case involved saving twice as many as would be harmed, while the probabilistic harm case involved, as a matter of expected value, harming just as many people as the action would save, with no net gain in lives at all. Participants thus exhibited risk-seeking in joint evaluations of plans that have a probability of harming some to save others, and displayed a preference based on expected value ratio under single evaluation.

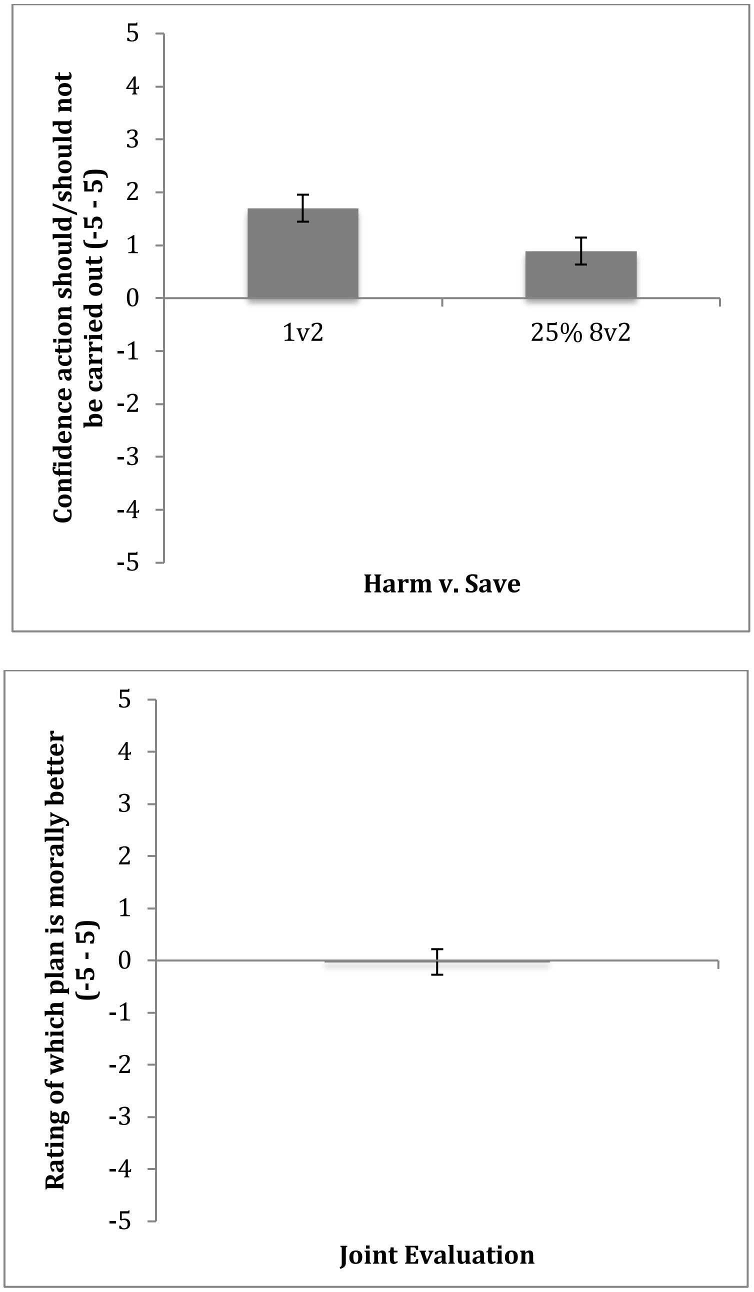


Figure 2.5. Sensitivity to expected value when value of lives saved to lives lost is not matched under single evaluation (a). Indifference between probabilistic harm with lower expected value and certain harm with higher expected value under joint evaluation (positive ratings indicate a preference for probabilistic harm; b). Error bars represent one standard error.

Study 5 Discussion

As in Study 4, participant preferences diverged in single and joint evaluations. Participants became indifferent when selecting between a plan that had an expected value ratio of 2 and a plan that had an expected value ratio of 1, because the plan with the lower expected value ratio spreads risk to more individuals, thus pitting sensitivity to expected value against a preference for spreading risk.

So far, while we have varied the number of people involved on both the harming side and the saving side of the dilemma, we have explored the effect of probability only on the harming side. We next explore how people treat ethical dilemmas where the saving is certain versus probabilistic.

Study 6

Study 6 turned to a different probability: probabilistic saving with certain harm. We adapted the scenarios from Study 5 to examine a parallel range of probabilities, this time on the saving side. For example, would an action that kills one to save two be judged differently from an action that kills one to save four who have a 50% chance of dying without the intervention?

Study 6 Materials and Participants

Study 6 examined the effect of probability across the same ratios as Study 5: 1v5 (R = 5), 1v2 (R = 2), and 4v5 (R = 1.25). While the expected value of the number of lives being saved was held constant, probabilities of the missile hitting the group on the saving side were systematically varied: the missile had a 1%, 10%, 20%, 50%, or 100% chance of hitting the group of people the agent was considering saving. Participants were asked, for example, *Should Bob set off a rocket that he knows will kill 1 person, but that he also* *knows will destroy a missile that will otherwise have an X % chance of killing Y people* [where XY = 200]*?* (-5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*). 697 participants were recruited (603 passed an attention check and were retained for analysis; mean age = 34.3, SD = 11.5, 57.7% female).

Study 6 Results

As in Study 1, An ANOVA with probability and expected value ratio entered as continuous factors revealed a significant effect of ratio, $F(1, 599) = 6.14$, $p = .014$, $\eta_p^2 = .01$ (mean 1v5 = 1.42, SD = 3.00; mean 1v2 = 1.12, SD = 3.11; mean 4v5 = .63, SD = 3.03); see Figure 6a. Unlike situations involving probabilities of harming, there was a significant effect of probability of saving while keeping expected value matched, $F(1, 599) = 59.5$, $p < .001$, $\eta_p^2 = .09$. The pattern is consistent with participants being risk averse to probabilistic saving: they were more likely to endorse the action when the chances of saving the group was high (e.g. 100%) and less likely to endorse it when the chances were low (e.g., 1%), even though the expected number saved was constant; see Figure 6b. This pattern occurred for each of the ratios: $F(1, 199) = 39.1$, $p < .001$, $\eta^2 = .16$, $F(1, 196) = 12.9$, $p < .001$, $\eta^2 = .06$, $F(1, 202) = 13.5$, $p < .001$, $\eta^2 = .06$ for 1v5, 1v2, and 4v5, respectively. The interaction of probability and ratio was marginal $F(1, 599) = 3.05$, $p = .08$, $\eta_p^2 = .01$.

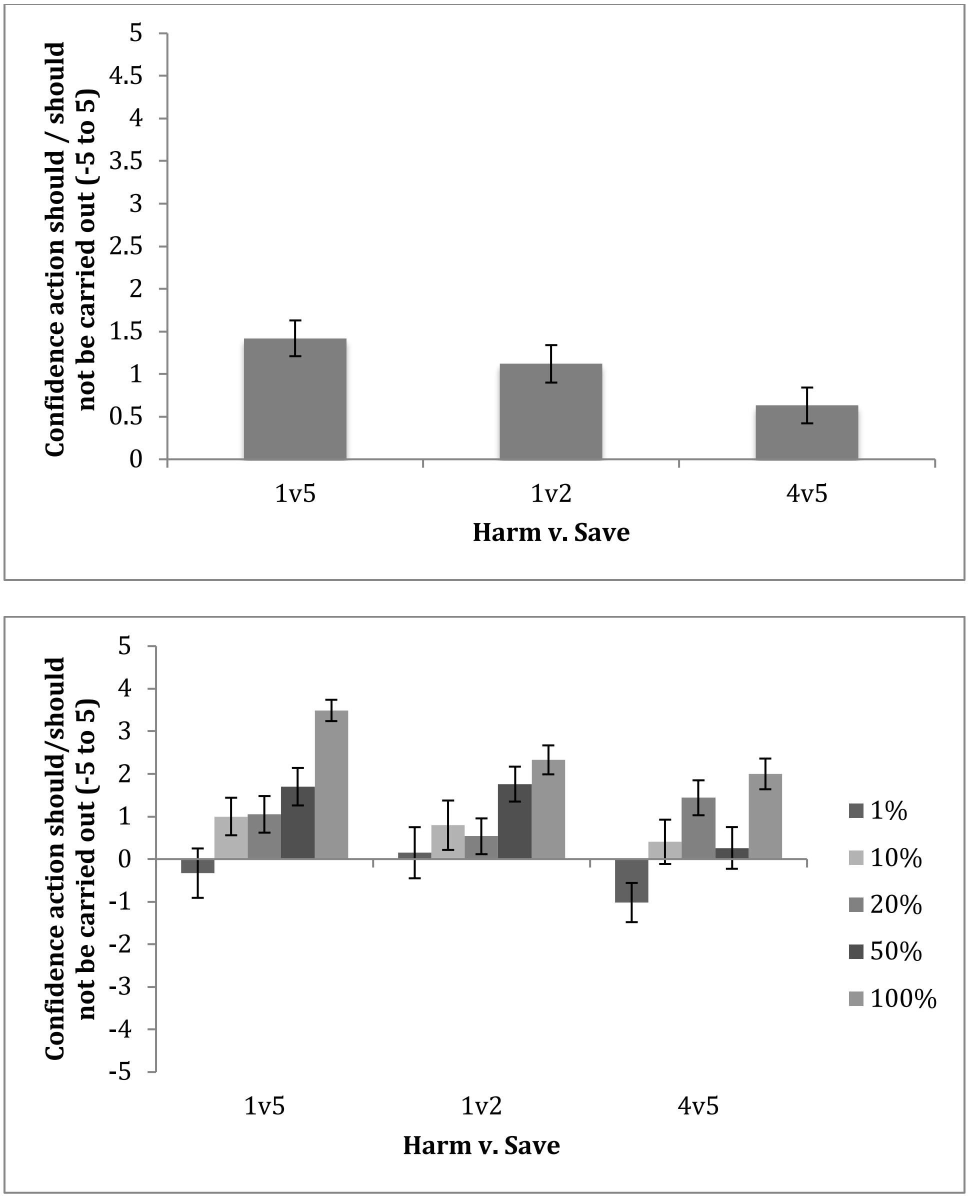


Figure 2.6. Sensitivity to expected value ratio of lives saved to lives lost in more concrete scenarios when probability of saving is varied (a). Sensitivity to various forms of the same expected value ratio (b). Error bars represent one standard error.

Study 6 Discussion

Participants continued to exhibit sensitivity to expected value ratio, being more confident of the action's rightness as the ratio of the number saved to the number sacrificed increased. However, unlike the findings for uncertain harming in single evaluation cases, we found a significant sensitivity to probability on the saving side.

Participants were averse to versions of plans that, though holding expected value fixed, probabilistically save lives. For example, when it came to sacrificing four to save an expected value of five, people were generally favorable when the saving of five was certain, and unfavorable when it was presented as a 1% chance of saving 500.

Study 7

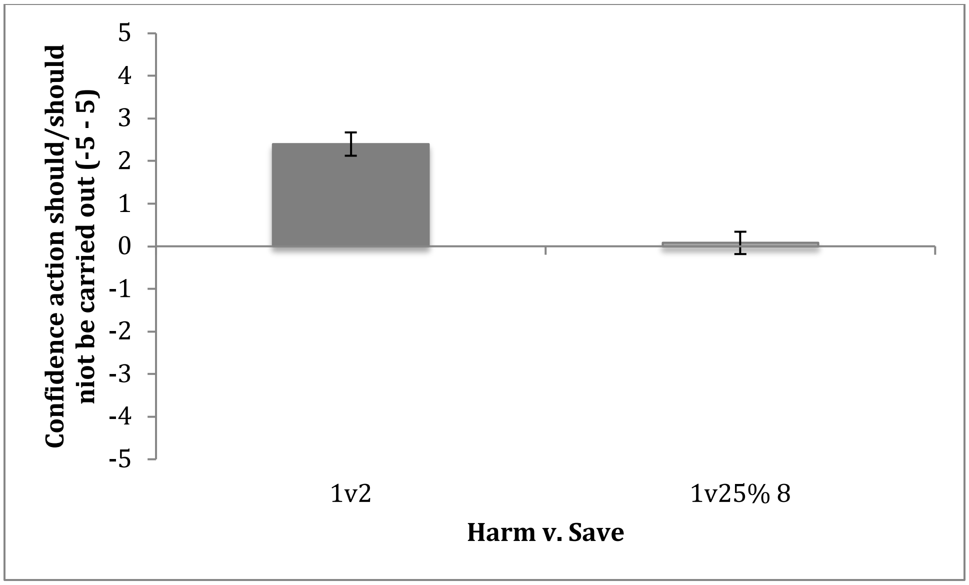
Study 6 shows that, under single evaluation, there exists a preference for certain saving over probabilistic saving. Study 7 explored what happens to the preference for certain saving under joint evaluation. That is, though people prefer an action that is certain to kill one to save two over an action that is certain to kill one to save eight from a 25% chance of death, do they become indifferent to the two plans under joint evaluation?

Study 7 Materials

In Study 7, half of the participants were exposed to a scenario where the plan would kill one to save two (R = 2), and half were exposed to a scenario where the plan would kill one for a 25% chance of saving eight (R = 2). After evaluating one of the two plans, participants saw both plans and were asked which one was morally better: *Which of the following two plans is morally better: firing the rocket at Missile X, which certainly kills 1 bystander and destroys a missile that will otherwise certainly kill 2 people, or firing the rocket at Missile Y, which certainly kills 1 bystander and destroys a missile that will otherwise have a 25% chance of killing 8 people?* (-5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*). Two hundred and thirty-five participants were recruited via Amazon's Mechanical Turk for participation in Study 7 (199 were retained for analysis after excluding participants who failed an attention check, mean age = 36.0, SD = 11.7, 65.3% female).

Study 7 Results

In the single evaluation, participants were again more confident that the agent should fire the missile when certainly saving two, than when doing so would have a 25% chance of saving 8, *t*(197) = 6.47, *p* < .001, *d* = .87 (mean certain saving 2 = 2.38, SD = 2.40; mean 25% chance of saving 8 = .08, SD = 2.89), see Figure 7a. When both plans were compared directly, the preference for certain saving remained: *t*(199) = 2.83, *p* = .005, *d* = .20 (mean =-.62, SD = 3.08), see Figure 7b. Participants thus exhibited risk-aversion to uncertain saving both under single and joint evaluation.



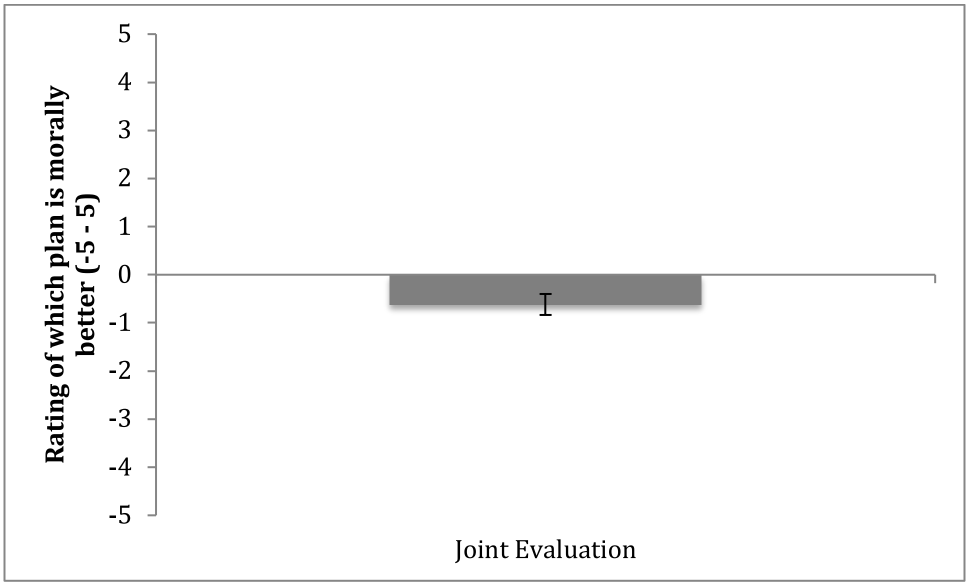


Figure 2.7. Sensitivity to whether saving is probabilistic or certain when expected value of lives saved to lives lost is fixed under single evaluation (a). Preference for probabilistic harm over certain harm when expected value is fixed under joint evaluation (negative ratings indicate a preference for the certain saving; b). Error bars represent one standard error.

Study 7 Discussion

Participants favored certain saving when evaluating cases individually and under joint evaluation. Thus, participant responses for probabilistic saving in neither case aligned with expected value, unlike for harm, where participants' responses were determined by expected value under single, but not joint evaluation. The joint evaluation case, when evaluating two plans that may save two groups of people, had a relatively small effect, given the magnitude of the single evaluation scale difference—this case is intrinsically more complex, as choosing not to pursue the plan that could or will save a group of people may be interpreted as letting the other group of people die. One way to account for this is to suppose that the pattern of responses in joint evaluation is not based only on whether saving is probabilistic or certain, but is rather a function of the probability of saving, the probability of letting die, and the number of people at risk of being let die. This raises the issue of whether and how people are sensitive to expected value and risk when it comes to letting die. In particular, will people's patterns of responses be risk-seeking, as in cases of killing, or risk-averse, as in cases of saving? This is an open question that requires further investigation.

General Discussion

Collectively, our studies show that people are sensitive to expected value in moral dilemmas, and they show this sensitivity across a range of probabilities. The particular kind of sensitivity to expected value participants display is consistent with the view that people's moral judgments are based on one single principle of threshold deontology. If one examines only participants' reactions to a single dilemma with a given ratio, one might naturally tend to sort participants' judgments into consequentialists (the ones who condone killing to save others) or non-consequentialists (the ones who do not). But this can be misleading, as is shown by the result that a number of participants who make judgments consistent with consequentialism in a scenario with ratio of 5:1 switch to an apparently deontological judgment when the ratio decreases. The fact that some judgments participants make are consistent with consequentialism does not entail that these judgments are expressive of a generally consequentialist moral theory. They consistently reflect a threshold deontological theory. On this theory, there is a general constraint against killing, but this constraint is overridden when the consequences of inaction are bad enough. The variability across people suggests that participants have different thresholds of the ratio at which the consequences count as “bad enough” for switching from supporting inaction to supporting action. This is consistent with the wide literature showing that people’s judgments can shift within the same ratio, depending on, for example, how the death of the one is caused.

Making the harms of action uncertain has a limited effect on participants’ moral choices. In a single evaluation case, people’s confidence about the moral rightness or wrongness of sacrificing one to save five was no different from their confidence about the moral rightness or wrongness of subjecting one hundred people to a 1% risk of sacrifice to save five. This suggests that people do not invoke distinct moral principles when it comes to harming others versus putting them at even slight risk of harm. It also indicates that, in this sort of dilemma at least, people do not show the usual risk-seeking tendency when it comes to losses. However, this tendency is apparent when people are asked not whether to act, but instead which action plan to implement. In such a joint evaluation case, participants show a preference for spreading the risk of harm across many people. In fact, this preference is strong enough that they are willing to incur twice as much expected harm when that harm is probabilistic.

Sensitivity to probability on the saving side revealed a somewhat different effect. In single evaluation cases, people are less confident about the rightness of bringing about a particular level of harm when the benefit is uncertain, even when it has the same expected value. This is consistent with prospect theory and risk aversion for gains. This preference for concentrating benefits is sufficiently strong that people are willing to take about half as much expected good for a given sacrifice if that good is certain rather than chancy.

One partial explanation for the asymmetry between responses on the saving side and responses on the harm side might be that subjects are bringing consistent deontological principles to bear. On some deontological theories, one's duty not to kill is stronger than one's duty to save or otherwise benefit others. Further, on such theories, one's duty to save or benefit others might be such that one can choose among a wide range of ways to fulfill the duty, and, in some situations, there may be no duty to save or benefit others at all. (See the distinction between perfect and imperfect duties in Kant 1785/2002, p. 222.) The scenarios are complicated in that they involve both imposing and reducing risk, but it is possible that in the case of benefiting others, given that there is no duty to benefit (or to benefit in any particular way) in the first place, one has no duty to distribute increased chances of living to more people as opposed to increasing even more the chances of living for a smaller group. Thus, with no such duty involved, but with a high value on certain saving, it makes sense in this case to prefer to save a smaller number with certainty than perform an act that at best will decrease others' chances of dying when they might not have died in any case. And this is what we find in subjects' responses to single evaluations in Studies 6 and 7. In contrast, when we vary whether the agent would cause certain death, or merely risk death in a single evaluation (as we do in Study 3), we do not find a difference in subjects' responses. In that case, it is only in the joint evaluation that we find differential responses. One reason for this might be that it is *prima facie* wrong to impose significant risk, just as it is *prima facie* wrong to kill, and so the salience of the comparison does not arise until one sees the two options in a joint evaluation. Thus, while there may be no duty to save (or to save in a particular way or on particular occasions), and so it is permissible to take into account factors such as probability of having an effect, there is a duty at all times not to kill or impose significant risk. And this could go some way toward explaining why it is only once subjects see a direct comparison of certain harming and spreading the risk of harm to a larger group that *this* difference becomes especially salient.

Our findings also contribute to a more nuanced understanding of deontology by comparing how participants respond to probabilistic and certain death. While deontologists may not be willing to kill one to save five, they may deem it acceptable to risk a 1% chance of harm to one to save five. Our data suggest that expected value calculation, rather than level of risk itself, accounts for this shift in judgment. An open question remains as to what determines a person's deontological weighting, or the value by which their expected value calculation is offset in deciding whether to act. Another open question concerns whether and to what extent people's deontological weighting about probabilistic harming and saving depends on the status quo. For example, our studies treated probabilistic saving, in the 25% case, as decreasing the probability of someone dying from 25% to 0%, but one might also probabilistically save someone by decreasing the probability of their dying from 100% to 75%. It requires further investigation whether people will respond differently according to the pre-existing level of risk in the circumstances.

Similar patterns of sensitivity to expected value and probability emerge in our findings with both concrete and abstract scenarios. The patterns also suggest that people are inclined to make more extreme moral judgments (e.g., being more confident that it is morally acceptable to kill one in order to save two) in our concrete scenarios than they are in our abstract scenarios. This difference is consistent with some recent work on Construal Level Theory (CLT) and moral judgment (Gong & Medin, 2012; Lammers, 2012). Based on CLT, one possible explanation of the difference in our findings is that people engage in low-level construals in concrete scenarios, and such low-level construals can intensify moral judgments by being easier to imagine (see Gong & Medin, 2012, p. 635). By contrast, people engage in high-level construals in abstract scenarios, and these high-level construals involve greater psychological distance that can mitigate the radicalness of moral judgments. But it is not obvious how to apply the theory in this case, since subjects are being asked to imagine both the possibility of two people dying and one person being killed intentionally by another. Since both aspects are made more vivid in the concrete scenario, it is not clear in which direction the moral judgment in this case we should expect to be intensified. We believe that the question of how responses differ with respect to concrete and abstract scenarios is an interesting one worth further exploration.

Our findings do not appear to fit clearly with the broader literature on evaluability, which finds that people make more logical decisions under joint evaluation than they do under single evaluation (e.g., Bazerman & Gino, 2012). For example, Bartels (2008) found that omission bias, readily apparent under single evaluation, can be reduced under joint evaluation. More broadly, it has been argued that normative positions should be derived from joint evaluation. Our Studies 4 and 5 find that people systematically deviate from expected value under joint evaluation, but not under single evaluation, when all other aspects of the scenario remain fixed, in deciding whether to spread risk of harm to many. It is not clear which position is normatively correct. One might view the simple calculation of the expected harm as correct, in which case it is the single evaluation cases that track the proper principle. On the other hand, there are some plausible normative arguments for spreading risk of harm when expected values are equal. That is, all else being equal, it might be better to commit the four lesser wrongs of imposing a 25% risk of harm than to commit the one greater harm of imposing certain death on one. Spreading risk might be preferable because it makes people more *evenly* suffer right-violation and is thereby a more just option when all else is equal. Alternatively, it might be the case that the aggregate of four deaths isn't exactly four times the moral worth of one death, but smaller than the latter due to diminished marginal (negative) utility. Finally, it might be the case that the norms of instrumental rationality already involve sensitivity to risk. Buchak (2013) has recently defended a "risk-weighted expected utility theory" as an alternative normative theory to the standard expected utility theory of instrumental rationality. If her theory is correct, then it is possible for us to both be risk-seeking and instrumentally rational. And since considerations about instrumental rationality can affect moral judgments—a point that even threshold deontologists would agree with—it is hardly surprising that moral judgments are sensitive to risk. It will then follow that, in the comparison between certain and probabilistic harming, the joint evaluation rather than the single evaluation is consistent with the proper principle.

This presents a complication for those using moral dilemmas to inform policy surrounding autonomous vehicles, as their findings outline moral preferences for how autonomous vehicles should be programmed under single evaluation (e.g., Awad et al., 2018). If such single evaluations are not coherent with joint evaluations of the multiple options that autonomous vehicles are likely to have in the face of a collision, policy may be better informed by the study of moral dilemmas involving joint evaluations.

Our data suggest that people seem, on the whole, not to embrace simple absolute consequentialist or non-consequentialist moral positions, but hold instead more nuanced views, balancing the harm done, the good achieved, and the value of rights, consistent with a principled threshold deontology. Our data also begin to shed light on the remarkably neglected domain of moral principles applied in an uncertain world. The normative ethical positions are largely silent on how such applications should be made, and so, given that almost every actual dilemma is likely to feature some degree of uncertainty at some level, data on how people view such dilemmas is especially valuable and potentially relevant to social policies and procedures. In order to give herself a high probability of saving a small group of people (or a low probability of saving a large group of people), a firefighter might need to break a window that will cause a fire to reach an elderly person who is unable to move. What should she do? Should an autonomous vehicle be programmed to avoid plowing into a school bus by moving to the left, where there is a low probability of colliding with a tandem, or by moving to the right, where there is a high probability of colliding with a pedestrian? Among the factors that appear to be important, and worthy of serious further scrutiny, are whether the uncertainty is on the harm side or the benefit side, and whether the dilemma is about whether to incur that harm or instead how to apportion it.

Chapter 2 has been submitted for publication of the material as it may appear in Cognitive Science, 2019. Ryazanov, Arseny; Wang, Tinghao; Nelkin, Dana; Christenfeld, Nicholas; Rickless, Samuel. The dissertation author was the primary investigator and author of this paper.

References

Agerström, J., & Björklund, F. (2009). Moral concerns are greater for temporally distant events and are moderated by value strength. *Social Cognition*, *27*(2), 261-282.

Alexander, L., & Moore, M. (2016). Deontological ethics. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science*, *23*(8), 861-868.

Arvan, M. (2013). Bad news for conservatives? Moral judgments and the Dark Triad personality traits: A correlational study. *Neuroethics, 6*(2), 307-318.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Barak-Corren, N., Tsay, C. J., Cushman, F., & Bazerman, M. H. (2018). If you're going to do wrong, at least do it right: Considering two moral dilemmas at the same time promotes moral consistency. *Management Science*, *64*(4), 1528-1540.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381-417.

Bazerman, M. H., & Gino, F. (2012). Behavioral ethics: Toward a deeper understanding of moral judgment and dishonesty. *Annual Review of Law and Social Science*, *8*, 85-104.

Brennan, S. (1995). Thresholds for rights. *Southern Journal of Philosophy*, *33*, 143-168.

Buchak, L. (2013). *Risk and rationality*. Oxford University Press.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, *104*(2), 216.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J.D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241-265

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363-366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Cushman, F., & Greene, J. D. (2011). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience, 7*(3), 269–279.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. In *The Oxford Handbook of Moral Psychology* (pp. 47–71). Oxford University Press.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgments: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian Disease problem: Time constraint, need, and disease type. *Judgment and Decision Making*, *13*(6), 529-546.

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

Fleischhut, N., Meder, B., & Gigerenzer, G. (2017). Moral Hindsight. *Experimental Psychology*, *64*(2), 110-123.

Fried, C. (1978). *Right and wrong*. Harvard University Press.

Gong, H., & Medin, D. L. (2012). Construal levels and moral judgment: Some complications. *Judgment and Decision Making*, *7*(5), 628–638.

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*(3), 364-371.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in Cognitive Sciences*, *6*(12), 517-523.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144-1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400.

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*(5537), 2105-2108.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & language*, *22*(1), 1-21.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247-257.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: a review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science*, *5*(4), 343-355.

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *Journal of Neuroscience*, *35*(36), 12593-12605.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193-209.

Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty*, *9*(1), 5-37.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making: Part I* (pp. 99-127). World Scientific Publishing Co.

Kant, I. (1785/2002). *Groundwork for the metaphysics of morals*. Translated by A. Zweig and edited by A. Zweig and T. E. Hill Jr. Oxford University Press.

Lammers, J. (2012). Abstraction increases hypocrisy. *Journal of Experimental Social Psychology*, *48*(2), 475–480.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press.

Moore, M. (1997). *Placing Blame: A Theory of the Criminal Law*. Oxford University Press.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological science*, *19*(6), 549-557.

Nagel, T. (1979). War and massacre. In *Mortal Questions* (pp. 53-74). Cambridge University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530-542.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163-177.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of doing and allowing. *Philosophical Review 98*(3), 287-312.

Rickless, S. C. (1997). The doctrine of doing and allowing. *Philosophical Review 106*(4), 555-575.

Ryazanov, A. A., Knutzen, J. , Rickless, S. C., Christenfeld, N. J., & Nelkin, D. K. (2018). Intuitive probabilities and the limitation of moral imagination. *Cognitive Science, 42*, 38-68.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of cognitive neuroscience*, *18*(5), 803-817.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, *34*(13), 4741-4749.

Shou, Y., & Song, F. (2017). Decisions in moral dilemmas: The influence of subjective beliefs in outcome probabilities. *Judgment and Decision Making*, *12*(5), 481–490.

Sinnott-Armstrong, W. (2015). Consequentialism. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/win2015/entries/consequentialism/

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, *4*, 1-8.

Thomson, J. J. (1990). *The realm of rights*. Harvard University Press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323.

Chapter 3: Sensitivity to Shifts in Probability of Harm and Benefit in Moral Dilemmas

Arseny A. Ryazanov, Tinghao Wang, Samuel Rickless, Craig McKenzie, Dana Nelkin

Correspondence concerning this article should be addressed to Arseny Ryazanov, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109

Contact: aryazano@ucsd.edu

Abstract

Psychologists and philosophers working on the nature of moral judgments rely on hypothetical moral dilemmas typically specify outcomes as certain to occur. This is in sharp contrast with real-life moral dilemmas or moral decision-making, which is almost always infused with outcome uncertainty. Six studies contribute to the growing literature on moral judgment under probabilistic outcomes by examining sensitivity in moral dilemmas to where shifts in probability of harm and saving occur. By pitting participant sensitivity to size of shift in harm probability against an observed sensitivity to location of probability shift, we identify that end-states of harm imposed by an action matter more than the magnitude the harm imposed, under both single evaluation and when the comparison between shifts is made explicit under joint evaluation. Our results also identify that moral judgments made under joint evaluation deviate further from expected value than judgments made under single evaluation. Consistent with the Dyadic Model of Morality, perceived harm partially mediates sensitivity to location of harm probability shift. Unlike for shifts in likelihood of harm, participants are found to be insensitive to where the shift in saving probability occur under both single and joint evaluation, suggesting an asymmetry between harm and benefit in moral reasoning.

Sensitivity to Shifts in Probability of Harm and Benefit in Moral Dilemmas

Despite being long-neglected by philosophers and psychologists alike, the role of probability in moral decision-making is garnering increasing attention (Shenhav & Greene, 2010; Ryazanov, Knutzen, Rickless, Christenfeld, & Nelkin, 2018; Shou & Song, 2017; Fleischhut, Meder, & Gigerenzer, 2017). Such research bridges the gap between the hypothetical scenarios that stipulate certain outcomes, through which moral dilemmas are traditionally studied, and the real-life scenarios that are infused with outcome uncertainty. For instance, instead of asking "should Tom certainly kill one person to certainly save five people?", the studies mentioned above ask questions such as "should Tom certainly kill one person for a 50% chance of saving ten people?" or "should Tom risk a 50% chance of killing two people for certainly saving five people?". Prior studies on uncertain moral dilemmas have identified systematic sensitivity to outcome likelihoods in moral judgments, both when it comes to probabilistic saving and probabilistic harming.

The Size and Location of Probability Shifts in Moral Dilemmas

We distinguish between the *size* of a probability shift and the *location* of a probability shift. The *size* of a probability shift concerns how much the probability of a certain outcome increases or decreases when one performs an action. But a probabilistic shift with the same size—e.g., a 25% increase—can occur in different *locations*, despite resulting in an identical change in expected value. For example, consider a plan that increases the probability of killing four people by 25% in order to save two others. The 25% increase in the probability of four people dying results in an expected loss of one life (.25 x 4), whereas certainly saving two people is an expected gain of two lives, leading to a favorable ratio of expected lives saved to expected lives lost (expected value ratio of 2). Note that the 25% increase in probability could occur anywhere in the 0-1 probability interval and not affect expected value. For example, increasing the probability of four people dying from 0% to 25% has the same expected loss of life as does increasing the probability from 75% to 100%.

Thus far, studies on moral dilemmas with uncertainty have all focused on the *size* of probability shifts, and, more specifically, on how the size of probability shifts contributes to the differences in expected value calculation where such shifts are presumed to be increases from 0% or decreases from 100% likelihood of an outcome. However, the *location* of probabilistic shifts could also matter for moral judgments. In the examples above, it might matter that increasing the probability of a group of four people dying from 0% to 25% means putting at risk a group that is otherwise facing no risk. Or it might matter that the increase from 75% to 100% means that everyone in the group will certainly die. If it turns out that the location of probability shifts has a robust effect on moral judgment *in addition to* the effect from the size of probability shifts, then it becomes puzzling how to characterize the nature of folk moral psychology. The finding would contradict the view that people's moral judgments are *consequentialist*, since being a consequentialist in the traditional sense is to recognize as morally relevant only the difference in expected value of an action, which is independent of the location of the probability shift. Though sensitivity to the location of probability shifts could be consistent with claiming that people's judgments are *deontological*, traditional deontological theories don't have the ready resources to explain the variations either. It is unclear why deontological constraints -- e.g., that we should not violate people's rights or use people as mere means -- are sensitive to *where* the probabilistic shifts occur. Moral dilemmas with different locations of probability shifts are thus worth systematic empirical investigation.

Relevant Research and Predictions

There is already some evidence suggesting the relevance of the location of probability shifts in decision making outside of the moral domain. In a study by Gonzalez and Wu (1999), participants were asked to select which of the following felt like a more significant change: increasing the odds of a lottery ticket that has a 65% chance of winning to 70%, or increasing the odds from 90% to 95%. Most participants chose the latter option. When offered a similar choice between increasing the odds of a ticket that has a 5% chance of winning to 10%, or from 30% to 35%, participants were more likely to choose the former option. This suggests that participants' responses in monetary decisions are sensitive to the location of probability shifts and that participants are more sensitive to probability shifts that occur closer to 0% and 100% than to probability shifts that occur closer to 50% for monetary decisions. It remains an open question whether participants would be similarly sensitive to probability shifts in moral dilemmas, and we thereby take up this question in the current project. We predict that there will be some kind of sensitivity to the location of probability shifts in moral dilemmas, given the sensitivity to location in other domains.

In the lottery ticket case, we might reasonably expect participants to be sensitive to the location of probability shifts regardless of whether the shift is an increase in likelihood of a beneficial outcome such as winning the lottery, or a negative outcome, such as the likelihood of losing an investment. But in moral dilemmas, which contain both negative outcomes (harm caused) and positive outcomes (benefit), there could be an asymmetry in sensitivity to the location of probability shifts for shifts likelihood of harm and shifts in likelihood of benefit. Intuitively, in the absence of other morally relevant factors, there is a *right* or *claim* to not have one's probability of dying increased whereas there is no right or claim to have one's probability of survival increased. More generally, there is a right or claim to not be *probabilistically harmed* whereas there is no right or claim to be *probabilistically benefited*. This parallels what moral philosophers refer to as the difference between the right and the good (Ross, 1930) or the distinction between justice and charity (Foot, 1978).

Empirical evidence adds further support to the philosophical notion of a harm/benefit asymmetry, which could possibly extend to an asymmetry in sensitivity to harm and benefit likelihood location shifts. Guglielmo and Malle (2019) find that blame is more differentiated than praise, more specifically, that mental states preceding negative actions are more finely-distinguished than mental states preceding positive actions (also see Monroe & Malle, 2019). Likewise, negative events have a larger range of linguistic representation than positive events (Rozin & Royzman, 2001; Peeters, 1971). In addition, some empirical theories on moral judgment give a much more central role to harm than they do to benefit. One notable example is Schein and Gray's (2018) theory of dyadic morality, which predicts the immorality of an act almost exclusively in terms of the perceived harm it causes. If such discernment extends to sensitivity to various ways of expressing the same change in expected value, but with the location of where the probability shift occurs varied, it suggests that participants may discern more among different locations in probabilistic harm (e.g., whether risk of harm is raised from 0% to

25%, or from 75% to 100%) than among different locations in probabilistic benefit (such as whether the risk of death is reduced from 100% to 75% or from 25% to 0%).

We examine whether moral judgments are more sensitive to the location of probabilistic shifts in the case of harm than in the case of benefit, as would be possible if the blame/praise differentiation asymmetry extended to location sensitivity for where shifts in probabilistic harm and benefit occur and would be consistent with the asymmetrical treatment of harm and benefit in some major moral theories. In light of Schein and Gray's (2018) theory of dyadic morality, we hypothesized that sensitivity to different locations of probabilistic shifts in harm could be attributable to differences in *perceived harm*. This is also consistent with Gonzalez and Wu's (1999) study, since if a shift of harm likelihood feels like a more *significant* change, it could be that the action feels *more harmful*. Though there are presumably deeper explanations for the differences in harm perception, we take it that perceived harm is a good starting point and already goes some way towards giving a more complete story of folk moral psychology in uncertain moral dilemmas.

Finally, moral judgments can be studied between-subjects, by posing different scenarios to different people, or within-subjects, by asking participants whether they find a particular difference to be morally relevant. Both approaches are useful: more consistent moral judgments are made under joint evaluation than when rating individual scenarios for two versions of trolley problems—where a person can be dropped onto a trolley track to save others ahead or where a trolley can be diverted onto a track with one person on it to save others ahead (Barak-Corren, Tsay, Cushman, & Bazerman, 2018). One reasonable explanation for this difference is that participants are motivated to reflect on whether their divergent reactions to the two scenarios are normatively defensible under joint evaluation. Thus, in this paper we explore the effect of the location of probability shifts in moral judgments under both single and joint evaluation. We hypothesize that the single evaluations will reveal people's unreflective preferences, while the joint evaluations will reveal people's considered preferences once the differences in individual cases are made particularly salient.

Studies

Six studies examine the role of shifts in probability in moral judgment. Study 1 finds that participants are sensitive to the location of probability shifts for harm, but not benefit, when the size of probabilistic shifts is held fixed. Study 2 examines whether the effect of location shifts in harm probability reflects sensitivity to end-state probability and insensitivity to the size of the shift. Studies 3a and 3b explore whether participants endorse the patterns observed in the between-subject studies upon reflection, under joint evaluation. Through mediation analyses, Study 4a examines the relationship between location of shift in likelihood of harming bystanders and the perceived harmfulness of the action, while Study 4b examines the relationship between location shift in likelihood of saving a group and perceived benefit.

Study 1

We begin by exploring whether participants are sensitive to the location of probability shifts for both harmful and beneficial outcomes in moral dilemmas, when the size of probabilistic shifts is held fixed.

Study 1 Methods

One thousand nineteen participants were recruited via Amazon's Mechanical Turk (862 passed an attention check; because results did not significantly differ between the full sample and those passing the attention check, all participants were retained for analysis; 61.8% female; mean age = 34.4, SD = 10.7). Participants were randomly assigned to read one of eight scenarios. Four of the scenarios concerned moral dilemmas in which two people could certainly be saved by increasing the risk of harming 4 bystanders by 25%. The four scenarios all had different starting and ending points for the probability shift, but the same expected value: 0% to 25%, 25% to 50%, 50% to 75%, and 75% to 100%, respectively, see Table 1. As an example, one of the scenarios was as follows:

Harry sees a group of two people who will certainly die without intervention. He knows the following facts. There is a group of four bystanders that is facing a 0% risk of death. Harry can carry out a plan that will certainly save the group of two people. However, in carrying out the plan, Harry will increase the risk of death for the group of four bystanders from 0% to 25%.

Participants responded to a single question regarding their confidence that the action should be carried out, adapted to each scenario, along an eleven-point scale. For example, participants were asked, *Should Harry carry out a plan that he knows with certainty will both save the group of two people and at the same time raise the risk of death for the group of four bystanders from 0% to 25%?* (-5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*).

The other four scenarios concerned moral dilemmas in which the probability of a group of eight people dying can be decreased by 25% as a result of certainly killing one bystander. Again, four scenarios all had different starting and ending points for the probability shift, but the same expected value: 25% to 0%, 50% to 25%, 75% to 50%, and 100% to 75%, respectively. For example, one of the scenarios was as follows:

Harry sees a group of eight people whose lives are in danger. He knows the following facts. There is a 25% chance of the group of eight people dying. Harry can carry out a plan that will reduce the risk of the group of eight dying from 25% to 0%. However, in carrying out the plan, Harry will certainly kill one bystander.

Participants responded to a single question regarding their confidence that the action should be carried out, adapted to each scenario, along an eleven-point scale. For example, participants were asked, *Should Harry carry out a plan that he knows with certainty will both reduce the risk of the group of eight dying from 25% to 0% and at the same time kill one bystander?* (-5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*).

Table 3.1. Study 1 scenarios.

Probability shift	Harm Scenarios	Save Scenarios	EV ratio of action
0% to 25%	Increase chance of 4 people dying from 0% to 25% to save 2 people	Kill 1 person to decrease chance of 8 people dying from 25% to 0%	2
25% to 50%	Increase chance of 4 people dying from 25% to 50% to save 2 people	Kill 1 person to decrease chance of 8 people dying from 50% to 25%	2
50% to 75%	Increase chance of 4 people dying from 50% to 75% to save 2 people	Kill 1 person to decrease chance of 8 people dying from 75% to 50%	2
75% to 100%	Increase chance of 4 people dying from 75% to 100% to save 2 people	Kill 1 person to decrease chance of 8 people dying from 100% to 75%	2

Study 1 Results

We first verified that participants were generally endorsing an action that had a positive expected value ratio (good done to harm done) of 2. Across all eight versions of the action, participants endorsed the action *t*(1018) = 7.01, p < .001, *d* = .22 (mean = .67, SD = 3.04). There was no significant difference between ratings assigned to scenarios where harm was probabilistic and scenarios where saving was probabilistic, *t*(1017) = 1.11, *p* = .27, *d* = .07 (mean harm = .77, SD = 2.97; mean save = .56, SD = 3.11).

Entering where the shift in harming/saving occurred as a linear factor (25-0, 50-25, 75-50, 100-75, recoded as 1, 2, 3, 4, respectively), and entering the type of probability shift (harm or save), as well as their interaction as factors into an ANOVA revealed an interaction between type of shift and sensitivity to probability shift *F*(1, 1015) = 27.1, *p* < .001, *r* = .18 , suggesting that participants were differentially sensitive to location of probability shifts for probabilistic harms and probabilistic saving (mean harm 0-25% = 1.55, SD = 2.54; mean harm 25-50% = 1.19, SD = 2.80; mean harm 50-75% = .89, SD = 2.81; mean harm 75-100% = -.55, SD = 3.29; mean save 25-0% = .35, SD = 3.20; mean save 50-25% = .33, SD = 2.91; mean save 75-50% = .76, SD = 3.13; mean save 100-75% = .81, SD = 3.20; see Figure 1a). Reverse coding shift location for save shifts (so that 1 = 100-75, 2 = 75-50, etc.) also yielded a significant interaction, *F*(1, 1015) = 8.16, *p* = .004, *r* = .09, see Figure 1b), suggesting that the interaction was not an artifact of the coding scheme used to compare harm shifts and saving shifts.

We then separately examined sensitivity to probability shifts on the harm side—is a plan that raises the risk of four bystanders dying from, e.g., 0% to 25% in order to save two people preferred to a plan that raises the risk of four bystanders dying from 50% to 75%? Entering where the shift in harm occurs as a linear factor revealed that participant judgments of whether the action should be carried out were influenced by where the 25% shift in likelihood of harm occurs, *F*(1, 510) = 33.4, *p* < .001, *r* = .25. In order to examine whether this linear effect was driven solely by an aversion to causing certain death—the 75%-100% shift, the linear model was rerun after excluding the 75%-100% shift and still yielded a significant linear relationship between being decreased confidence in action and the location of the probability shift 1, *F*(1, 510) = 33.4, *p* < .001, *r* = .25. Thus, as the 25% increase in likelihood of harm occurred closer to 100%, participant confidence in carrying out the action decreased.

We next examined sensitivity to probability shifts on the saving side—for example, is a plan that will certainly kill one bystander in order to reduce the risk of a group of 8 dying from 25% to 0% viewed more favorably than a plan that reduces the risk of a different group of 8 dying from 75% to 50%? Entering where the shift in saving occurs as a linear factor (25-0, 50-25, 75-50, 100-75, recoded as 1, 2, 3, 4, respectively), revealed that participants judgments were not influenced by where the 25% shift in likelihood of saving occurred, *F*(1, 506) = 2.14, p = .14, *r* = .065. To verify that the observed lack of effect on the saving side was not the result of framing the saving likelihood as a decrease in the likelihood of dying, rather than an increase in the likelihood of survival, a separate study compared the two versions for several probability shifts and found no evidence of sensitivity to probability shift for either frame (see supplementary materials).

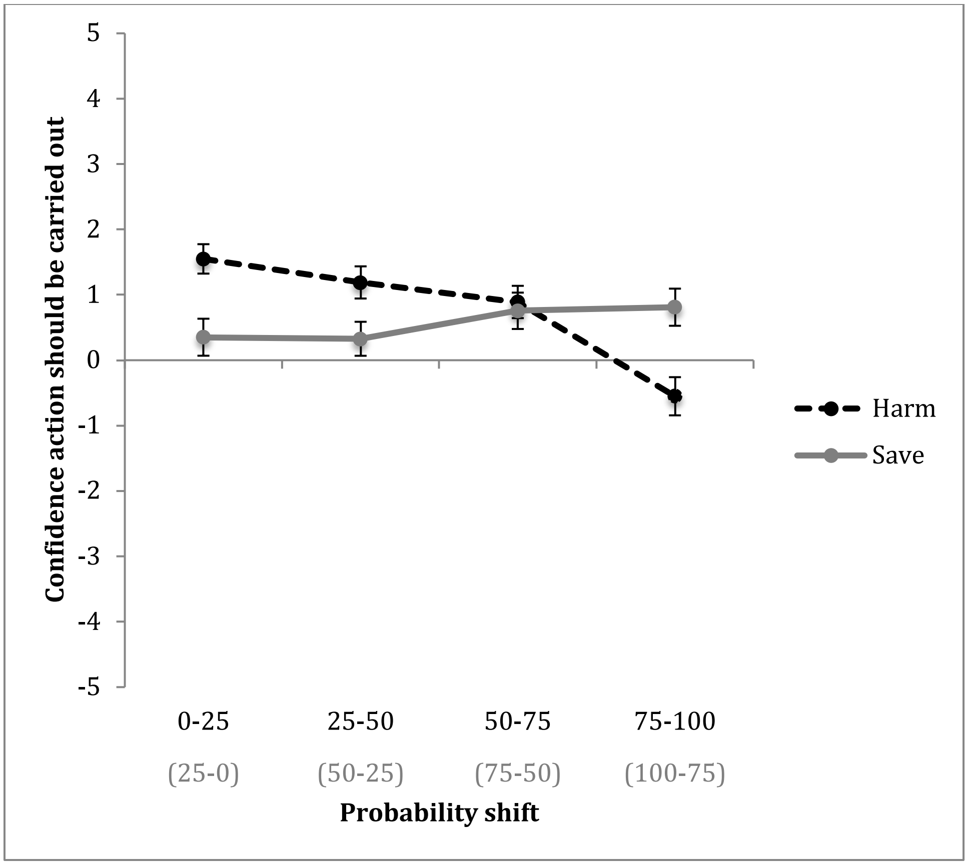


Figure 3.1. Sensitivity to probability shift location for an action that increases in likelihood of four bystanders dying by 25% in order to save two people (harm), and to probability shift location for an action that deceases the likelihood of eight individuals dying by 25% by killing one bystander. Error bars represent one standard error.

Study 1 Discussion

Study 1 found that that the location of probability shifts affects moral reasoning independently from the ways in which the numerical value of the shift contributes to the differences in expected value calculation for probabilistic harm. Harry's plan in each of the eight scenarios concerns exactly the same ratio in expected value—the equivalent of 2 lives saved and 1 life lost. Though participants, on the whole endorse action, and we did not observe overall differences in endorsing actions between actions that involve probabilistic harm or benefit, participants were sensitive to location of the shift in harm likelihood, but not the location of the shift in benefit likelihood.

Thus, there appears to be a harm/benefit asymmetry in sensitivity to the location of probability shifts, when the size of the shift is held fixed. On the harming side, participants are keenly sensitive to where the shift in probability of harm occurs. On the saving side, however, participants are largely insensitive to where the shift in probability of saving occurs. The finding of differential sensitivity between harm and benefit location shifts is consistent with the more general phenomenon that judgments of negative actions being more fine-grained than moral judgments of beneficial actions (Guglielmo & Malle, 2019; Monroe & Malle, 2019).

The effect of shifts in probability of harm appears to be linear, rather than showing increased sensitivity when the shifts occur close to 0 and 1, as found by Gonzalez and Wu (1991) for monetary decisions. Furthermore, the sensitivity to shifts of harm does not appear to simply reflect a specific to an aversion to certain harm, as excluding the 75%-100% shift from the linear model still yielded a significant linear relationship between being decreased confidence in action and the location of the probability shift.

It's not obvious what the sensitivity to the location of probability shifts shows about the nature of folk moral psychology. A common view of moral judgment depicts it as either *consequentialist* or *deontological*. Consequentialist judgments reflect an exclusive concern for the expected values in outcomes, whereas deontological judgments reflect a further concern regarding certain deontological constraints, e.g., constraints against right violation or using people merely as means. But neither consequentialism nor deontology can easily characterize the sensitivity observed in the current study. On one hand, the relevant moral judgments cannot be said to be consequentialist, since to be a consequentialist in the traditional sense is to recognize as morally salient only differences in expected values irrespective of where the probability shifts occur. Thus, the studies clearly reveal that something other than expected value is operating for probabilistic harms. On the other hand, it is not obvious what tools deontologists have to accommodate the relevance of probability shifts, since it's not totally clear how the legitimacy of any deontological constraints, e.g., those regarding right violation or treating people as ends, depends on where a probability shift occurs. Thus, the sensitivity observed in the current study constitutes a puzzling phenomenon that cannot be easily accommodated in the traditional consequentialism/deontology framework, though we make some suggestions about how to address this puzzle in the general discussion. This pattern of results may, however, be accounted for psychologically: harm that occurs closer to certainty feeling more harm-like could drive moral judgment, consistent with the role of perceived harm identified by the Dyadic Theory of Morality. We test this possibility in Study 4, after first examining whether participants attend to both start-state and end-state probabilities, and examining their judgments under single and joint evaluation.

Study 2

In Study 1, we held fixed the size of probability shifts and explored the role of the location of probability shifts. A further question concerns what will happen if the size of probability shifts is *not* held fixed. Will the end-state of probability shifts affect moral psychology in a way that *overrides* the differences in size and differing starting point? For example, we found that increasing risk of harm from 75% to 100% mattered more than increasing risk from 50% to 75%. To the extent that end state is all that matters, then an increase from 50% to 100% will have the same effect as an increase from 75% to 100%, even though the expected loss of life in the former case is twice that of the latter. Studies 2 and 3 aim to explore this general question from different perspectives.

Study 2 independently varied the size of the probability shift and the end point of the probability shift for probabilistic harms. It could be the case that the end-state probability was driving the effect in Study 1. In the event that end-states influence moral judgment, we also wanted to verify that our participants paid adequate attention to both the starting points and the ending points (i.e., the size) of the probability shifts when responding to the moral dilemmas we developed, in order to rule out the possibility that our findings in Study 1 were due to participants not paying attention to initial-state probabilities.

Study 2 Methods

Three hundred twenty nine participants were recruited via Amazon's Mechanical Turk (59.6% female; mean age = 35.2, SD = 10.7). Participants were randomly assigned to read one of four scenarios in which two people could certainly be saved by increasing the risk of harming four bystanders. In two of the scenarios, the shifts were from 0% to 50% and from 25% to 50%; in the other two scenarios, the shifts were from 50% to 100% and from 75% to 100%. Participants then responded to a single question regarding their confidence that the action should be carried out along an eleven-point scale, as in Study 1. Immediately after the question, participants were asked to recall the initial probability of harm to the four bystanders and the probability of harm to the four bystanders should the plan be carried out (*What was the **initial** risk of death for the group of four* *bystanders?; What would the risk of death be for the group of four bystanders if Harry carried out his plan?*). This was used to verify that participants paid adequate attention to the initial, not just final, probability of harm.

Table 3.2. Study 2 scenarios

Probability shift	Harm Scenarios	EV ratio of action
0% to 50%	Increase chance of 4 people dying from 0% to 25% to save 2 people	1
25% to 50%	Increase chance of 4 people dying from 25% to 50% to save 2 people	2
50% to 100%	Increase chance of 4 people dying from 50% to 75% to save 2 people	1
75% to 100%	Increase chance of 4 people dying from 75% to 100% to save 2 people	2

If participants were pure consequentialists, they should be more likely to endorse the actions that involve 25% shifts in harm (EV ratio = 2; 2 lives saved / 1 ended), compared to actions that involve 50% shifts in harm (EV ratio = 1; 2 lives saved / 2 ended). We would expect to see the main effect of shift size, but no effect of end point, since only shift size matters to EV. However, if participants are instead sensitive to endpoint rather than size of shift, we would expect to see similar ratings of actions with the same end point (e.g., 100%), regardless of shift size (50%-100% vs. 75%-100%).

Study 2 Results

A 2x2 ANOVA analyzing the full set of participants, with size of shift and end-point entered as factors revealed no significant effect of size of shift, $F(1, 324) = 1.95$, $p = .16$, $r = .07$ (mean 50% shift = -.18, SD = 3.15; mean 25% shift = .26, SD = 3.06, suggesting that participants were not sensitive to expected value, and therefore not making purely consequentialist judgments. A significant main effect of end-state revealed that participants were sensitive to whether the final end-state was a 50% chance of the four bystanders dying, or a 100% chance of the four dying, *F*(1, 324) = 59.9, *p* < .001, *r* = .39 (mean 50% end-state = 1.27, SD = 2.73; mean 100% end-state = -1.18, SD = 2.99), inconsistent with pure consequentialism. There was no significant interaction of end point and size of shift, *F*(1, 324) = .01, *p* = .93, *r* < .01. This suggests two possibilities: participants may have mistakenly attended to only end points in the experiment, or participants may genuinely care more about endpoints than sizes of shift in harm.

By employing a stringent attention check that asked participants to identify starting and ending probabilities using free recall, we could explore whether size of shift mattered for participants who had correctly identified both probabilities. Two hundred thirty five people (71%) correctly identified the starting and ending probabilities after being asked to respond to our dependent variable. Notably, the pattern of results among those participants was similar to the full sample: A 2x2 ANOVA, with size of shift and end-point entered as factors again revealed no significant effect of size of shift, *F*(1, 233) = 2.70, *p* = .10, *r* = .11 (mean 50% shift = -.63, SD = 3.14; mean 25% shift = .05, SD = 3.12). Thus, participants who had verifiably attended to both start-state and end-state probabilities, and were therefore aware of the shift size, were nonetheless insensitive to shift size. Consistent with the larger sample, we continued to see a significant main effect of end point, *F*(1, 233) = 66.6, *p* <.001, *r* = .47 (mean 50% end-state = 1.19, SD = 2.84; mean 100% end-state = -1.77, SD = 2.71). There was no significant interaction between end point and size of shift, *F*(1, 231) = .03, *p* = .86, *r* = .01. Planned contrasts between shifts that resulted in the same end-state level of risk to the bystanders were all non-significant (25-50 vs. 0-50: *t*(119) = 1.28, *p* = .20, *d* = .31 (mean 25-50% = 1.45, SD = 2.70; mean 0-50% = .84, SD = 3.00); 50-100 vs. 75-100: *t*(113) = .92, *p* = .36, *d* = .18, (mean 50-100% = -2.02, SD = 2.60; mean 75-100% = -1.54, SD = 2.81)). Thus, our findings suggest that ending probabilities, rather than the difference in the sizes of probability shifts or starting point, matter to participants, and that this is not the result of having only attended to endpoints.

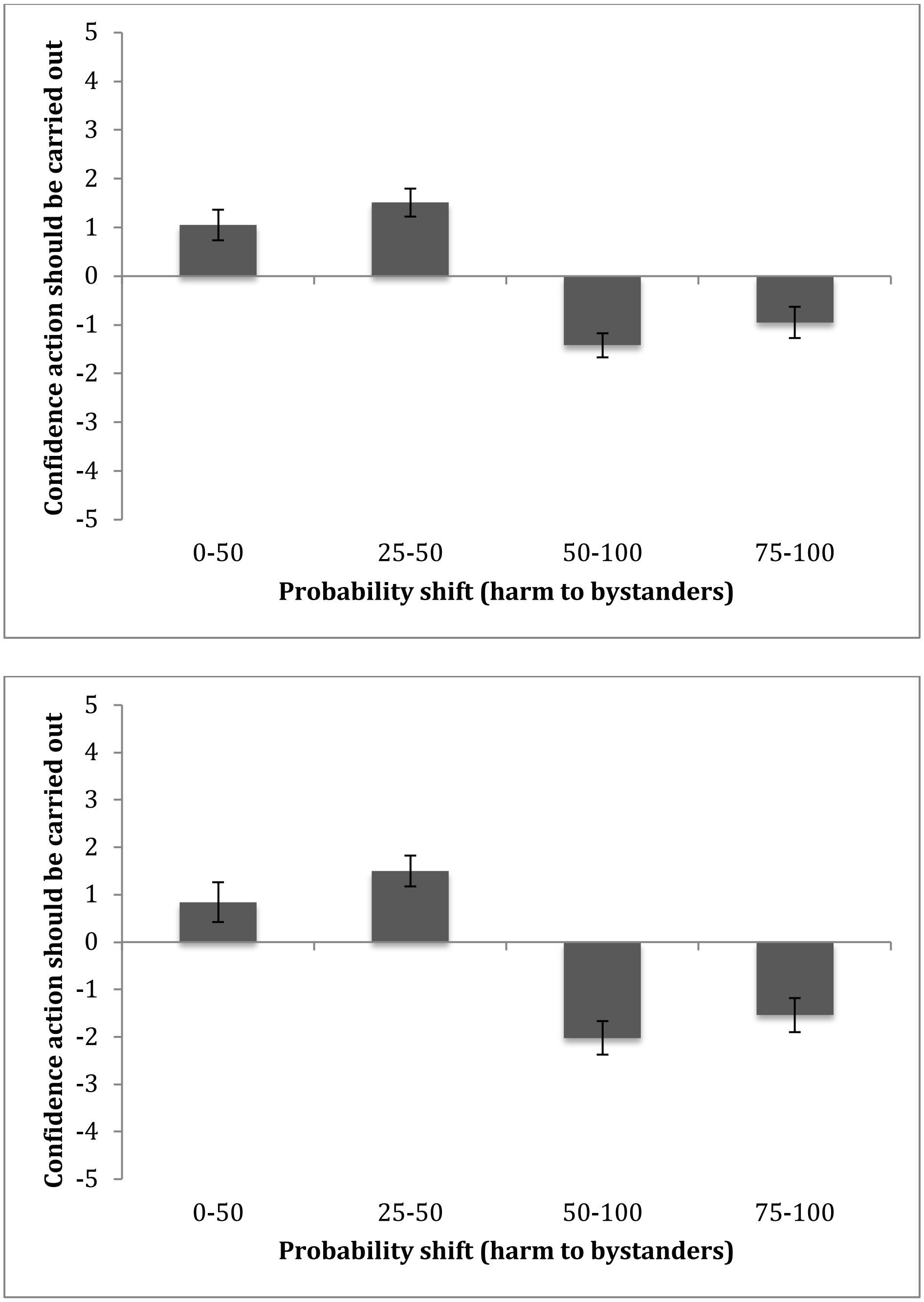


Figure 3.2. Sensitivity to size of shift and end-state likelihood of harm to bystanders for all participants (a) and participants correctly recalling the initial and end state probability of harm (b). Error bars represent one standard error.

Study 2 Discussion

Participants were more sensitive to the end-state probability than to the size of the shift in probability of harm to the bystanders, despite being able to recall initial and final probabilities for the scenarios they read. This suggests that participants care more about what level of risk of harm the bystanders end up at than how much the risk of harm is increased for the bystanders.

Because end points matter more than shift size, it raises the question of just how much people are willing to trade off the two. Large increases in probability of harm with relatively low end points (e.g., 0%-90%) might be viewed as more acceptable than small increases in probability of harm with high endpoints (e.g., 90%-100%). And, if so, we were also interested in whether participants' more reflective preferences under joint evaluation would be consistent with their preferences under single evaluation.

Study 3a and 3b

Studies 3a and 3b continued exploring how much the location of probability shifts matters. To examine the extent to which participants would prefer a larger probability increase in harm with a lower end state to a small probability increase with a higher end-state probability, we set out to estimate a balance point value, X, such that participants would no longer prefer the 0% to X% plan over the X% to 100% plan. For a consequentialist, X would be 50 (i.e., indifferent between an increase in harm from 0%-50% and an increase from 50% to 100%). However, sensitivity to end points suggests that X will be greater than 50%. We examine these preferences under single and joint evaluation, in order to see whether judgments made under single evaluation would withstand reflection under joint evaluation.

Study 3a

Study 3a Methods

Three hundred and ninety five participants were recruited via Amazon's Mechanical Turk (327 passed an attention check; because results did not significantly differ between the full sample and those passing the attention check, all participants were retained for analysis; 55.4% female, mean age = 33.9, SD = 11.6). The study was comprised of a single evaluation task and joint evaluation task. In the single evaluation task, each participant was randomly assigned to read one of eight scenarios. Four of the eight scenarios involved an increase from 0% to x% (x = 50, 75, 85, 95), whereas the other four involved an increase from x% to 100% (x = 50, 75, 85, 95). For example, the single evaluation task that involved an increase from 0% to 95% went as follows:

Harry sees a group of two people who will certainly die without intervention. He knows the following facts. There is a group of four bystanders that is facing a 0% risk of death. Harry can carry out a plan that will certainly save the group of two people. However, in carrying out the plan, Harry will increase the risk of death for the group of four bystanders from 0% to 95%.

Participants were asked, for example, *Should Harry carry out a plan that he knows with certainty will both save the group of two people and at the same time raise the risk of death for the group of four bystanders from 0% to 95%?* (-5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*). After completing the single evaluation, participants then read a joint evaluation scenario in which participants had to choose between two plans, labeled as Plan X and Plan Y. One of these two plans was the plan participants had read under single evaluation, the second was its matched pair, such that participants saw matched 0% to X% and X% to 100% plans (see Table 3 for scenarios and pairings). For example, participants who had rated either the 0% to 95% plan or the 95% to 100% plan under single evaluation, chose between a 0% to 95% plan and 95% to 100% plan in the joint evaluation task, as follows:

Harry sees a group of two people who will certainly die without intervention. He knows the following facts. There are two groups, A and B, with four bystanders in each group. Group A is facing a 0% risk of death, and Group B is facing a 95% risk of death. Harry can carry out a plan, Plan X, that will save the group of two people, but raise the risk of death for the A group of four bystanders from 0% to 95%. Alternatively, he can carry out a plan, Plan Y, that will save the group of two people, but raise the risk of death for the B group of four bystanders from 95% to 100%. He only has time to carry out one of his plans.

Participants were then asked, *Assuming that Harry must carry out one of the two plans, which should he carry out: Plan X, which he knows with certainty will both save the group of two people and at the same time raise the risk of death for the A group of four bystanders from 0% to 95%; or Plan Y, which he knows with certainty will both save the group of two people and at the same time raise the risk of death for the B group of four bystanders from 95% to 100%?*

Table 3.3. Study 3a scenarios.

Probability shift	Harm Scenarios	Scenario Pairing	EV ratio of action
0% to 50%	Increase chance of 4 people dying from 0% to 25% to save 2 people	A	1
50% to 100%	Increase chance of 4 people dying from 50% to 100% to save 2 people	A	1
0% to 75%	Increase chance of 4 people dying from 0% to 75% to save 2 people	B	.66
75% to 100%	Increase chance of 4 people dying from 75% to 100% to save 2 people	B	2
0% to 85%	Increase chance of 4 people dying from 0% to 85% to save 2 people	C	.59
85% to 100%	Increase chance of 4 people dying from 85% to 100% to save 2 people	C	3.33
0% to 95%	Increase chance of 4 people dying from 0% to 95% to save 2 people	D	.53
95% to 100%	Increase chance of 4 people dying from 95% to 100% to save 2 people	D	10

Study 3a Results:

In this study, we were interested in whether, upon reflection, participants would endorse an x-100 plan over a 0-x plan. We could thus identify the balance point (X) at which people flip from preferring a 0-x plan to an x-0 plan under joint evaluation, and whether this tracks their preferences under single evaluation.

First, we examined preferences under single evaluation. We observed a significant interaction between where the balance point was set (x) and whether the shift in harm resulted in a probabilistic or certain harm end-state (0-x or x-100), $F(1, 391) = 28.4$, p <.001, $r^2 = .0725$ (see Figure 3). Consistent with Study 2 findings, though participants endorsed a plan that raises the likelihood of four bystanders dying from 0-50% to save a group of two, they rejected a plan that raised the likelihood of the four bystanders dying from 50-100% to achieve the same result, $t(96) = 3.48$, $p < .001$, $d = .70$ (mean 0-50% =

.92, *SD* = 2.96; mean 50-100% = -1.10, *SD* = 2.79). However, participants preferred a plan that increases the chances of harm from 95-100% over a plan that shifts risk of harm to the four from 0-95%, *t*(98) = 4.55, p <.001, *d* = .91 (mean 0-95% = -1.85, *SD* = 2.80; mean 95-100% = .75, *SD* = 2.90. There was no preference between plans that raised the likelihood of harm from 0-75% and 75-100%, *t*(101) = .04, *p* = .96, *d* = .01 (mean 0-75% = -.58, SD = 2.95; mean 75-100% = -.60, SD = 2.95), nor between plans that raised the likelihood of harm from 0-85% and 85-100%, *t*(92) = .69, *p* = .49, *d* = .14 (mean 0-85% = -.69, SD = 3.04; mean 85-100% = -.24, SD = 3.25). Thus, under single evaluation, sensitivity to expected value and to end-state were tied when the balance point (X) was set at 75.

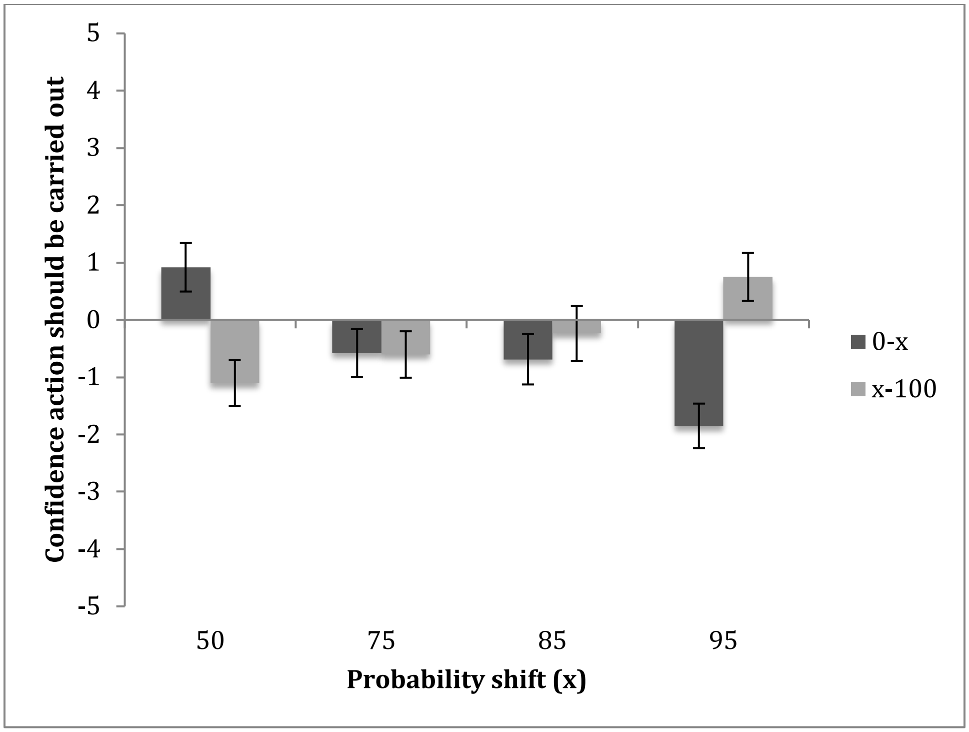


Figure 3.3. Sensitivity to size of shift in probability of harm to bystanders for shifts that increase the likelihood of harm from 0-x%, or from x-100%, under single evaluation. Error bars represent one standard error.

Under joint evaluation, we were interested in whether participants would endorse the patterns observed under single evaluation (e.g., whether participants endorse the idea that raising the risk of harm to four bystanders from 0-50% is preferable to raising the risk of harm to four bystanders from 50-100%). While a preference for 0-50% over 50-100% could be normatively defensible, in that it would not be unreasonable for avoiding certain harm to be a tie-breaking preference between two actions with matched expected values, a preference for 0-85% over 85-100% (or indifference between 0-95% and 95-100%) is less clearly defensible, in that the expected value of the 0-85% action is such that it does much more harm than good (EV ratio = .59), whereas the 85-100% action does much more good than harm (EV ratio = 3.33).

Participants endorsed the pattern of results observed under single evaluation for 0-50% and 50-100% plans, by endorsing 0-50% plans under joint evaluation, despite their equivalent expected value, $t(97) = 6.19$, $p < .001$, $d = .63$ (difference from 0; mean = -1.89, SD = 3.01), see Figure 4. Despite not having a preference under single evaluation between a plan that raises the risk of harm to four bystanders from 0-75% to one that raises the risk of four bystanders dying from 75-100%, participants preferred the 0-75% plan under joint evaluation, $t(102) = 2.78$, $p = .007$, $d = .27$ (mean = -.86, SD = 3.15). Likewise, despite not having a preference under single evaluation between a plans under single evaluation, participants preferred the 0-85% over the 85-100% plan under joint evaluation, $t(93) = 3.15$, $p = .002$, $d = .32$ (mean = -1.03, SD = 3.18). Perhaps most surprisingly, despite clearly endorsing the 95-100% plan and rejecting the 0-95% plan under single evaluation, under joint evaluation participants became indifferent between the 95-100% plan, which had an expected value ratio of 10 (causing much benefit for little harm), and the 0-95% plan, which had an expected value of .53 (causing much more harm than benefit), *t*(99) = 1.50, *p* = .13, *d* = .15 (mean = -.49, SD = 3.28). Thus, participants were willing to pay a tremendous expected value premium to avoid raising the likelihood of death to the bystanders to 100%: Sensitivity to expected value and to end-state, under joint evaluation, was achieved only once the balance point (X) was set at 95.

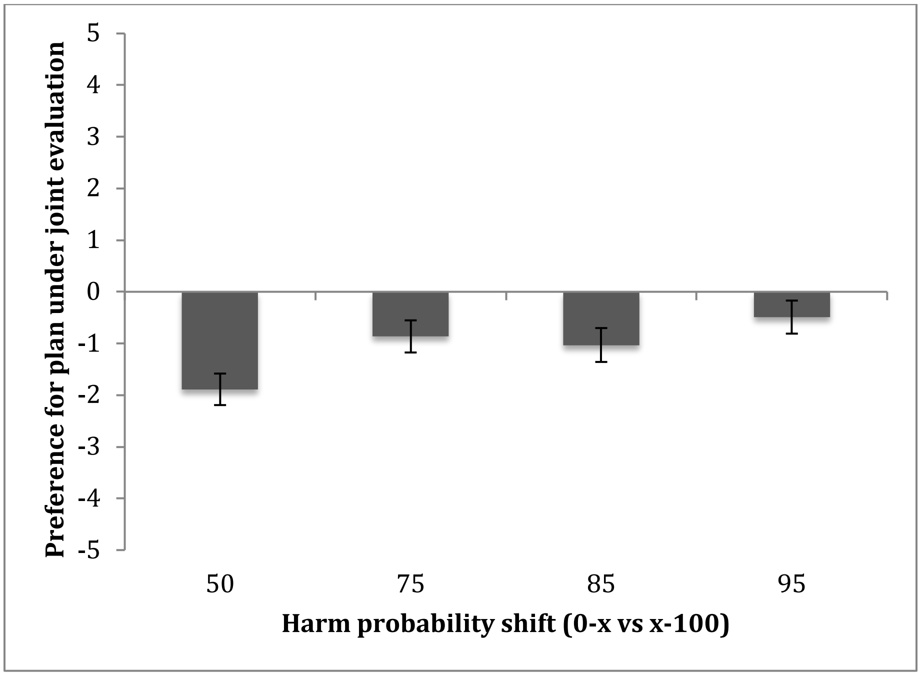


Figure 3.4. Preference for plans that increase harm from 0-x% or x-100% to four bystanders in order to save two people, under joint evaluation. Negative numbers indicate a preference for 0-x plans. Error bars represent one standard error.

One unexpected finding was that an order effect emerged, such that participants who had seen 0-x plans were more likely to endorse them *F*(1, 387) = 8.76, *p* < .01 (mean 0-x% = -1.52, SD = 3.04; mean x-100% = -.60, SD = 3.27), across all scenarios. Supplementary studies were conducted using the x = 50 and x = 95 conditions, where the joint evaluation task was presented without preceding single evaluations. The results showed the same pattern, with participants being undecided at x = 95 (*t*(112) = .84, *p* = .40, *d* = .08 (mean = -.26, *SD* = 3.24)) and having a strong preference at x = 50 (*t*(106) = 13.9, *p* < .001, *d* = 2.34, (mean = -2.89, SD = 2.15). See supplementary materials for details.

Study 3a discussion

Under joint evaluation, not only do participants endorse the position that it is better to raise the risk of one dying from 0-50% than to raise the risk of another dying from 50% to 100%, but they are willing to pay a premium to avoid raising the risk to the group already at risk. Specifically, they prefer a plan that raises the risk of death for four bystanders from 0% to 85% over a plan that raises the risk of death for another group of four bystanders from 85% to 100%. By contrast, single evaluation tasks don’t show a strong preference for the 0% to 85% plan. Furthermore, under joint evaluation participants are tied between 0-95% plans and 95-100% plans, despite having a strong preference for 95-100% plans under single evaluation. We interpret the joint evaluation tasks as revealing people’s considered judgments, and thereby should be taken seriously for the purpose of understanding lay moral theory, as well as for moral theorists’ project of engaging in reflective equilibrium in defense of a correct set of moral principles that mutually support and explain our reactions to particular scenarios. But this is subject to further scrutiny. Both the single and the joint evaluation results add support to our Study 2 finding that the end-state of probability shifts matters more than the size of the probability shifts. Surprisingly, joint evaluations (balance point = 95) deviate further from sensitivity to expected value than single evaluations (balance point = 75) for shifts in likelihood of harm. We next examined whether joint evaluation judgments of shifts in location on the saving side would be consistent with single evaluation insensitivity to location observed in Study 1.

Study 3b

We have competing predictions regarding what will happen under joint evaluation for varying shifts on the saving side. It could be that people continue to be insensitive to probability shifts under joint evaluation, if people genuinely do not believe that it is better to definitely save than to reduce definite risk. Alternatively, it could be that a preference for certain saving emerges under joint evaluation, such that, for example, when choosing between a plan that can decrease risk of death for one group of people from 50% to 0%, or decrease the risk of death for another group from 100% to 50%, people prefer the plan that results in certain saving.

Study 3b methods

One hundred seventy five participants were recruited for this study (141 passed an attention check, because results did not significantly differ between the full sample and those passing the attention check, all participants were retained for analysis, 63.4% female, mean age = 35.3, SD = 11.0). Study 3b was structurally similar to Study 3a: Participants were asked to perform a single evaluation task and a joint evaluation task. But the actions in the scenarios now involved shifts in the likelihood of a group *dying*, at the cost of killing a bystander. Though we had not observed sensitivity to location of saving probability shift in Study 1, it is possible that a preference for certain saving could emerge under joint evaluation. In order to examine this, we adapted the Study 3a scenarios to describe an action that could decrease the likelihood of the group of eight dying by certainly killing a bystander. We selected new balance point values (X), such that the action could reduce the likelihood of death for the eight from 100-x% or from x-0%. Selecting x = 50 as a balance allowed us to examine this preference at tied expected values (EV = 4 for both plans). A second balance point, x = 25%, allowed us to determine whether any potential emergent joint evaluation preference for certain saving, observed under tied expected values, could outweigh sensitivity to the expected value of the action when expected values are unmatched (25- 0% EV = 2; 100-75% EV = 6). Participants were randomly assigned one of four plans: two of the plans involved killing a bystander to decrease the risk of eight dying from x% to 0% (x = 50, 25), and two involved killing a bystander in order to decrease the chance of eight people dying from 100% to x% (x = 50, 25), see Table 4. One scenario, for instance, was as follows:

Harry sees a group of eight people whose lives are in danger. He knows the following facts. There is a 50% chance of the group of eight people dying. Harry can carry out a plan that will reduce the risk of the group of eight dying from 50% to 0%. However, in carrying out the plan, Harry will certainly kill one bystander.

After evaluating one of the two plans, (e.g. *Should Harry carry out a plan that he knows with certainty will both reduce the risk of the group of eight dying from 50% to 0% and at the same time kill one bystander?* ; -5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*), participants were exposed to the plan they had seen and its matched pair, such that they would see both a 0-X and corresponding X-100 plan, labeled as Plan X and Plan Y:

Harry sees two groups, A and B, with eight people in each group whose lives are in danger. He knows the following facts. Group A is facing a 50% chance of dying, and Group B is facing a 100% chance of dying. Harry can carry out a plan, Plan X, that will reduce the chance of Group A dying from 50% to 0% but certainly kill one bystander. Alternatively, he can carry out a plan, Plan Y, that will reduce the chance of Group B dying from 100% to 50% but certainly kill one bystander. He only has time to carry out one of his plans.

Participants were then asked to compare the two plans as follows: *Assuming that Harry must carry out one of the two plans, which should he carry out: Plan X, which he knows with certainty will both reduce the risk of the A group of eight dying from 50% to 0% and at the same time kill one bystander; or Plan Y, which he knows with certainty will both reduce the risk of the B group of eight dying from 100% to 50% and at the same time kill one bystander?* (-5 – 5; *Very confident Harry should carry out Plan X –Not at all confident either way—Very confident Harry should carry out Plan Y*).

Table 3.4. Study 3b scenarios.

Probability shift	Save Scenarios	Scenario Pairing	EV ratio of action
50% to 0%	Kill 1 bystander to decrease chance of 8 people dying from 50% to 0%	A	4
100% to 50%	Kill 1 bystander to decrease chance of 8 people dying from 100% to 50%	A	4
25% to 0%	Kill 1 bystander to decrease chance of 8 people dying from 25% to 0%	B	2
100% to 25%	Kill 1 bystander to decrease chance of 8 people dying from 100% to 25%	B	6

Study 3b results:

Under single evaluation, overall, plans that reduced the likelihood of the eight dying from 100% to x were rated more favorably than plans that reduced the likelihood of the eight dying from x to 0%, *F*(1, 170) = 5.74, *p* = .018, *r* = .15 (mean 100-x% = 1.77, SD = 2.46; mean x-0% = .71, SD = 3.19), see Figure 5. There was no main effect of where the balance point x was set (whether x = 25 or 50), *F*(1, 170) = .30, p = .58, *r* = .04 (mean 25 = 1.08, SD = 2.88; mean 50 = 1.33, SD = 2.95), or interaction between where the balance point was set and whether the probability shift resulted in probabilistic or certain saving (100-x% or x-0%), *F*(1,168) = .62, *p* = .43, *r* = .06. The main effect, however, was driven primarily by the 25-0% vs. 100-25% comparison, *t*(82) = 2.26, *p* = .026, *d* = .50 (mean 25-0% = .47, SD = 2.53; mean 100-25% = 1.87, SD = 2.53) rather than the 50-0% vs 100-50% comparison *t*(86) = 1.12, *p* = .27, *d* = .24 (mean 50-0% = .98, SD = 3.39; mean 100-50% = 1.68; SD = 2.43). These results suggest that, under single evaluation, participants were indifferent between plans with identical expected values but that varied in probability shift end-state (50-0% and 100-50%). Participants, were, however, sensitive to the differences in expected value between the 100-25% (EV = 6) plan and the 25-0% plans (EV=2), preferring the plan that resulted in probabilistic saving over the plan that resulted in certain saving, but at a lower expected value. Consistent with Study 1 findings, we observed insensitivity to equivalently-valued likelihood shifts on the saving side. We additionally found sensitivity to size of shift on the saving side, which did not appear to compete with a preference for a certain saving end-state.

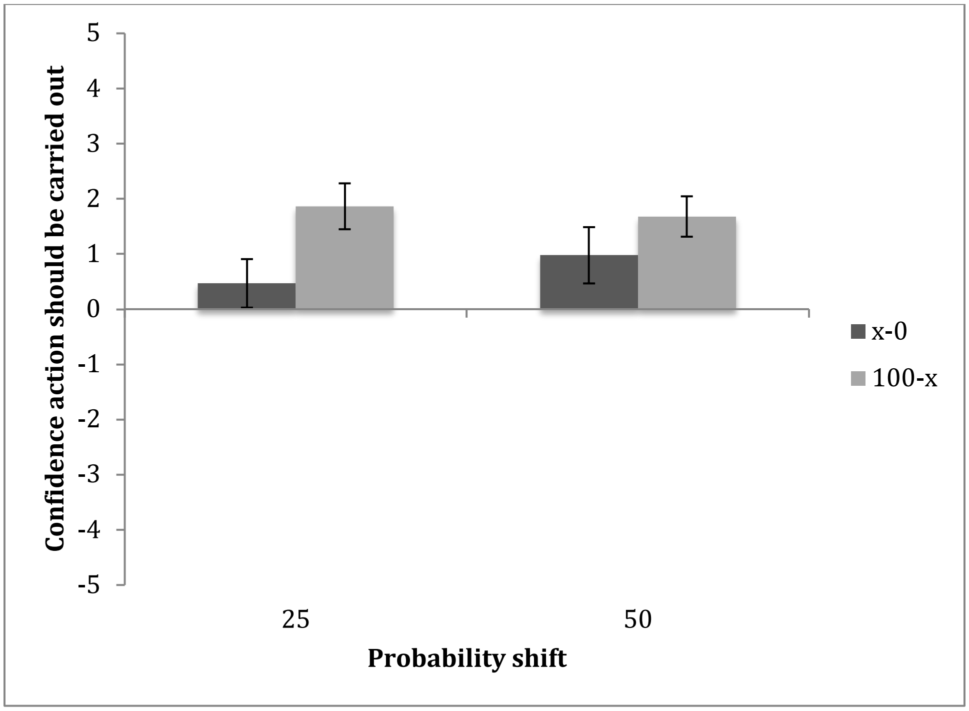


Figure 3.5. Preference for plans that reduce the likelihood of a group of eight dying from 0-x% or x-100% but will kill a bystander, under single evaluation. Error bars represent one standard error.

Under joint evaluation, participants were indifferent between the plan that reduced the likelihood of 8 dying from 50-0% and 100-50%, *t*(83) = .43, *p* = .66, *d* = .05 (mean = -.15, SD = 3.17), consistent with insensitivity to end-state of probability shift on the saving side observed under single evaluation, see Figure 6. Unlike under single evaluation, however, participants were indifferent between the 25-0% and 100-25% plans, despite the 100-25% plan having a higher expected value than the 25-0% plan *t*(83) = .97, *p* = .33, *d* = .11 (mean = .36, SD = 3.36). Thus, under joint evaluation, we again saw no preference for certain saving end-states, and saw a decreased sensitivity to expected value, as compared to judgments made under single evaluation.

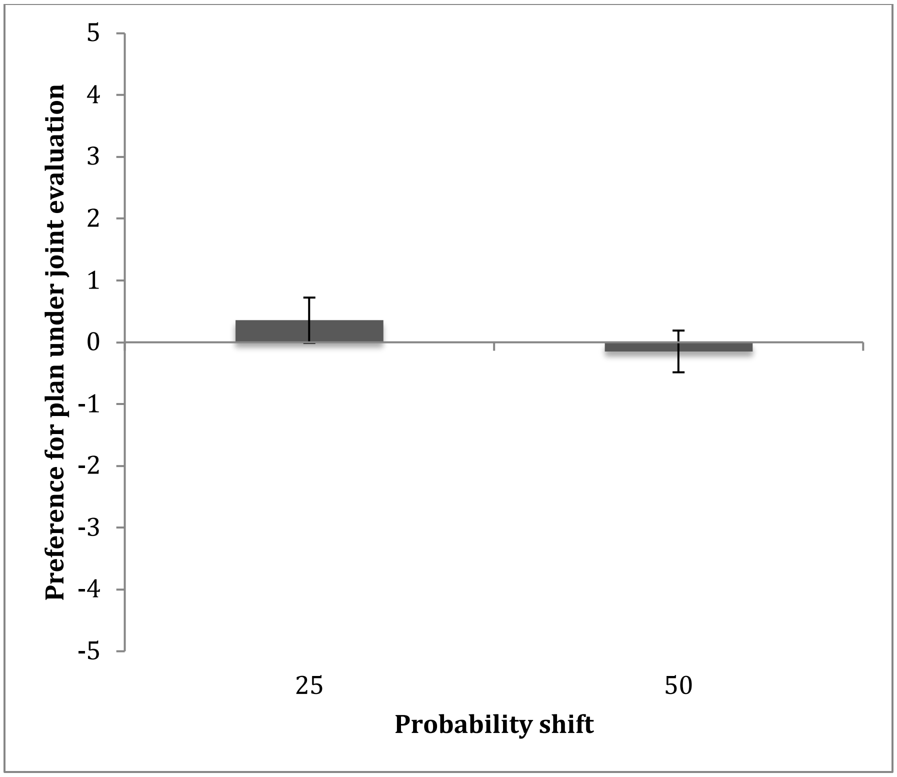


Figure 3.6. Preference for plans that reduce the likelihood of a group of eight dying from 0-x% or x-100% but will kill a bystander, under joint evaluation. Error bars represent one standard error.

Study 3b Discussion

Under joint evaluation, participants do not appear to endorse a preference we observed under single evaluation for a 100-25% shift in in likelihood of the group of eight dying. This preference can be interpreted as a sensitivity to size of shift, as participants Study 1 were insensitive to location shifts involving saving probabilities that varied start-states and end-states. The overall findings, however, are consistent with the decreased sensitivity to saving shifts, as compared to harming shifts that we observed in prior studies. Again, here we take the results from joint evaluations to be participants' considered preferences and to reflect their lay moral theory. Though the lack of sensitivity to the location of probabilistic benefit confirms our hypothesis that there is more sensitivity on the side of harm than on the side of benefit, it comes as a surprising finding that other than size of shift being relevant under single evaluation, we observe no sensitivity under joint evaluation for this. Though the result is broadly consistent with existing findings on the harm/benefit asymmetry, according to which moral reasoning may be more fine-grained in sensitivity to shifts in harm than in shifts in benefit, it is still surprising that folk moral psychology is so coarse-grained that it doesn't even differentiate between, for example, decreasing the probability of people dying from 100% to 25% and decreasing the probability of the same number of people dying from 25% to 0% under joint evaluation.

The general harm vs. benefit asymmetry as observed in Study 1, and reconfirmed in Studies 3a and 3b, is consistent with the hypothesis we started with: There should be more sensitivity on the harm side, because we have a right or claim to not have our probability of dying increased whereas we do not have a right to have our probability of living increased.

Studies 4a and 4b: Mediations

Throughout the preceding studies, we observed a sensitivity to where shifts in likelihood of harm occur, and general insensitivity to where shifts in saving likelihood occur. One question that emerges is why such a shift occurs. One possibility is that participants perceive different harm probability shifts as differentially harmful, consistent with the dyadic model of morality (Schein and Gray, 2018). Though Schein and Gray's model focuses almost exclusively on perceived harm, it is plausible that perceived benefit might play a mediating role as well, especially in cases involving probabilistic saving. In the following studies, we explore whether differences in perceived harm and perceived benefit are causal in divergent moral judgments of the action.

Study 4a Design

Study 4a asked participants to evaluate one of two plans, as they had in Study 1. The first plan involved increasing the risk of dying to a group of four bystanders from 0-25% in order to save two people. The other plan involved a 75-100% increase in the chance of the group of four bystanders dying to save the two. After evaluating one of the two plans, participants were asked to answer two questions: one regarding whether Harry should carry out the plan (*Should Harry carry out a plan that he knows with certainty will both save the group of two people and at the same time raise the risk of death for the group of four bystanders from 0% to 95%?* ; -5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*), and a second regarding how harmful the action would be to the group of bystanders (e.g. *How harmful is Harry's plan for the four bystanders?;* 0: *not at all harmful*, to 10: *extremely harmful*). Questions were presented in a randomized order. Two hundred twenty two participants were recruited using Amazon’s Mechanical Turk (163 passed an attention check; because results do not significantly differ between full sample and those passing attention check, all participants were retained for analysis; 54.1% female, mean age = 35.1, SD = 11.5).

Study 4a Results

In order for us to test whether the perceived harm to the bystanders mediates moral judgment of actions that carry a variable shift in likelihood of harm, we first tested the significance of each individual path. The relationship between harm probability shift (0-25% or 75-100%) and confidence in the morality of action was partially mediated by the perceived harmfulness of the action for the group of four bystanders. The regression of probability shift on confidence in the morality of action was statistically significant ($\beta = .39$, $t(220) = 6.34$, $p < .001$), as was the regression of probability shift on perceived harmfulness ($\beta = .53$, $t(220) = 9.40$, $p < .001$), and the regression of perceived harmfulness on confidence in the morality of action ($\beta = .37$, $t(220) = 5.91$, $p < .001$), see Fig. 1. The standardized indirect effect was (.53)(.37) = .44. The significance of this indirect effect was tested using bootstrapping procedures: unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect partially mediated the relationship between probability shift and judgments of the action (mediated effect = 1.06, $p < .001$, 95% CI [.50, 1.66]; direct effect = 2.39, $p < .001$, 95% CI [1.38, 3.43]).

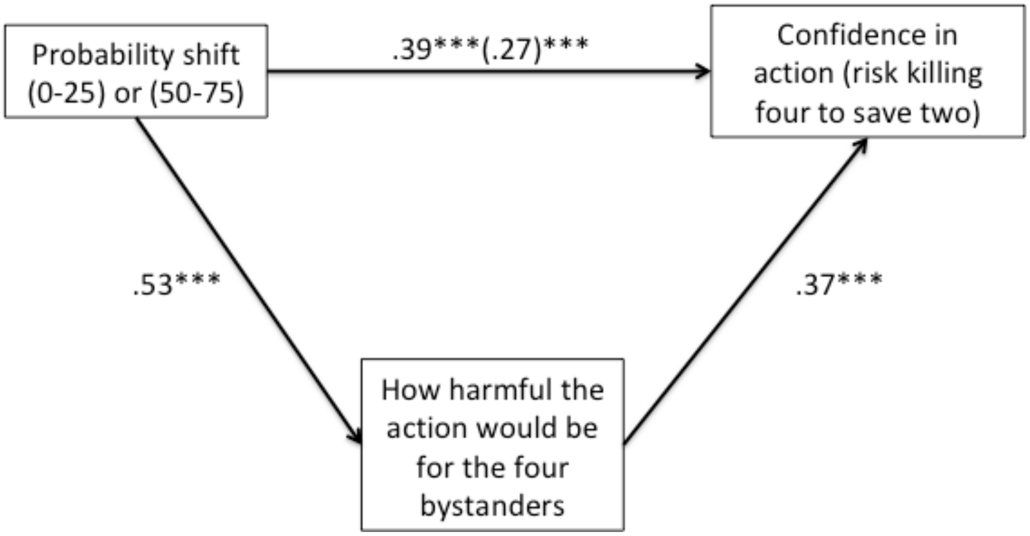


Figure 3.7. Standardized regression coefficients for the relationship between harm probability shift and confidence in the morality of carrying out an action that raises the risk of four bystanders dying to save people as mediated by perceived harmfulness of the action for the four bystanders. The standardized regression coefficient between probability shift and confidence in the morality of action, controlling for perceived harmfulness, is in parentheses.
****p* < .001.

Study 4a Discussion

Perceived harm partially mediated the difference between plans that increase the risk of bystanders dying from 0-25% and plans that increase the risk of bystanders dying from 75-100%, consistent with predictions made by the theory of dyadic morality. We next explore whether we see a similar effect for saving.

Study 4b

Study 4b Design

Study 4b adopted the approach of Study 4a to scenarios in which the saving is probabilistic, and in which shifts in saving likelihood vary. Study 4b asked participants to evaluate one of two plans. The first plan would result in the death of one bystander in order to decrease the likelihood of a group of eight people dying from either 25-0% or 100-75%. After evaluating one of the two plans, participants were asked to answer two questions: one regarding whether Harry should carry out the plan (*Should Harry carry out a plan that he knows with certainty will both reduce the risk of the group of eight dying from 25% to 0% and at the same time kill one bystander?* ; -5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*), and a second regarding how beneficial the action would be for the eight people (e.g. *How beneficial is Harry's plan for the eight people whose lives are in danger?;* 0: *not at all beneficial*, to 10: *extremely beneficial*). Questions were presented in a randomized order. Two hundred twenty nine participants were recruited using Amazon's Mechanical Turk (165 passed an attention check; because results do not significantly differ between full sample and those passing attention check, all participants were retained for analysis; 56.9% female, mean age = 32.4, SD = 10.7).

Study 4b Results

In order for us to test whether perceived benefit mediates confidence in the morality of actions that carry a variable shift in likelihood of benefit, we first tested the significance of each individual path. The regression of probability shift on confidence in the morality of action was not statistically significant ($\beta = .07$, $t(223) = 1.08$, $p = .28$). The non-significant effect observed here is consistent with Study 1 findings, and precludes mediation analyses. However, we can still examine whether perceived benefit varies where the shift in outcome likelihood for saving occurs. The regression of probability shift on perceived benefit was significant, ($\beta = .20$, $t(221) = 3.05$, $p < .01$, as was the regression of perceived benefit on confidence in the morality of action (β = .28, t(221) = 4.32, p < .001), suggesting that perceived benefit is sensitive to outcome probability shifts, and that perceived benefit does relate to confidence in the morality of action, though not in a way that contributes to overall differences between the two versions of the scenario, see Figure 8.

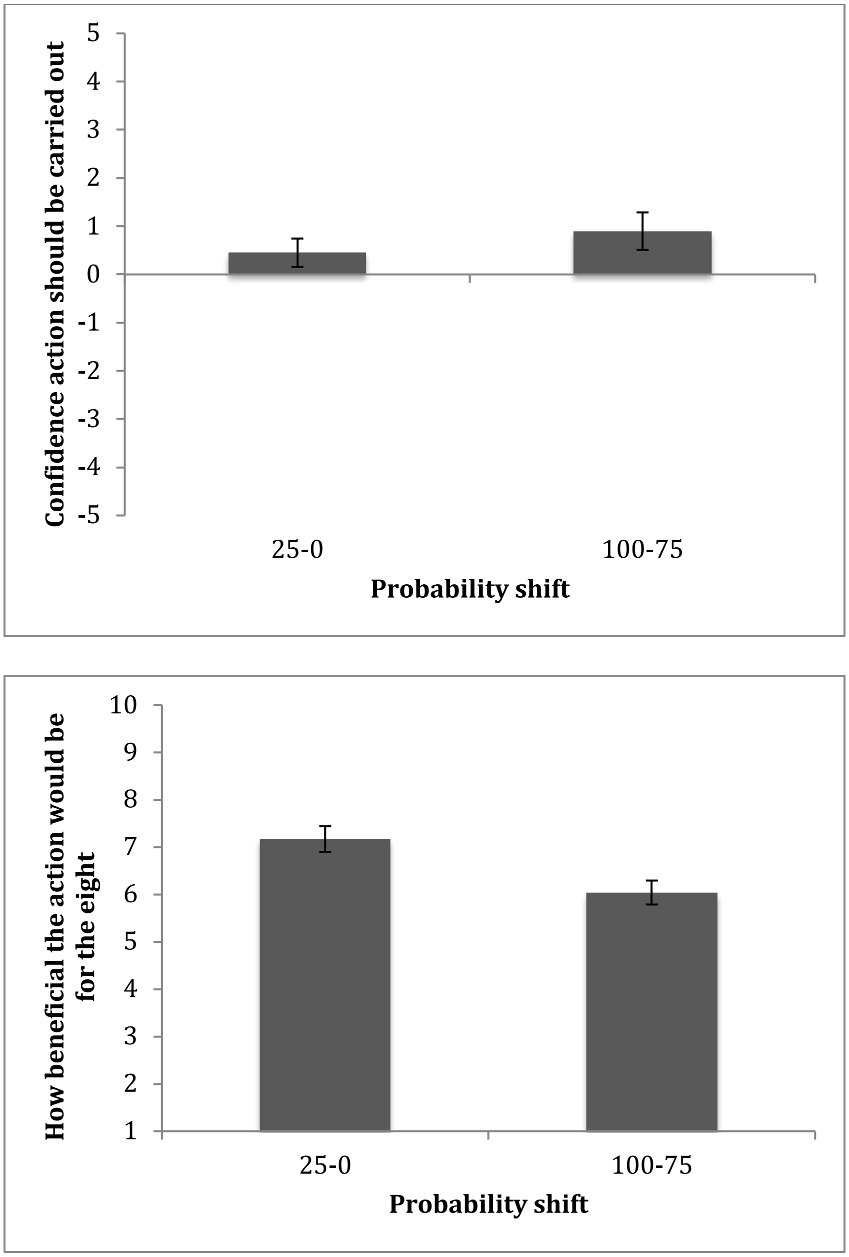


Figure 3.8. Confidence in action for plans that reduce the likelihood of a group of eight dying from 25-0% or 100-75% but will kill a bystander (a). Ratings of how beneficial the action would be for the eight (b). Error bars represent one standard error.

Study 4b Discussion

Though perceived benefit does correlate with confidence in the morality of action more generally, it does not mediate a difference between moral judgments where the saving probability differs between two otherwise identical moral dilemmas.

Studies 4a-b Discussion

We see that divergent moral judgments arise between saving probability shifts because shifts are perceived as differentially harmful. The effect size of where the shift in outcome likelihood for saving occurs on perceived benefit ($\beta = .20$) was less than half the effect size of where the shift in outcome likelihood for harming occurs on perceived harm ($\beta = .53$), suggesting that participants were differentiating more between location-shifts for harm than for benefit, consistent with a more general greater differentiation of negative events than positive events other researchers have observed (e.g., Guglielmo & Malle, 2019).

General Discussion

Our results suggest that participants are sensitive not only to expected value and probability in moral dilemmas, but also to where the shift in probability occurs, for increases of risk to bystanders. Study 1 identified that participants are sensitive to where shifts in the likelihood of harm to a group of bystanders occur, but not to where shifts in the likelihood of saving a group occur. Study 2 found that participants are sensitive to end-state probability, rather than to the size of the shift or start-state probability, for shifts in likelihood of harm to bystanders. Study 3a explored whether participants would endorse such a position upon reflection, under joint evaluation, and found that participants prefer plans that raise the likelihood of harm for a group of four bystanders from 0-85%, in order to save two people, over plans that raise the likelihood of harm from 85-100% for a different group of four bystanders. In Study 3b, participants were found to continue being indifferent between location of shifts in likelihood of saving a group under joint evaluation. Study 4a identified perceived harm as a mechanism behind divergent moral judgments for actions that cause mathematically equivalent harm likelihood shifts that occur at different parts of the probability distribution. Study 4b found that, though saving probability shifts are perceived as differentially beneficial, differences in perceived benefit between probability shifts do not result in divergent moral judgments.

In general, the results from these six studies show that participants are strongly sensitive to the location of probability shifts for harm. This phenomenon makes it difficult to render folk moral judgments consistent with the traditional consequentialist framework which takes expected value to be the single currency for moral permissibility and obligation. At the same time, the traditional tools of the deontologist, which presuppose the existence of moral constraints on certain harming, do not capture this phenomenon without further elaboration. We thus need a more nuanced way to describe the normative principle that can explain why participants reason in this way. We think one promising strategy is to develop more fine-grained deontological constraints that concern moral decision making under uncertainty. Such a strategy would need to, for example, explicate the force of deontological constraints in terms of the location of probability shifts. For example, it may be that the right against raising a person’s risk of death to 100% is more stringent than the right against raising a person’s risk of death to

50%, even when the probability shift in the latter case is greater. We are optimistic that this kind of framework can be worked out, but a complete defense has to be left for another occasion.

The more practical applications of the sensitivity to locations of probability shifts can be seen, for example, in the design of autonomous vehicles. Autonomous vehicles all face "decisions" under uncertainty; it is therefore worth examining how complicated the probabilistic computing would need to be. And our results show that the computation has to involve at least two distinct probability estimates: an estimate of initial (or final) probabilities of various outcomes and an estimate of the size of probability shifts. That is, it is not enough for autonomous vehicles to simply calculate *how much* change in the probability of harming people would be involved in alternative courses of action. One implication of the finding that participants care more about end-state harm likelihoods than sizes of shifts is that initial risk levels are largely irrelevant to the endorsement of action in moral dilemmas. Such findings help inform where resources should be directed in the design of detectors in autonomous vehicles that will serve as inputs for ethical decisions these vehicles will be programmed to make: whether the pedestrian that can be sacrificed to save a family in a vehicle ahead is already at some risk of dying is less important than what level of risk redirecting the autonomous vehicle would bring her, outside of situations where the pedestrian is already at great risk of dying from the impending collision.

Our results also show that locations of probability shifts matter in ways that can even make participants indifferent as between probability shifts of the same size, and even prefer a *larger* size of increase of the probability of people dying just because the increase occurs in the preferable location. Thus, the location isn't merely a "tie-breaker" for participants but rather plays a substantial role in their decision-making process.

We also see a general harm/benefit asymmetry in the sensitivity to locations of probability shifts. This pattern of asymmetry adds further support to the claim that we are sensitive to probability shifts in harm in a more fine-grained way than we are to probability shifts in benefit. However, the finding that there is no significant variation according to the locations of probability shifts on the benefit side is surprising to us, potentially suggesting a deeper harm/benefit asymmetry than first appears plausible. We speculate that the explanation is that we have a right or claim to not be harmed but no right or claim to be benefited. This is consistent with the suggestion of dyadic morality theory, according to which perceived harm plays a much more central role in moral judgments, in contrast to perceived benefit.

Finally, single and joint evaluations result vary in our studies. Like others that have advocated for joint evaluations to be taken more seriously, since they indicate participants' considered preferences under reflection (Barak-Corren et al., 2018), we find moral judgment under single and joint evaluation to diverge, calling into question the common method both in philosophy and in psychology to simply test for moral permissibility by appealing to intuitions about single cases. However, we do not find judgments under joint evaluation to be more clearly defensible, though we can say that joint evaluation judgments of actions involving shifts in harm or saving deviate further from expected value calculation than single evaluations for some actions. A complete moral theory would benefit from further comparison across both kinds of evaluation to come to an eventual moral verdict, in order to help us arrive at a more nuanced and complete moral theory. This, again, has crucial implications for designing autonomous vehicles. The relevant moral principles will be nuanced in a way that reflects whether the vehicle is facing a single alternative course of action to a collision, or multiple options, in addition to the end-state probability for each outcome. More broadly, such research informs decision-making in complicated real-life cases in which there are multiple actions that might be taken, each matched with a variety of probabilities regarding outcomes.

Chapter 3 is currently being prepared for submission for publication of the material. Ryazanov, Arseny; Wang, Tinghao; Nelkin, Dana; McKenzie, Craig; Rickless, Samuel. The dissertation author was the primary investigator and author of this paper.

References

Barak-Corren, N., Tsay, C., Cushman, F., & Bazerman, M. (2018). If you're going to do wrong, at least do it right: The surprising effect of considering two moral dilemmas at the same time. *Management Science*, *64*(4), 1528-1540.

Berlin, I. (1969). Two Concepts of Liberty, in I. Berlin, *Four Essays on Liberty*, London: Oxford University Press.

Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian Disease problem: Time constraint, need and disease type. *Judgment and Decision Making*, *13*(6), 529.

Fleischhut, N., Meder, B., & Gigerenzer, G. (2017). Moral hindsight. *Experimental psychology*.

Foot, P. (1978). *Virtues and Vices*. Oxford: Blackwell.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, *38*(1), 129-166.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one*, *14*(3), e0213544.

Peeters, G. (1971). The positive-negative asymmetry: On cognitive consistency and positivity bias. *European Journal of Social Psychology*, *1*(4), 455-474.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Oxford University Press.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, *5*(4), 296-320.

Ryazanov, A. A., Knutzen, J., Rickless, S. C., Christenfeld, N. J., & Nelkin, D. K. (2018). Intuitive Probabilities and the Limitation of Moral Imagination. *Cognitive science*, *42*, 38-68.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32-70.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

Shou, Y., & Song, F. (2017). Decisions in moral dilemmas: The influence of subjective beliefs in outcome probabilities. *Judgment & Decision Making*, *12*(5).

Chapter 4: Gambling With Lives and Money: Incidental Affect Shifts Gamble Preferences For Moral and Monetary Decisions

Arseny A. Ryazanov[1], Carsten Erner[2], & Piotr Winkielman[1]

[1] Department of Psychology, University of California, San Diego

Alexandia, VA[2] ,

Abstract

Current evidence for whether incidental affect can shift moral judgment is equivocal. In a large study (N = 365), we examined whether risk preferences in moral decisions are sensitive to incidental affect. Participants made a series of parametrically-varied moral or monetary decisions, adapted from the Asian disease problem, choosing between certain losses and probabilistic losses, as well as between certain gains and probabilistic gains, with varying expected values. Each decision was preceded by positive, negative, or neutral food stimuli, putatively as part of a separate reaction time task. Both moral and monetary gambles were sensitive to incidental negative affect, suggesting that incidental affect can influence moral decision-making through a domain-general process. Additionally, moral decisions were found to overall be less sensitive to expected value than monetary decisions are.

Keywords: moral cognition; incidental affect; prospect theory; morality

Gambling With Lives and Money: Incidental Affect Shifts Gamble Preferences For Moral and Monetary Decisions

Whether morality has an affective or cognitive basis is a fundamental question faced by philosophers and psychologists alike. Hume famously argued that reason was the “slave of the passions”, while Kant believed moral authority to stem from pure reason (Hume, 2003; Wilson & Denis, 2018). The role affect plays in morality has likewise been contested by different schools of psychologists, from Kohlberg’s (1971) rationalist model that rejected the role of affect to Haidt’s (2001) social-intuitionist model which holds affect to be central to moral judgment, likening morality to an emotional dog with a rational tail. Contemporary dual-process models suggest that there are distinct affective and deliberative paths in moral cognition, without taking a stance on whether one of these paths accounts for the majority of moral judgments (Crockett, 2013; Cushman, 2013; Greene 2007; Greene 2009).

Theories that incorporate a causal role of affect in moral judgment, such as the social-intuitionist model, initially enjoyed the support of a large volume of research on incidental disgust, in which moral judgments appeared to be amplified when made in disgusting contexts (e.g., while seated at a messy desk; Wheatley & Haidt, 2005; Schnall, Haidt, Clore, & Jordan, 2008; Horberg, Oveis, Keltner, & Cohen, 2009; Seidel & Prinz, 2013). However, a subsequent meta-analysis of 50 studies that manipulated incidental disgust found a very-small-to-negligible overall effect of incidental disgust on moral decision-making (Landy & Goodwin, 2015; though see Schnall, Haidt, Clore, & Jordan, 2015 for response). While other studies have provided evidence for integral affect playing an important role in moral judgment (e.g., Shenhav & Greene, 2014; Greene, 2017), the question of whether incidental affect matters for moral cognition has become unclear in light of this meta-analysis.

One possible reason for the inconsistency in the observed impact of incidental affect on moral judgment is that incidental affect may only play a role in judgments that people are unsure about, i.e. for boundary cases, rather than amplifying moral judgment more generally. This would be consistent with Payne, Brown-Iannuzzi, and Loersch's (2016) finding that primes affected gambling decisions in a blackjack task only when the value of the gamble was uncertain, rather than across all hands dealt. Adopting such a model, incidental affect could influence judgment of boundary transgressions, for example the permissibility of stealing a box of pens from one's employer for personal use, but not transgressions that are clearly wrong, such as the permissibility of murdering one's coworker for having stolen a box of pens. Thus, the study of how incidental affect shifts moral judgment may benefit from parametrically varying the ambiguity of the morality of the act being judged.

One moral judgment that is particularly amenable to such parametric variation is termed the Asian disease problem. In the Asian disease problem, participants decide between the certain loss of a group of individuals or the probabilistic loss of a larger group, and between the certain survival of a group of individuals and the probabilistic survival of a larger group of individuals (Tversky & Kahneman, 1981). Despite being initially developed to explore the framing effects of losses and gains, with risk aversion observed for gains, but risk seeking observed for losses, subsequent research has identified that scenario details can shift risk preferences. For example, Ginges and Atran (2011) examined a modified Asian disease problem in which they pitted deontological preferences for military action, as opposed to non-violence, against risk preferences in decisions regarding rescuing hostages from a foreign country. Their findings showed that, in decisions where the expected value of a certain option and risky option were tied, deontological preferences for military action could reverse general risk preferences. In another set of studies, the specific nature of the disease involved in the problem (whether the disease was AIDS, leukemia, or an unusual infection) was also found to influence participant risk preferences (Diederich, Wyszynski, & Ritov, 2018). Though exactly why risk preferences differ for different diseases remains unclear, one possibility is that the diseases evoke different affective reactions, which in turn influence risk preferences. If integral affect, such as the kind of disease or type of action undertaken can influence risk preference, it is possible that incidental affect, or affective reactions to unrelated stimuli, could likewise shift risk preferences.

Importantly, like Payne, Brown-Iannuzzi, & Loersch, (2016), Diederich, Wyszynski, and Ritov (2018) departed from the traditional paradigm of asking a few questions to many participants, and instead parametrically varied a series of factors, such as the probability of harm, that could matter in the Asian disease problem. Since prior research has identified that participants are sensitive to the expected value of such decisions (e.g., Shenhav & Greene, 2014), we chose to parametrically vary the expected value of risky options presented alongside certain options to examine the role of incidental affect in moral decision-making across a range of expected values. This allows us a range of moral ambiguity over which to examine incidental affect, and to see whether incidental affect selectively influences moral decision-making involving expected values that participants are uncertain about, consistent with the Payne et al. (2016) task.

While much of the prior research on how incidental affect could amplify moral judgment focused specifically on incidental disgust, recent evidence suggests that disgust is not a unique affective predictor of moral evaluation. For example, Cusimano, Royzman, Leeman, and Metas (2018) found that the role of disgust can be overstated by methodologies that limit participant responses to disgust measures, and Landy and Piazza (2019) found that general negative affect correlates with moral evaluation. Furthermore, it may be that affect, regardless of whether it has a positive or negative valence, shifts moral judgment: Cheng, Ottati, and Price (2013) found that in addition to grief, fear, and disgust, excitement could also shift moral judgment compared to a neutral condition, and that all of these affective states shifted moral judgment indistinguishably from one another. Thus, we examine the role of incidental positive affect, in addition to incidental negative affect, in shaping moral risk preferences.

Our design also affords the opportunity to design an analogous monetary decision task, to directly compare the effects of incidental affect on moral and monetary decision-making. In doing so, we are able to examine whether any observed effect of incidental affect is general, applying to both monetary and moral decisions, or domain-specific. Some evidence suggests that moral and monetary decisions should be made similarly. Moral decisions appear to recruit a general valuation mechanism responsible for monetary decisions in fMRI studies (Shenhav & Greene, 2010). Furthermore, distributive justice preferences regarding how to allocate resources among orphans correlate with gamble preferences (Palmer et al., 2013). Similarly, Landy and Piazza (2019) find that trait disgust sensitivity predicts not only moral judgments, but unrelated judgments including aesthetic and competence judgments as well, and thus could potentially extend to monetary decisions.

However, other evidence suggests that monetary judgments may be distinct from nonmonetary judgments. Hsee and Rottenstreich (2001) found that affect-rich prizes, such as kisses and shocks, do not have the same s-shaped value function that monetary rewards and punishments do. Decisions involving lives may inherently be more affective than decisions involving money, and as such could be more scope insensitive (Hsee & Rottenstreich, 2004), or less sensitive to the specific number of people an action saves. Ginges and Atran (2011) found support for this notion in that participants appeared to be insensitive to expected value in war-time scenarios, in which participants had to specify the minimum number of hostages that would need to be savable to justify military intervention in an unspecified foreign country. Consistent with this, McGraw, Shafir, & Todorov (2010) found that gambles with non-monetary consumer objects were less sensitive to probability range than monetary gambles were. Together, these studies suggest that gambles involving lives could be less sensitive to expected value than gambles involving monetary outcomes.

Incidental Affect Outside Morality

Through comparing the role of incidental affect on analogous moral and monetary gambles, we are able examine whether moral and monetary decisions are differentially sensitive to incidental affect. There are several competing models of how affect influences general decision-making (e.g., Damasio, 2005; Finucane, Alhakami, Slovic, & Johnson, 2000; Forgas, 1995; Lerner & Keltner, 2000; Loewenstein, 1996; Loewenstein, Weber, Hsee, & Welch, 2001; Mellers, Schwartz, Ho, & Ritov, 1997; Schwarz & Clore, 1988). One possibility is that affect interferes with decision-making, and makes decisions more erratic, or less sensitive to expected value (Easterbrook, 1959; Evans, 2003; Pham, 1996). A second possibility is that affect has a general main effect on risk preferences, shifting them across the entire gamble distribution (Damasio, 1994; Schwarz & Clore, 1988; Slovic, Finucane, Peters & MacGregor, 2007). A third is that affect only influenced decision-making only when the subjective value of the choices is close-to-tied (Lowenstein & Lerner, 2003). Recently, negative incidental affect was found to play a role in monetary decisions (Hoffree et al., under review), consistent with affect having a general main effect on risk preferences for money, but with a more pronounced effect for gambles where subjective value is closer to equal.

Thus, our study examines 1) whether moral decisions involving choices between certain and probabilistic outcomes are sensitive to incidental affect; 2) whether any sensitivity to incidental affect observed in moral decision-making is unique to moral judgment, or reflects a general interplay of affect and cognition that can be observed outside the moral domain; 3) whether any observed sensitivity to incidental affect more strongly influences decisions between subjectively-tied decisions outcomes; 4) whether overall sensitivity to expected value varies between moral and monetary decisions.

Study

Methods

Positive, neutral, or negative affective images were presented incidentally before either moral or monetary gambles (the overall logic of the task and stimuli was adapted from Studies 4 and 5 in Hoffree et al., under review). The images contained foods that had been previously pretested to be positive (e.g., a chocolate dipped strawberry), neutral (e.g., peas), or negative (e.g., monkey meat). Each trial was comprised of two sections, a picture segment and a gamble segment, that participants were explicitly told were unrelated. Pictures were randomly assigned without replacement to gambles. Participants were assigned to either the moral or the monetary condition, and responded to the respective gamble following the presentation of the images.

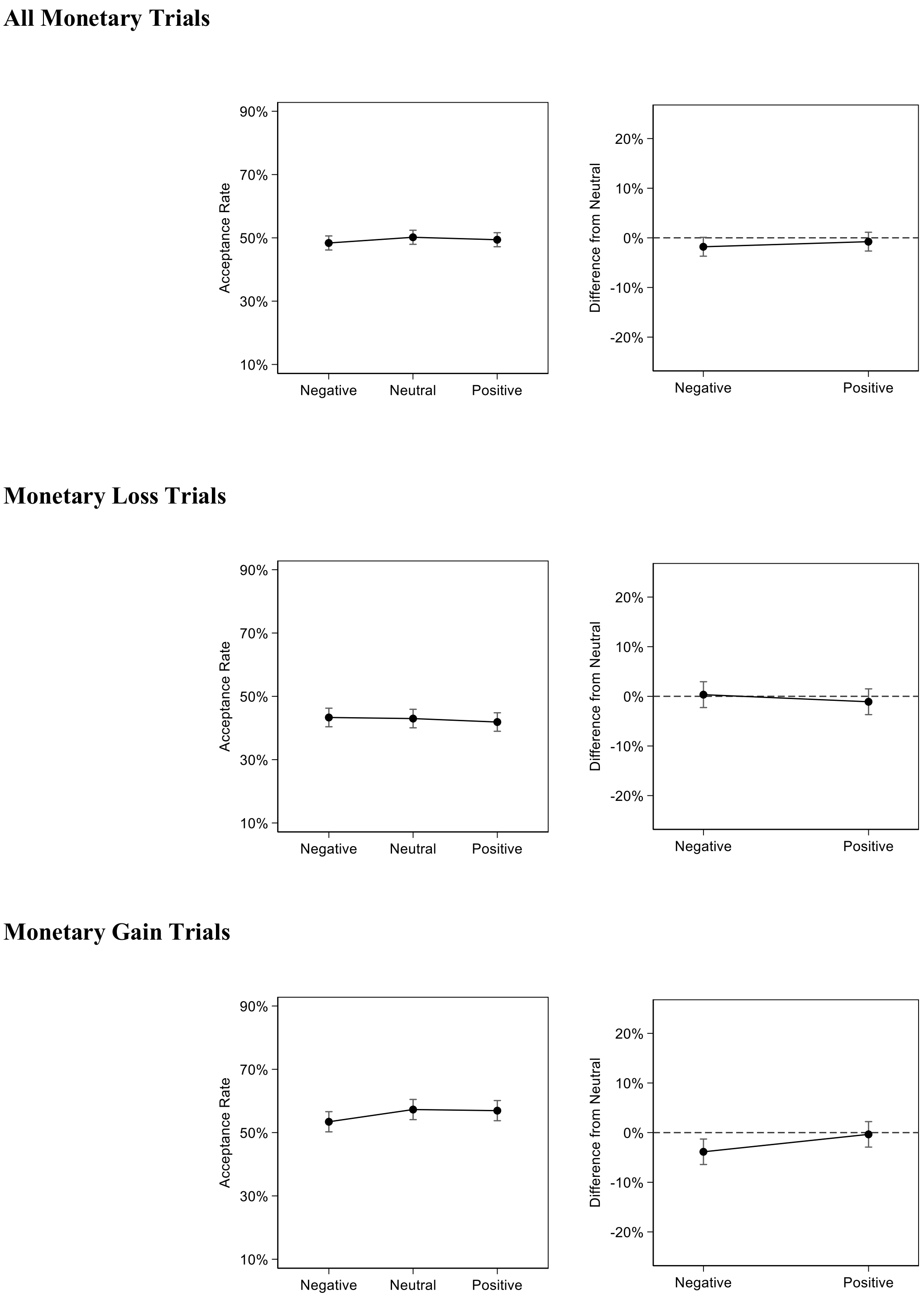
Gambles consisted of a loss block, and gain block, presented in a randomized order. There were 32 trials: 16 loss trails, and 16 gain trails. Each trial was presented with a unique, randomly selected, neutral, negative, or positive stimulus (randomly assigned to avoid demand characteristics of each gamble always being presented with one of each stimulus), for a total of 96 trails. In the case of moral gambles, participants were instructed that, in the context of the outbreak of a rare disease, they needed to pick between two outcomes: 10 people definitely survive, or a 50% chance of 10-40 people surviving (in increments of 2). In the loss condition, participants picked between the certain deaths of 10 people, or a 50% chance of 10-40 people dying (in increments of 2). The monetary gambles were analogous to the moral gambles: gain trials consisted of choosing between certainly gaining 10 points and a 50% chance of gaining 10-40 points, (in increments of 2); loss trials consisted of choosing between chance certainly losing 10 points or a 50% chance losing 10-40 points, in increments of 2. Participants responded to a 1-4 scale of degree to which they accepted the gamble option (strongly accept gamble, weakly accept gamble, weakly reject gamble, strongly reject gamble).

Each trial had the following format: a shape was presented for up to 2 seconds that corresponded to the valence of the stimuli (e.g. circle corresponds to a positive stimulus). Next, the stimulus photograph was presented for up to 2 seconds, with 1.5 seconds of feedback. In this portion of the task participants had been instructed to press a specific key as soon as they saw a shape, and a different key as soon as they saw a photograph. Next a gamble was presented for 4 seconds, after which participants had up to 5 seconds to respond to the gamble using the 4-point scale, and 3 seconds of feedback on their decision.

Participants were told that their responses would help inform public health policy for moral gambles, and economic policy for monetary gambles. Responses were collapsed into accepting/rejecting the gamble for analyses. Three hundred sixty five (75% female) University of California, San Diego undergraduates participated in the study for course credit. Individual trials with missing gamble responses were dropped – 2.1% of trials (mean) were dropped.

Acceptance Rate

We first examine changes in overall rates of acceptance in the affective conditions for moral and monetary decisions. As shown in Figure 1, across the loss and gain procedures, contrasting the negative stimuli against neutral stimuli, a two-level linear model with subject-level random intercepts revealed that participants were less likely to accept gambles following negative stimuli for moral gambles (c = -0.034, $p < .01$, 95% CI [-.057, -.011]), and marginally less likely to accept monetary gambles (c = -0.018, $p = .063$, 95% CI [-.037, .001]), see Figures 1-2. Gamble acceptance did not change following positive stimuli, compared to neutral stimuli (moral gambles: c = -0.013, $p = .289$, 95% CI [-0.036, 0.011]; monetary gambles: c = -0.008, $p = .800$, 95% CI [-0.027, 0.11]). Note that these contrast estimates can be directly interpreted as percentage point differences (e.g., -0.12 equals an 12%-point decrease).

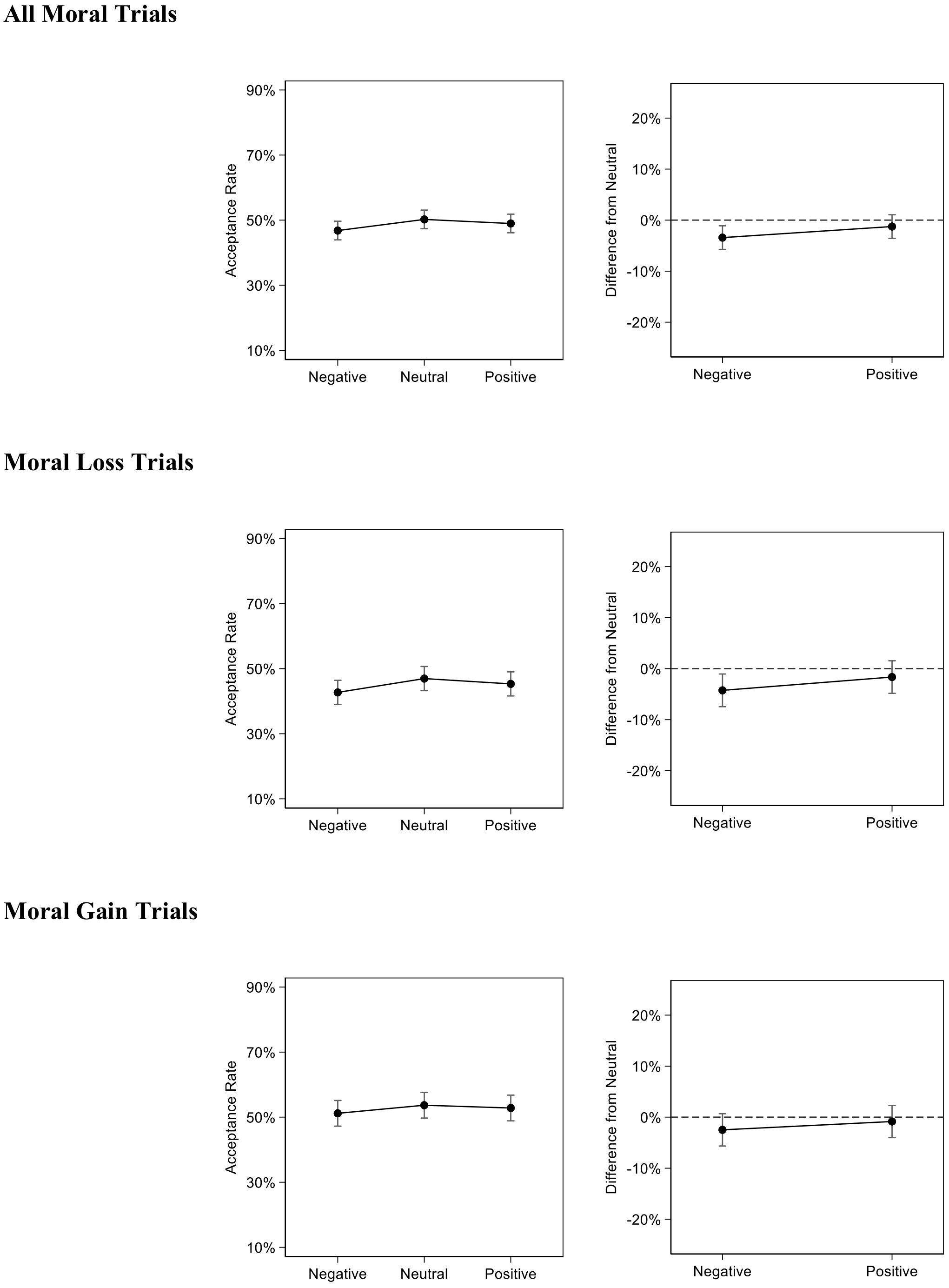


All Monetary Trials

Monetary Loss Trials

Monetary Gain Trials

Figure 4.1. Acceptance rate across affective categories for all trials (collapsed across procedure), loss trials only, and gain trials only for monetary gambles. This figure shows acceptance rates across affective categories (left) and contrasts in acceptance rate relative to the neutral stimuli (right). Reported are adjusted predictions and 95% confidence intervals from a two-level linear model with subject-level random intercepts and gender as covariate.



All Moral Trials

Moral Loss Trials

Moral Gain Trials

Figure 4.2. Acceptance rate across affective categories for all trials (collapsed across procedure), loss trials only, and gain trials only for moral gambles. This figure shows acceptance rates across affective categories (left) and contrasts in acceptance rate relative to the neutral stimuli (right). Reported are adjusted predictions and 95% confidence intervals from a two-level linear model with subject-level random intercepts and gender as covariate.

Next we examined whether sensitivity to incidental affect was limited to either losses or gains for moral gambles and for monetary gambles. We observed that the effect of incidental negative affect was limited to gains for monetary gambles, (monetary gain gambles: c = -.039, *p* = .003, 95% CI [-0.064, -0.013]; monetary loss gambles: c = .003, *p* = .800, 95% CI [-0.023, 0.029]). Unlike for monetary gambles, we observed an effect of incidental negative affect on gambles involving losses, but not gains, (moral gain gambles: c = -.025, *p* = .12, 95% CI [-0.056, 0.007]; monetary loss gambles: c = -.0425, *p* = .009, 95% CI [-0.074, -0.105]).

We then examined whether the effects of negative stimuli were amplified for decisions were more morally ambiguous, see bottom panels of Figure 3 for visualization of the confidence bands for the effects of stimuli valence. In order to examine moral ambiguity, we looked at distance of a particular gamble from a tied expected value with the certain option. If negative stimuli have a stronger effect at more closely-tied expected values, we would expect to see an interaction of absolute distance from tied expected value and affect. Model comparisons for the significance of this interaction were conducted separately for the monetary and moral tasks, with gender entered as a covariate. Absolute value of the distance from tied expected value and stimulus valence were entered as main effects, and compared with a model containing their interaction (note: these analyses assume a linear interaction effect). For both monetary gains and moral losses, the interaction of distance from tied expected value and the effect of negative stimuli did not yield a significantly improved model, suggesting that the effect of incidental affect was not amplified by moral ambiguity: $\chi^2(2) = 3.32$, *p* = .19 for moral losses, $\chi^2(2) = 2.10$, *p* = .35 for moral gains. For monetary gains, there was likewise no interaction of distance and sensitivity to incidental affect: $\chi^2(2) = 1.13$, $p = .57$ for monetary gains, $\chi^2(2) = 4.04$, $p = .13$ for monetary losses. Thus we did not observe increased sensitivity to incidental affect at subjectively tied expected values.

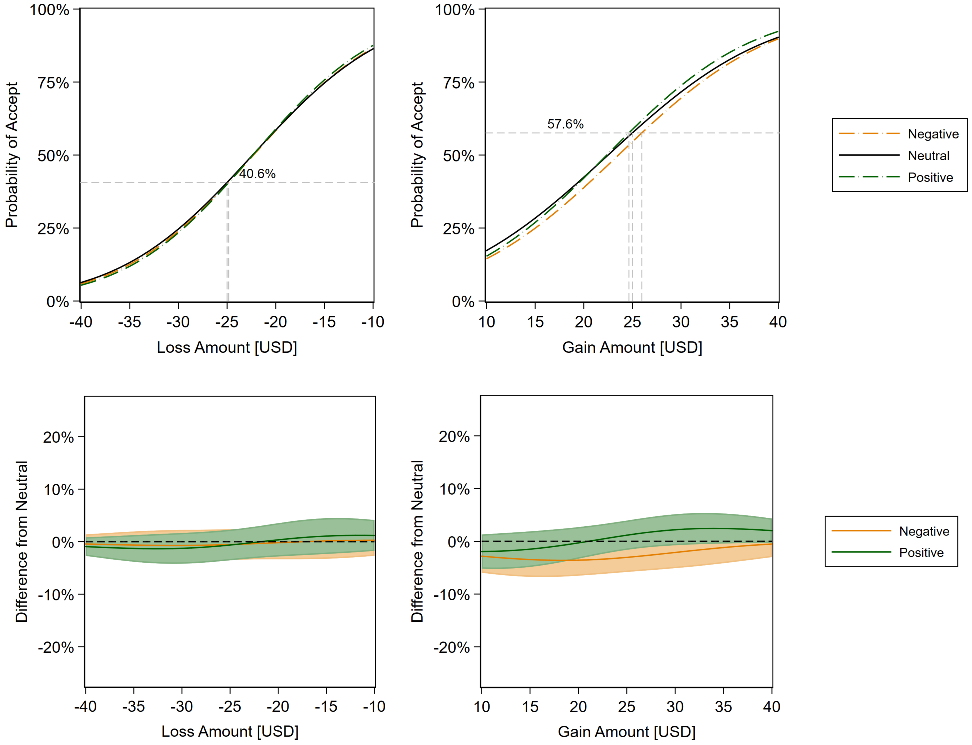


Figure 4.3. Probability of accepting a gamble by expected value, across affective categories for monetary gambles. Results for pure loss (top left) and pure gain (top right). The top panels show probabilities of accepting a gamble across affective categories. The bottom panels show difference from neutral category (shaded areas are confidence bands). Reported are adjusted predictions from a two-level logit model with subject-level random intercepts and gender as covariate.

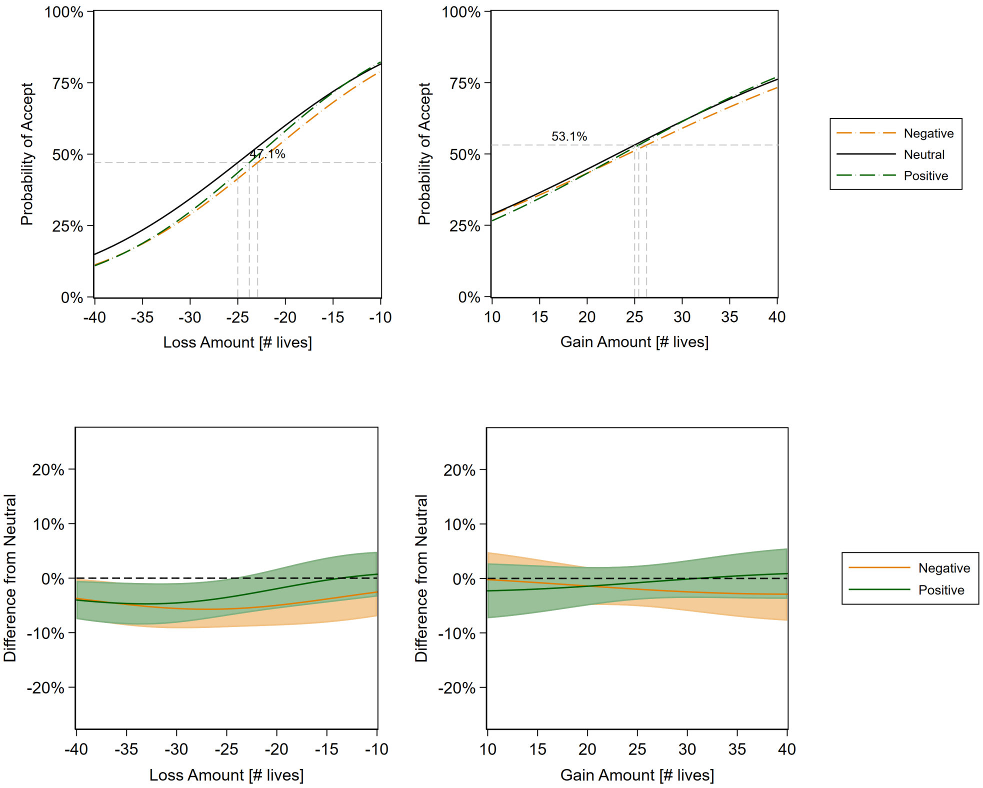


Figure 4.4. Probability of accepting a gamble by expected value, across affective categories for moral gambles. Results for pure loss (top left) and pure gain (top right). The top panels show probabilities of accepting a gamble across affective categories. The bottom panels show the difference from neutral category (shaded areas are confidence bands). Reported are adjusted predictions from a two-level logit model with subject-level random intercepts and gender as covariate. The x-axis shows the expected value of the gamble option (in lives lost/saved).

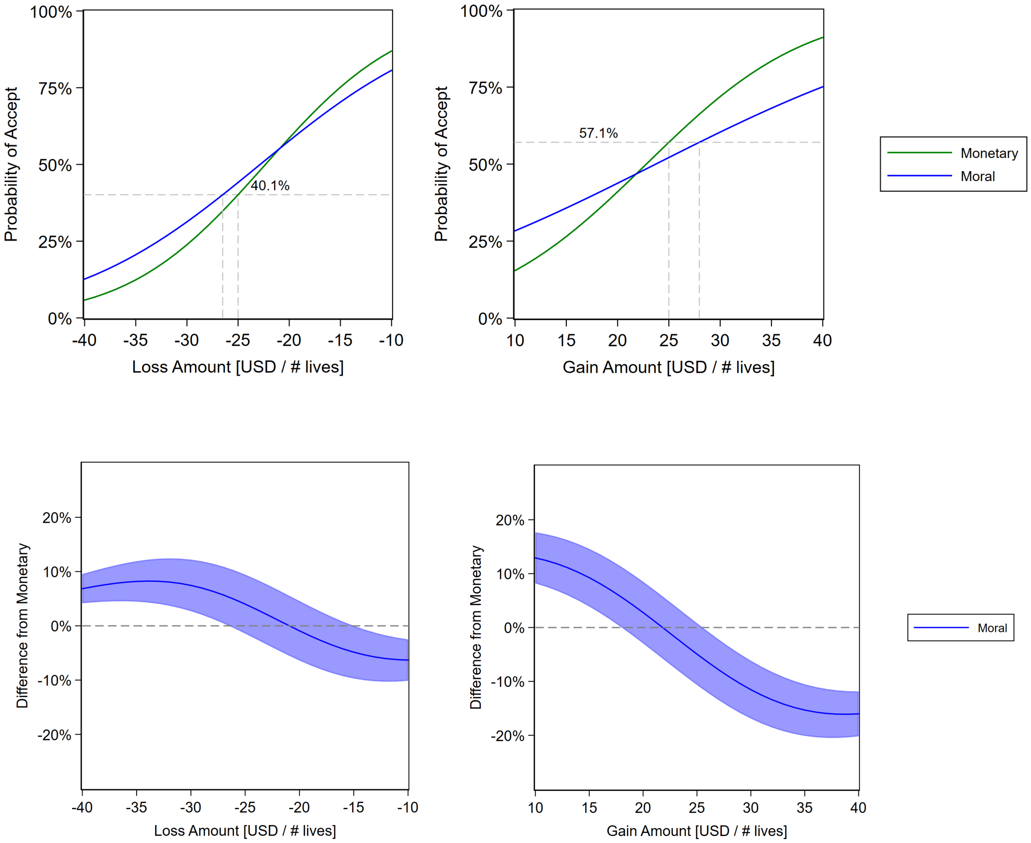


Figure 4.5. Probability of accepting a gamble by expected value, contrasting monetary and moral domain for losses (left) and gains (right). The top panels show probabilities of accepting a gamble across domains. The bottom panels show a difference between domains (shaded areas are confidence bands). Reported are adjusted predictions from a two-level logit model with subject-level random intercepts and gender as covariate.

Lastly, we examined whether participants were more sensitive to expected value for monetary gambles than moral gambles, consistent with scope insensitivity for affective decisions, see Figure 5 for visualization of the difference between domains across the range of expected values. Model comparison using a model that specified the effect of decision type (moral or monetary) and the interaction of expected value and gamble type (gain or loss) better fit the data after the inclusion of an interaction of decision type (moral or monetary) and expected value, $\chi^2(1) = 15.5$, $p < .001$, β for the interaction of decision type and expected value = .25, SE = .06 (note: model comparison assumes a linear interaction effect).

Discussion

We found that moral decisions are sensitive to incidental affect: Exposure to negative stimuli prior to deciding whether to certainly save or lose 10 individuals, or probabilistically save or lose a larger number of individuals, resulted in a decreased likelihood of taking the risky option across moral gambles involving losses and gains. By moving beyond single decisions, and exploring the effects of negative stimuli on a series of decisions with parametrically varied expected values, we see that affect can shift decision-making for decisions involving lives across a range of expected values. Our results point to a small but consistent effect of incidental affect on moral decision-making.

By examining sensitivity to incidental affect in analogous monetary and moral tasks, we were also able to explore whether a general process can account for sensitivity to incidental affect in moral decision-making. We found that incidental affect shifted both monetary and moral decisions, suggesting a domain-general process, though in somewhat different ways. In our monetary decision, incidental affect had stronger effects on decisions involving monetary gains than on decisions involving monetary losses. However, for the moral decisions, we observed a stronger effect for moral losses than moral gains. Why this is the case is unclear. One speculative possibility is that incidental affect amplifies integral affect, and that monetary gains and moral losses carry more integral affect than monetary losses and moral gains in a task such as ours, where decisions are not tied to actual earnings or lives. That is, the gain of money may be more emotionally salient than the loss of money in a task without real payoffs, and the loss of lives may likewise be more salient than the gain of lives. Future studies should explore whether the difference in sensitivity to incidental affect between moral and monetary gambles reflects a difference between monetary and moral cognitive processes despite their overall similarity, and also further examine the relationship between incidental and integral affect (see Västfjäll et al., 2016 for overview on integration of incidental and integral affect).

Overall, participants were more sensitive to expected value in monetary decisions than in moral decisions, showing less agreement for gambles that had extremely high or low expected values in moral decisions. This is consistent with prior research showing that affective decisions are less sensitive to scope. It is possible that by making the moral decision more affective, such as providing pictures of the people to be killed or saved, participants could become even more insensitive to scope.

Our study has several limitations: namely that the task required many trials, and could potentially cause participants to attune to features that matter less under single evaluation. However, differential sensitivity for losses and gains suggests that responses to the negative stimuli were not the result of demand characteristics, as, had they been, we would not expect to see differential sensitivity to losses and gains, particularly in the monetary condition, where we observer a stark difference between the effect of negative valence on losses and gains.

Incidental affect can influence moral cognition: inducing negative affect in our study shifted moral risk preferences. Though we did not find reason to be the slave of passions, that moral decisions are influenced by whether they are made in the context of disgusting food suggests that moral decisions are not immune to irrelevant passions.

Chapter 4 is currently being prepared for submission for publication of the material. Ryazanov, Arseny; Erner, Carsten; Winkielman, Piotr. The dissertation/thesis author was the primary investigator and author of this material.

References

Cheng, J. S., Ottati, V. C., & Price, E. D. (2013). The arousal model of moral condemnation. *Journal of Experimental Social Psychology*, *49*(6), 1012-1018.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363-366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273-292.

Cusimano, C., Royzman, E. B., Leeman, R. F., & Metas, S. (2018). Methodology is the core disgust problem: Response to Inbar and Scott (2018). *Judgment and Decision Making*, *13*(6), 639.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Penguin.

Damasio, A. (2005). The neurobiological grounding of human values. In *Neurobiology of human values* (pp. 47-56). Springer, Berlin, Heidelberg.

Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian Disease problem: Time constraint, need and disease type. *Judgment and Decision Making*, *13*(6), 529.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, *117*(1), 39–66.

Ginges, J., & Atran, S. (2011). War as a moral imperative (not just practical politics by other means). *Proceedings of the Royal Society B: Biological Sciences*, *278*(1720), 2930-2938.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in cognitive sciences*, *11*(8), 322-323.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. *The cognitive neurosciences*, *4*, 1-48.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66-77.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of personality and social psychology*, *97*(6), 963.

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: on the affective psychology of value. *Journal of Experimental Psychology: General*, *133*(1), 23.

Hume, David. *A treatise of human nature*. Courier Corporation, 2003.

Kohlberg, L. (1971). Stages of moral development. *Moral education*, *1*(51), 23-92.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*(4), 518-536.

Landy, J. F., & Piazza, J. (2019). Reevaluating moral disgust: sensitivity to many affective states predicts extremity in many evaluative judgments. *Social Psychological and Personality Science*, *10*(2), 211-219.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion*, *14*(4), 473–493.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272–292.

Lowenstein, G., & Lerner, J.S. (2003). The role of affect in decision making. In R. Davidson, K. Scherer, & H. Goldsmith (Eds.), *Handbook of affective science, pp. 619-642*. New York: Oxford University Press.

Loewenstein, G., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267–286.

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*(5), 816-830.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*(6), 423–429.

Palmer, C. J., Paton, B., Ngo, T. T., Thomson, R. H., Hohwy, J., & Miller, S. M. (2013). Individual differences in moral behaviour: a role for response to risk and uncertainty?. *Neuroethics*, *6*(1), 97-103.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General*, *145*(10), 1269.

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological science*, *12*(3), 185-190.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and social psychology bulletin*, *34*(8), 1096-1109.

Schwarz, N., & Clore, G. L. (1988). How do I feel about it? The informative function of affective states. In K. Fiedler & J. P. Forgas (Eds.), *Affect, Cognition, and Social Behavior* (pp. 44–62). Hogrefe.

Slovic, P., Finucane, M.L., Peters, E., & MacGregor, D.G. (2007). The affect heuristic. *European Journal of Operational Research, 177*(3), 1333-1352.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2015). Landy and Goodwin (2015) Confirmed Most of Our Findings Then Drew the Wrong Conclusions. *Perspectives on Psychological Science*, *10*(4), 537-538.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, *34*(13), 4741-4749.

Seidel, A., & Prinz, J. (2013). Sound morality: Irritating and icky noises amplify judgments in divergent moral domains. *Cognition*, *127*(1), 1-5.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

Västfjäll, D., Slovic, P., Burns, W. J., Erlandsson, A., Koppel, L., Asutay, E., & Tinghög, G. (2016). The arithmetic of emotion: Integration of incidental and integral affect in judgments and decisions. *Frontiers in Psychology*, *7*, 325.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780-784.

Wilson, E. E., & Denis, L., (2018). Kant and Hume on Morality. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL: https://plato.stanford.edu/archives/sum2018/entries/kant-hume-morality

Chapter 5: How Enough Helping Becomes Enough: The Moral Accounting Model

Arseny A. Ryazanov, Dana Kay Nelkin, Samuel C. Rickless, Nicholas J. S. Christenfeld

University of California San Diego

Under review at *Journal of Experimental Social Psychology*

Correspondence concerning this article should be addressed to Arseny Ryazanov, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109

Contact: aryazano@ucsd.edu

Abstract

When does it become acceptable to decline to help a cause? We propose a model of moral accounting, which outlines lay theory of how an individual's prior beneficence increases the acceptability of them declining further helping. We identify two pathways through which moral demands decrease: having helped can make further helping more costly and can also bank moral credit (Studies 1 and 2). We then uncover the cognitive structure of moral credit: Moral credit is sensitive to effort, luck, domain, and time, and is taken into consideration in addition to the increasing marginal cost of continued helping, in ways consistent with mental financial accounting (Studies 3-5). As a result, beneficence completed over longer time periods grants extended credit (Study 6). Study 7 compares how judgments made in a within-subjects design, which we interpret as normative judgments of how much these factors should matter, generalize to judgments of individual actions.

Keywords: morality; helping, altruism; decision-making; prosocial

How Enough Helping Becomes Enough: The Moral Accounting Model

From hungry children to oil-soaked seabirds, countless needs call for benevolent action. What prompts people to beneficence and where people's moral responsibilities lie is the subject of much inquiry in both philosophy and psychology. A related question is when it becomes acceptable to quit. Is there some point where one has done enough?

There is intense philosophical debate about the demandingness of duties of beneficence (Beauchamp, 2016). Some argue that such duties are extremely demanding, requiring sacrifice even for complete strangers (Singer, 1972; Unger, 1996; Arneson, 2004). Such views follow from act-consequentialism, which requires people to do what has the most beneficial consequences, impartially considered. When needs are great and individuals have resources to address those needs, the demands of beneficence will be extremely high. Singer (1972) suggests that one ought to give up to the point at which further beneficence would make one worse off than it would help the beneficiary. The problem of over-demandingness has been taken by some as an objection to act-consequentialism (Scheffler, 1994; Vallentyne, 2006). In response, act-consequentialists have found various ways of accommodating the problem (Kagan, 1989; Singer, 2011; Arneson, 2004), including distinguishing between ground-level morality and the less-demanding rules that should be publicly advocated for practical reasons.

Other philosophers advocate more modest principles of beneficence (Nagel, 1989; Murphy, 1993; Scheffler,1994; Herman, 2001; Miller, 2004; Cullity, 2006). One challenge for such views is to explain what justifies departures from full impartiality. Another is to provide some principled rationale for where to draw the line between what is required in the way of beneficence and what is not. Various considerations are invoked to explain and justify limits, such as collective responsibility, that it would be wrong for individuals to be required to give more than their fair share (Murphy, 1993), and that morality must accommodate our natural psychological proclivity to invest in special relationships (Nagel, 1989; Scheffler, 1994; Herman, 2001; Miller, 2004). Complications aside, these ethicists share the ordinary conviction that there are limits to the duty of beneficence.

But if ordinary moral intuition tends to favor a modest principle of beneficence, what do people think reduces the moral requirement to make further contributions? Beyond what people often explicitly offer as justifications, a large literature shows that situational factors can influence whether a person chooses to engage in helping. For example, people care more about single victims than they care about statistical victims (Hsee & Rottenstreich, 2004; Jenni & Loewenstein, 1997; Small & Loewenstein, 2003; Small, Loewenstein, & Slovic, 2007); are less inclined to help those who are, or feel, far away (e.g., Touré-Tillery, & Fishbach, 2017); and differentially value lives (Goodwin & Landy, 2014). Given the lack of clear normative guidance about the limits of beneficence, we set out to uncover a guiding lay theory of how people actually understand the ethical problem of when people have done enough.

Our model first breaks apart prior beneficence into increased future cost of acting and moral credit. Having acted can make future action more costly: swimming out to save the tenth capsized boater will be more difficult than was swimming out to save the first. This increased marginal cost need not be in physical fatigue, but could be in any finite resource. Cost to the agent is frequently cited in explaining limits on the demands of beneficence (Scheffler, 1994; Cullity, 2006; Vallentyne, 2006; Beauchamp, 2016), with moral demand inversely proportional to cost. But prior action will not always make future action more costly. Beneficence that does not deplete a finite resource may tap a different resource—moral credit.

Often in work related to moral demands, moral credit from a prior contribution is confounded with increased marginal cost of further action (e.g., Bell, Grekul, Lamba, Minas, & Harrell, 1995; Barnes, Ickes, & Kidd, 1979). Many ethicists likewise talk about "costs" and "burdens," without carefully distinguishing between different kinds of costs or addressing the question of whether cost and objective prior contribution come apart (e.g., Singer, 1972; Murphy, 1993; Unger, 1996; Miller, 2004; Cullity, 2006; Scheffler, 1993 and 1994).

Research on moral licensing suggests that people may endorse a moral credit path in moral accounting. Moral licensing occurs when having done a good thing, people feel licensed to misbehave (Sachdeva, Iliev, & Medin, 2009; Merritt, Effron, & Monin, 2010; Blanken, van de Ven, & Zeelenberg, 2015), which could be thought of as a reduction below zero in need for further moral action. For example, participants prompted to recall a time when they had behaved morally reported decreased intention to give to charity, donate blood, and volunteer, relative to a control group (Jordan, Mullen, & Murnighan, 2011). Though research on moral licensing generally concerns what subjects themselves feel and do, Effron and Monin (2010) found it to extend to judgments of others. Though a mental moral credit account has been discussed as a possible explanation for why moral licensing occurs (Merritt, Effron, & Monin, 2010), its structure remains underexplored.

Moral credit could arise from the effort, or from the effect, of prior contributions. From a normative standpoint, one might think that only effort, and not ultimate success or failure, should count toward moral credit: One ought to be assessed only for what is in one's control; the rest is, after all, just luck. There is a great deal of debate about whether moral judgment should depend on the results of one's actions that are not in one's control. Many philosophers, and even the authors of the Model Penal Code, take it that while we often do assess failed attempts at harm less negatively than successful harmings, this is a mistake (e.g., Nagel, 1979; American Law Institute, 1985). For example, once we hold all else fixed in a case (e.g., two equally skilled assassins aim to shoot their victim, pull the trigger with equal malice, and by a fluke, one's gun jams and the other's does not), we see that each is equally blameworthy. But not all moral and legal theorists agree (e.g., Moore, 1997), and participants do factor in moral luck when attributing blame (Young, Nichols, & Saxe, 2010; Royzman & Kumar, 2004; Cushman, Dreber, Wang, & Costa, 2009; Gino, Shu, & Bazerman, 2010). No parallel debate about moral luck has arisen in discussions concerning obligations of beneficence. At the same time, even simply expressing the intention to help can reduce future helping (Tanner & Carlson, 2008), suggesting that outcomes are not all that matters. Based on such research on moral luck, we predict moral credit to factor in success, in addition to effort.

Subsequently, we examine the generalizability of moral credit. If moral credit is mentally represented the way money is, moral credit for beneficence in one domain may not excuse from helping in another. Money is not treated as entirely fungible—mental financial budgets consist of buckets, such that an expense incurred in one account does not result in another account being depleted (Thaler, 1999). For example, participants told that they had already attended a basketball game that cost $50 indicated being less likely to attend a play, compared to participants told that they had received a $50 parking ticket, because only the former had already depleted their mental entertainment budget (Heath & Soll, 1996). There is some evidence to support the idea that people may treat money and moral credit similarly. Moral decisions recruit the same neural circuitry as financial decisions (Shenhav & Greene, 2010), gamble preferences correlate with distributive justice concerns (Palmer et al., 2013), and lives are susceptible to the same framing effects as money (Tversky & Kahneman, 1986).

The moral licensing literature, however, suggests that there should be some domain generalizability of the credit. The positive intentions that, for example, past recall of immoral acts engenders often belong to a different domain than the historical deed (e.g., Cascio & Plant, 2016; Mazar & Zhong, 2010, Effron & Monin, 2010). Whether moral credit is likely to be greater if that overlap is greater, and reduced if the two acts are more disparate, has not been explored. We hypothesize a domain-sensitive limited fungibility of moral credit, similar to the compartmentalization that occurs for financial mental accounting, but some generalizability, consistent with moral credit.

Our model also includes a temporal component to the moral credit—that the reduction in demand for future action from prior action should dissipate over time. Research on how long financial expenditures endure provides a framework for thinking about the temporal generalizability of beneficence: One example comes from a study of cab drivers, who drive for longer on days they are making less money, suggesting that the intuitive temporal endurance of their earnings is one day (Camerer, 1997). Another example comes from Gourville & Soman's (1998) findings that gym attendance spiked right after payment, enduring just as long after an annual payment as after monthly payments. If a similar process occurs for good deeds done, it is unclear how long that beneficence is thought to endure. Gourville & Soman's (1998) findings also suggest that the duration of the financial debit seems to be less sensitive to size than it is to recency. If moral credit operates similarly, a diffuse prior effort could likewise produce more enduring moral credit.

Research on moral licensing generally involves exploring the immediate effects of prompts or experimental manipulations (see Merritt, Effron, & Monin, 2010). As a result, the temporal gradient of moral licensing, or how long such an effect could persist, remains underexplored. While recalling prosocial behavior has been shown to cause immoral behavior in the present, whether recency matters is unclear (Jordan, Mullen, & Murnighan, 2011). Our hypothesized model of lay theory, which is shown in Figure 1, suggests that prior beneficent acts reduce the moral demands for further action (Study 1). Such an effect can occur through two pathways. First, having acted can make further action more costly. Second, having acted can bank moral credit (Study 2). Moral credit can accrue both from the degree of effort put into the prior action and from the efficacy of that action (Study 3). Moral credit applies most strongly to domains most similar to the initial act (Study 4) and decreases over time (Study 5). Temporally diffuse beneficence produces more enduring credit than a single act (Study 6).

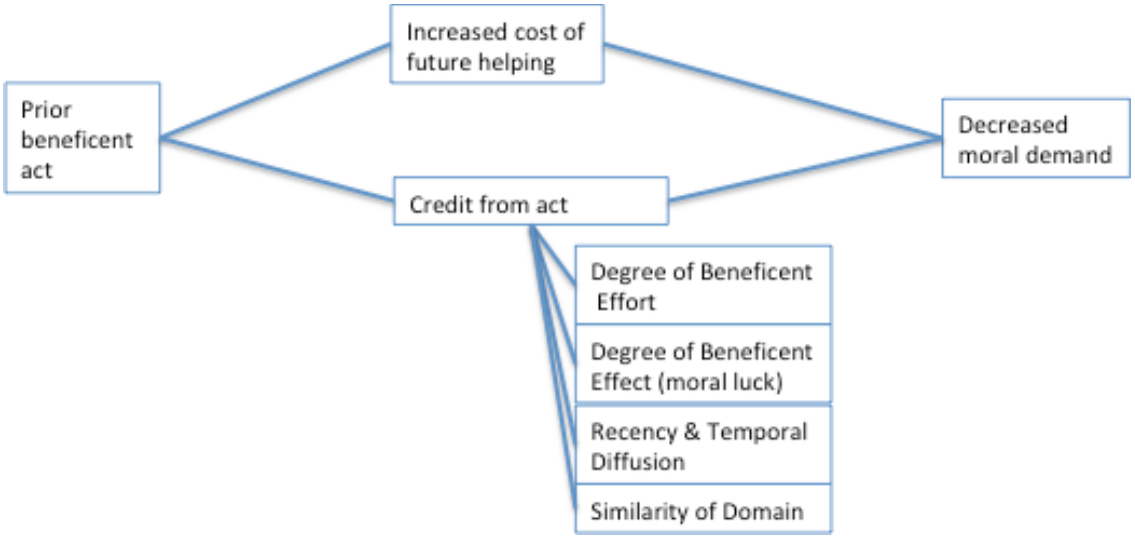

Prior beneficent act
Increased cost of future helping
Credit from act
Decreased moral demand
Degree of Beneficent Effort
Degree of Beneficent Effect (moral luck)
Recency & Temporal Diffusion
Similarity of Domain

Figure 5.1. The Moral Accounting Model

Our model concerns how participants *think* credit should be assigned, and as such, we utilize a within-subjects design. By making the differences salient to participants, we see whether participants think these differences *should* matter. This approach has been used to resolve inconsistencies in moral judgment. For example, the divergence between two versions of trolley problems—whether to drop one person onto a trolley track to save others ahead or divert the trolley onto a track with one person on it to save others ahead—is reduced when the evaluations are made jointly rather than individually, suggesting that participants do not think that the differences are morally relevant (Barak-Corren, Tsay, Cushman, & Bazerman, 2018). Rather than exaggerating differences between scenarios, the within-subjects design prompts deliberation as to whether differences should matter. Furthermore, between-subject evaluation can sometimes obscure preferences that exist when participants are presented with all options (Kahneman & Ritov, 1994). Thus, we uncover a lay theory of moral accounting using a within-subjects design. However, Study 7 compares participant responses when only seeing one level of each scenario used in Studies 1-6 to examine the extent to which individual judgments map on to the lay theory of moral accounting.

Study 1: Prior Beneficent Work

Study 1 Materials

Study 1 examined participants' intuitions regarding whether prior beneficence reduces the moral demand for further beneficence. (*How* these intuitions reflect views about the costliness or creditability of the agent's prior moral acts is examined subsequently.) Participants saw two scenarios, randomly ordered, which varied whether the protagonist had already done beneficent work:

Thousands are homeless and stranded after Hurricane Katrina. Bill is thinking about volunteering by rescuing people in his boat for the next hour. He would very much like to take a nap. Instead, he could spend the hour looking for people. By doing so, he could rescue about 10 people.

Thousands are homeless and stranded after Hurricane Katrina. Tony is a volunteer helper, who is rescuing people in his boat. He has spent 12 hours today finding people and ferrying them to nearby shelters. He has probably rescued about 10 people per hour or around 120 people today. Tony is tired and desperately wants to go home to rest. Before he goes home, he could spend one more hour rescuing 10 people.

After seeing each scenario, subjects were asked, "*How morally acceptable would it be for Bill [Tony] to decline spending the next hour rescuing 10 people?".* Subjects rated the demand on a scale from -3 (not at all acceptable) to 3 (completely acceptable).

We used, here and throughout our studies, a continuous scale of the moral acceptability of declining to help, for example, instead of a dichotomous measure of whether to help, because an actual decision about whether action is morally required would likely require much more background information. The moral acceptability of declining an action can be judged weaker or stronger based on the presented information, in the absence of background knowledge potentially relevant to moral decision-making. We report all measures, manipulations, and exclusions in this and all subsequent studies other than additional demographic measures. Sample size was determined prior to data collection, and was selected to be similar to other studies using within-subject designs to examine moral judgment (e.g., Ryazanov, Knutzen, Rickless, Christenfeld, & Nelkin, 2018). Sensitivity power analyses were conducted for Study 1(using alpha = .05 and beta = 80%) to determine the minimum detectable effect size (MDES; *d*) using Gpower software (Erdfelder, Faul, & Buchner, 1996).

Study 1 Results

Eighty-eight American adults were recruited via Amazon's Mechanical Turk (9 excluded for failing a basic attention check; 64.6% female; mean age = 33.6, SD = 10.4). Subjects rated the moral demands on Bill (no prior beneficence) as greater than the moral demands on Tony (prior beneficence), *t*(78)= 10.6, *p* < .001, *d* = 1.47 (mean prior good = 1.47, SD = 1.60; mean no prior good = -1.04, SD = 1.81; see Figure 2), MDES *d* = .30.

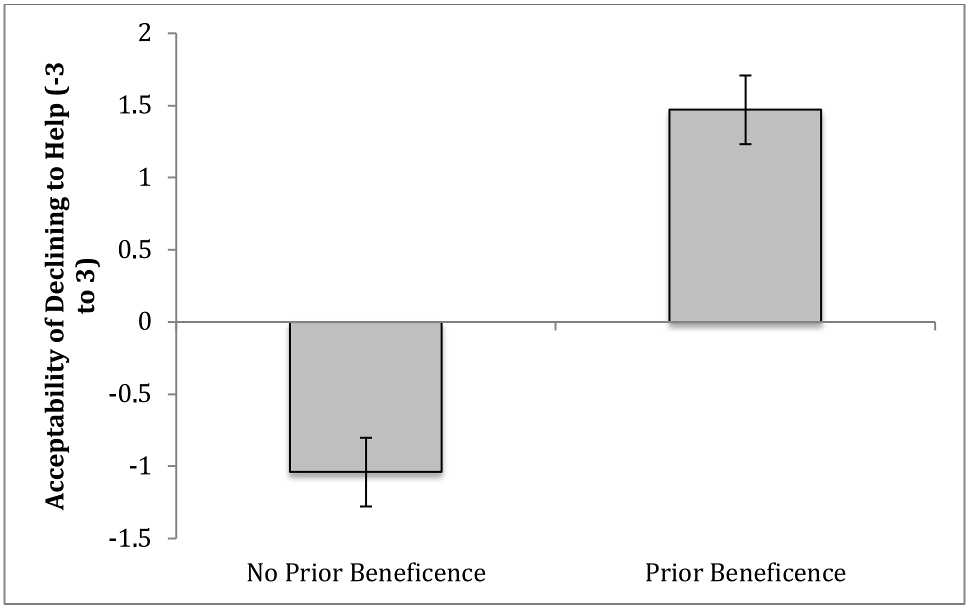


Figure 5.2. The effect of prior beneficence on increasing the acceptability of declining to help. Error bars represent one standard error.

Study 1 Discussion

Our first study indicates that people think moral demands for beneficent action are reduced for those who have already worked to do good. We thus have evidence for accounting, the most basic aspect of our model, and can next examine how the perceived acceptability of declining to help arises. We begin by testing two potential paths of moral accounting: (i) the cost of continuing to help and (ii) the moral credit accrued from prior beneficence.

Study 2: Marginal Cost and Prior Contributions

Study 2 Materials

Three sets of scenarios explored the cost and prior contribution paths: the boat rescue scenario, modified from Study 1, and new scenarios involving volunteering at a soup kitchen and donating a kidney. We started with the same comparison as in the first study – someone who has worked to do good in the area versus someone who has not (e.g., worked a long soup kitchen shift versus spent the same time having a quiet day at home). A reduction in the demand on the former to volunteer an hour in the soup kitchen might reflect some combination of the moral credit for prior work and increased cost of further effort. We aimed to unconfound these two factors by creating scenarios in which the agent has done good in a way that would not make the further effort harder (e.g., the agent donated money rather than worked the soup kitchen shift) and scenarios in which the agent had not done good but further effort would be costly (e.g., the agent had put in a long shift as a chef, but not in a soup kitchen). If the perceived reduction in moral demand from prior work is entirely due to the increase in the cost of future work, then, when the cost is eliminated, there should be no reduction. On the other hand, if the perceived reduction in moral demand is due to the moral credit from having accomplished good, then there should be no reduction when the prior work was not beneficent. Scenarios were presented in a random order, within-subjects. For each scenario, subjects responded to, "*How morally acceptable would it be for NAME to decline doing XX?*" *(-3 to 3; completely unacceptable – completely acceptable)*. Sensitivity power analyses were conducted for Studies 2-6 (with alpha = .05 and beta = 80%) to determine the minimum detectable effect size (MDES; *d*) using Westfall's (2015) PANGEA program.

Study 2 Results

One hundred fourteen undergraduates participated in the study for partial course credit, (22 excluded for failing a basic attention check; 69.6% female, mean age 20.1, SD = 2.23). Subject responses were analyzed using a repeated measures ANOVA with scenario set, prior effort, and moral credit entered as factors. Across the three scenario sets, there was a main effect of cost $F(1, 91) = 232.5$, $p < .001$, $\eta_p^2 = .72$ (mean low cost = -.03, SE = .08, mean high cost = 1.28, SE = .07), MDES $d = .10$, such that it was more acceptable for people who had already worked, regardless of whether the work was in the aid of others, to decline helping, see Figure 3. There was also a main effect of moral credit $F(1, 91) = 91.7$, p < .001, $\eta_p^2 = .50$ (mean low prior contribution = .30, SE = .09, mean high prior contribution = .95, SE = .07) MDES $d = .12$, such that it was more acceptable for people who had already helped, whether that help was effortful or not, to decline helping. There was a main effect of scenario, $F(2, 182) = 68.4$, $p < .001$, $\eta_p^2 =$ .43 (mean boat rescue = -.38, SE = .10, mean kidney = .94, SE = .10, mean soup kitchen = 1.30, SE = .08), and no interaction of prior effort and prior contribution $F(1, 91) = .044$, $p = .83$, $\eta_p^2 < .001$, though there was a significant interaction of scenario and prior effort $F(2, 182) = 11.8$, $p < .001$; $\eta_p^2 = .12$ (main effect of effort for each scenario set significant, p <.001), as well as a significant interaction of scenario and prior contribution $F(2, 182) = 28.0$, $p < .001$, $\eta_p^2 = .24$ (main effect of prior contribution for boat rescue and soup kitchen scenarios individually significant, $F(1, 91) = 128$, $p < .001$, $\eta_p^2 = .58$ and $F(1, 91) = 24.5$, $p < .001$, $\eta_p^2 = .21$, respectively; trend for kidney scenarios, $F(1, 91) =$ 2.78, $p < .001$, $\eta_p^2 = .03$).

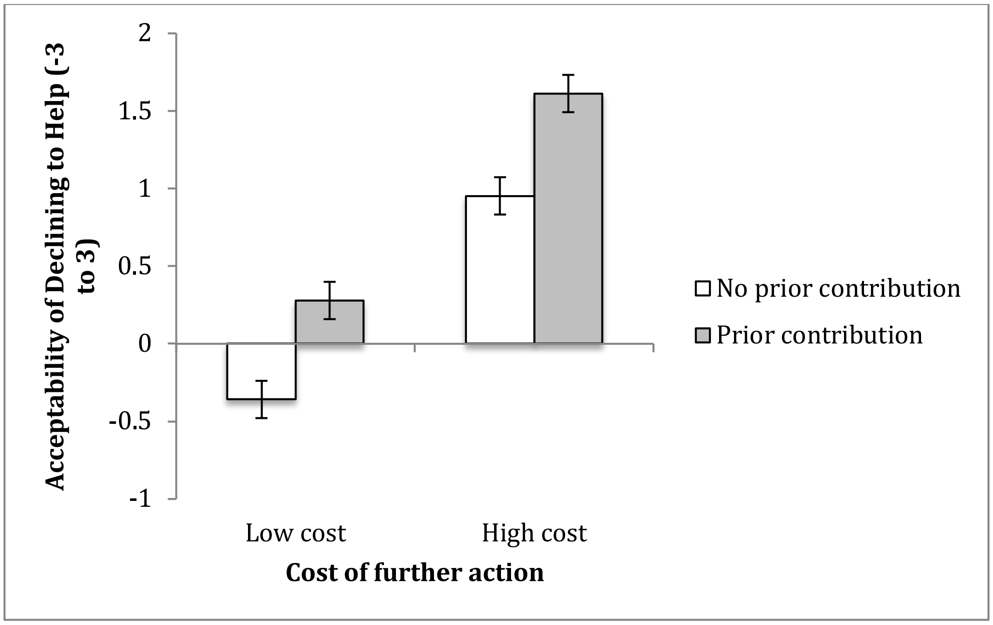


Figure 5.3. The effects of cost of further action and prior contribution on the acceptability of declining to help. Error bars represent one standard error.

Study 2 Discussion

Study 2 provides evidence that people take both increased cost and moral credit as reducing the demand to do more good. That there was no interaction between these factors suggests they are additive, with no special reduction in demand if one is tired from moral labor. Our results therefore suggest that most people are not (consistent) act-consequentialists with a view that moral demand depends on the good left to be done and is independent of prior effort.

Study 3: Effort vs. Effect

We explored the contributions of luck and effort to moral credit, to examine whether moral luck applies to positive outcomes, as it does for negative ones.

Study 3 Materials

We adapted versions of the soup kitchen and boat rescue scenarios and varied effort (e.g., the agent worked for hours to make enough food for 20 people, or used the last few minutes of his shift to prepare enough leftovers for 20 people in the soup kitchen scenarios, and luck (e.g., the food actually fed the 20 homeless, or could not be delivered). The moral act inquired about was a new one, related to the first (e.g., giving money to the homeless shelter) but different enough that the cost of performing it should not depend on the prior effort. Both scenarios controlled for the total effort (e.g., time spent working in a kitchen), as well as the outcomes of further beneficence. The design was entirely within-subjects. See Appendix B for scenarios.

Study 3 Results

One hundred thirteen undergraduates participated in the study for partial course credit (13 participants excluded for failing an attention check; female = 67.0%, mean age = 19.7, SD = 1.76).

Analyzed as a 2 (effort) x 2 (luck) x 2 scenario repeated measures ANOVA, the impact of effort was significant, $F(1, 99) = 41.2$, $p < .001$, $\eta_p^2 = .29$, 95% CI [.15, .42] (mean low effort = -.05, SD = 2.06; mean high effort = .35, SD = 1.96; see Figure 4), MDES $d = .08$. Likewise, the effect of luck was significant, $F(1, 99) = 36.4$, $p < .001$, $\eta_p^2 = .27$, 95% CI [.13, .40] (mean unlucky = -.055, SD = 1.97; mean lucky = .35, SD = 2.05), MDES = .10. There was also a main effect of scenario $F(1, 99) = 142$, $p <.001$, $\eta_p^2 = .59$ (mean boat rescue = -.98, SD = 1.70, mean soup kitchen = 1.28, SD = 1.65). There was a significant interaction of scenario by effort, such that prior effort played a larger role in the boat rescue scenario than in the soup kitchen scenario, $F(1, 99) = 6.00$, $p =$

.016, $\eta_p^2 = .06$ (mean boat low effort = -1.27, SD = 1.67; mean boat high effort = -.70, SD = 1.67; mean soup low effort = 1.17, SD = 1.65; mean soup high effort = 1.40, SD = 1.64; main effect of effort for each scenario set significant—boat: $F(1, 99) = 15.6$, $p < .001$, $\eta_p^2 = .25$; soup: $F(1, 99) = 7.06$, $p = .009$, $\eta_p^2 = .07$). No other interaction was significant (interaction of scenario and efficacy: $F(1, 99) = 1.28$ $p = .26$, $\eta_p^2 = .01$; interaction of effort and efficacy $F(1, 99) = .02$ $p = .89$, $\eta_p^2 < .01$; interaction of effort and efficacy $F(1, 99) = .09$ $p = .76$, $\eta_p^2 < .01$).

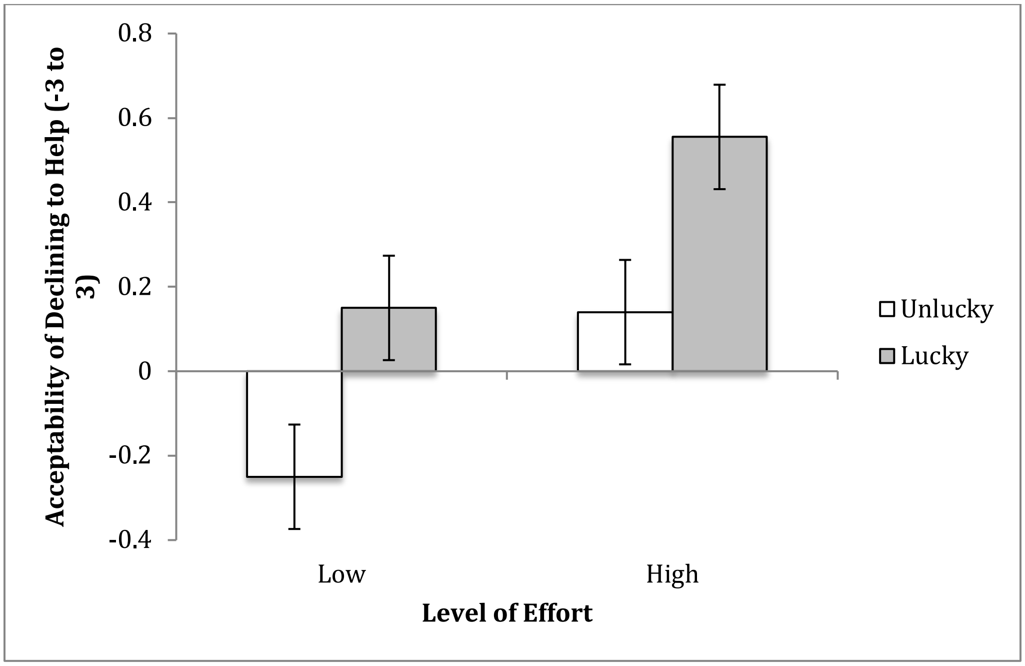


Figure 5.4. The effect of effort and luck on the acceptability of declining to help. Error bars represent one standard error.

Study 3 Discussion

Study 3 indicates that people take into account both how hard the agent tried, and her luck, consistent with our model. Results of past efforts that are not in the agent's control clearly matter to participants' judgments, so it appears that luck is taken into account. It is possible that moral luck for blameworthy acts is normatively different from moral luck in obligations of beneficence. If one rescues people with astonishing ease, the demand for further action could be reduced because there are actually fewer people needing rescue. Likewise, trying hard and failing leaves the total need undiminished. Such considerations are not relevant in the case of harm. Thus, with beneficence one could argue that moral luck normatively ought to be factored in.

Study 4: The Domain Generalizability of Moral Credit

We next examine whether moral credit diminishes as the next beneficent need becomes less similar. We tested this by asking about the acceptability of declining to help in a way that is different but contributes to the same cause (e.g., whether having worked in a soup kitchen reduces moral demand to donate money to the soup kitchen), a related cause (e.g., donating to a local homeless shelter), and an entirely independent cause (e.g., contributing to wildlife rescue). This allowed us to explore whether, as in the case of financial mental accounting, people have different buckets for different kinds of beneficence or a single moral credit account.

Study 4 Materials

We adapted versions of the boat rescue, soup kitchen, and kidney scenarios to ask about the acceptability of declining to help to a cause that is the same, similar to, or entirely different from the original cause helped (3 (scenario) x3 (domain) x2 (helped or not)). The moral act inquired about was new, related to the first (e.g., giving money to the homeless shelter) but different enough that the cost of performing it should not depend on prior effort, so any reduction in moral demand would be due to some generalizability of the moral credit from prior contributions. Subjects saw all levels of all scenario sets. See Appendix C for scenarios.

Study 4 Results

One hundred fifteen undergraduates participated in the study for partial course credit, (26 excluded for failing a basic attention check; 73.0% female, mean age 19.4, SD = 2.10). A repeated measures ANOVA with similarity (same, similar, different domains recoded as a continuous) and scenario (boat rescue, kidney donation, soup kitchen) entered as factors revealed a main effect difference of acceptability of not helping for the different scenario sets, $F(2, 176) = 20.1$, $p < .001$. $\eta_p^2 = .19$ (mean boat rescue = .57, SD = 1.84; mean kidney donation = 1.03, SD = 1.60; mean soup kitchen = 1.06, SD = 1.60). There was also a main effect of similarity, such that dissimilar causes were ones where it would be generally more acceptable to decline helping, $F(1, 88) = 32.5$, $p < .001$. $\eta_p^2 = .21$, (mean same = .62, SD = 1.81; mean similar = .82, SD = 1.72; mean different = 1.22, SD = 1.48). It was also overall more acceptable for those who had already helped to decline helping $F(1, 88) = 56.6$, $p < .001$, $\eta_p^2 = .39$ (mean no prior contribution = .45, SD = 1.70, mean no prior contribution = 1.32, SD = 1.58). The central question, however, is whether prior helping produced less moral credit for dissimilar appeals, after controlling for the main effect differences of the acceptability of declining to help the various causes. A significant interaction of similarity and prior helping suggested this, $F(1, 88) = 12.9$, $p = .001$, $\eta_p^2 = .13$ (mean same no prior contribution = .11; SD = 1.72; mean same prior contribution = 1.29; SD = 1.59; mean similar no prior contribution = .35; SD = 1.73; mean similar prior contribution = 1.13; SD = 1.76; mean different no prior contribution = .90; SD = 1.55; mean different prior contribution = 1.54; SD = 1.34), MDES $d = .22$. A three way interaction revealed that the extent to which moral credit generalizes across domains differs for the different scenarios, $F(2, 176) = 3.53$, $p = .031$, $\eta_p^2 = .04$: moral credit did not fully generalize for the soup kitchen and boat rescue scenarios, but did for kidney donation, see Figure 5. While the boat rescue cases showed a generalizability gradient ($F(1, 88) = 19.4$, $p < .001$, $\eta_p^2 = .18$), and soup kitchen cases trended towards this gradient ($F(1, 88) = 3.11$, $p = .08$, $\eta_p^2 = .03$), kidney donation cases did not ($F(1, 88) = .20$, $p = .88$, $\eta_p^2 < .01$).

Overall, the data are consistent with there being some specificity of moral credit, rather than a single moral account. While the credit does diminish for the most dissimilar case, some still exists, as revealed by the still-significant increase in the perceived moral permissibility of declining to help in the “different” condition, $F(1, 88) = 38.6$, $p < .001$, $\eta_p^2 = .30$ (mean no prior contribution = .90, SD = 1.55; mean prior contribution = 1.54; SD = 1.34). Other significant effects included an interaction of scenario by prior helping, $F(2, 176) = 23.3$, $p < .001$, $\eta_p^2 = .21$, such that prior helping had an overall greater effect on moral credit in boat rescue scenarios than in soup kitchen or kidney scenarios, and an interaction of scenario by similarity, $F(2, 176) = 11.7$, $p < .001$, $\eta_p^2 = .12$, such that similarity had the greatest effect on overall ratings of the acceptability of declining to help in boat rescue scenarios.

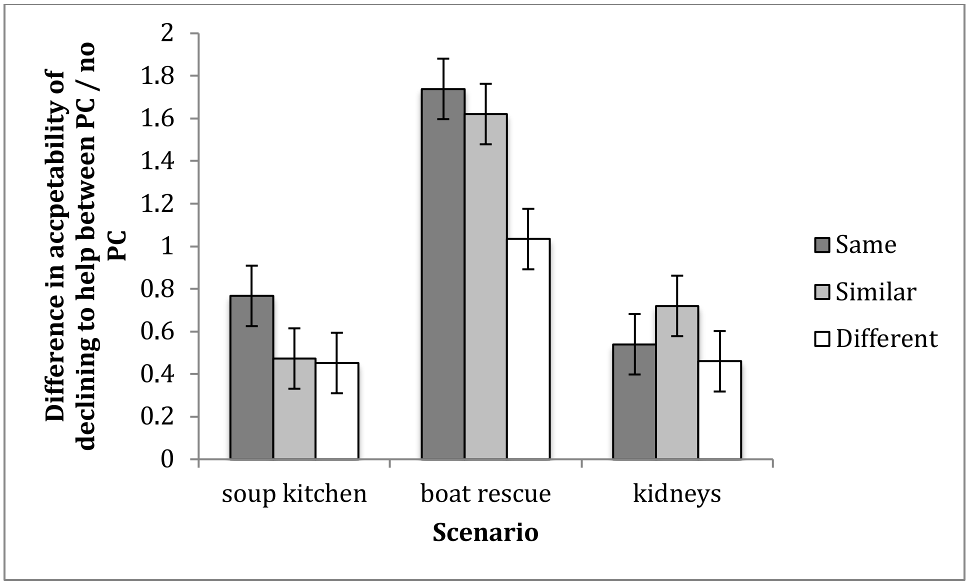


Figure 5.5. The domain generalizability of moral credit from prior contributions for each individual scenario set. Error bars represent one standard error.

Study 4 Discussion

In Study 4 we find limited generalizability of moral credit. People who have helped in a soup kitchen face a reduction in perceived moral demand to help further there, and this falls off if the need is helping at a homeless shelter. Kidney donation does not show the same specificity as boat rescues and soup kitchen volunteering, with moral credit from kidney donation generalizing across monetary donations to blood banks, and even to oil-soaked wildlife. At first glance, this result might appear in some tension with a general moral principle to promote the good of others. Should the particular recipients or causes matter? However, it is possible that people assume that spreading one's giving among causes is more likely to have greater overall effects. Thus, if some charities spend unwisely, or are unable to get resources to victims, then it might be thought better to spread that risk among a variety of causes. Against such a background assumption, being told that an agent in a scenario has only given in one area might lead to an inference that they have only discharged part of their moral debt by contributing in a single area.

Study 5: Temporal Duration of Moral Credit

Study 5 explored the temporal endurance of moral credit from prior beneficent behavior. Does the moral credit from having donated a kidney last a day? A week? A decade? Findings on the endurance of financial expenditures suggest that there could be a similar limited endurance to prior beneficence.

Study 5 Materials

We adapted versions of the soup kitchen and kidney scenarios to ask about the level of moral demand for action that is different from but contributes to the same cause (e.g., whether having worked in a soup kitchen reduces moral demand to donate money to a homeless shelter), the same day, a week, a year, and a decade later. Because of the temporal specificity of rescuing people in the wake of a hurricane, the boat rescue scenarios were replaced by scenarios involving helping at an afterschool program. We chose a different moral request from the initial action, so that having contributed would not increase the cost of future action, and there would be no presumption of a necessary physical recovery enabling further action. We also asked about the acceptability of declining to help for a person who had not yet contributed, to see at what point people no longer have moral credit for an action they have performed. Subjects saw all levels of all scenario sets. See Appendix D for scenarios. A power analysis using the correlation between measures in Study 1 and the GPower software package revealed that at least 54 participants would be needed to detect a small effect (η_p^2=.02) with an alpha of .05 and power of .80 (Erdfelder, Faul, & Buchner, 1996).

Study 5 Results

One hundred seventeen participants were recruited for partial course credit from a participant pool (13 excluded for failing attention check; 66.3% female; M_{age} = 20.4, SD = 3.25). Data was analyzed using a repeated measures ANOVA with time specified as a continuous factor and scenario set specified as a categorical factor. There was a significant effect of time on the acceptability of declining to help, such that as time since the first act increased, the acceptability of declining to help decreased $F(1,103) = 29.0$, $p < .001$, $\eta_p^2 = .23$, (mean today = 1.46, SD = 1.65; mean week = 1.31, SD = 1.62, mean year .96, SD = 1.74; mean decade = .84, SD = 1.79) , MDES $d = .19$, see Figure 6a. There was no main effect of scenario on the acceptability of declining to help, F(2, 206) = 2.23, p = .11, $\eta_p^2 = .02$ (mean soup kitchen = 1.24, SD = 1.72; mean kidney = 1.04, SD = 1.79; mean school = .93, SD = 1.71). However, there was an interaction of scenario and time, such that moral credit endured more for some scenarios than for others, $F(2, 206) = 9.43$, $p < .001$, $\eta_p^2 = .08$, see Figure 6b. Kidney donation moral credit had a slower temporal decline ($F(1, 103) = 4.17$, $p = .04$, $\eta_p^2 = .04$) than moral credit from helping at an afterschool program ($F(1, 103) = 19.0$, $p < .001$, $\eta_p^2 = .15$) or volunteering at a soup kitchen ($F(1, 103) = 26.5$, $p < .001$, $\eta_p^2 = .20$). We next examined whether there was any remaining moral credit a decade later. For kidney scenarios there was $t(103) = 2.22$, $p = .029$, $d = .38$ (mean kidney decade = 1.02, SD = 1.77, mean kidney no prior contribution = .654, SD = 1.84), unlike for afterschool helping $t(103) = 1.18$, p = .24, d = .21 (mean school decade = .740, SD = 1.75; mean school no prior contribution = .596, SD = 1.88), or soup kitchen helping, where there was an effect in the opposite direction *t*(103) = 2.74, *p* = .007, *d* = .48 (mean soup decade = .769, SD = 1.87, mean soup no prior contribution = 1.12, SD = 1.67).

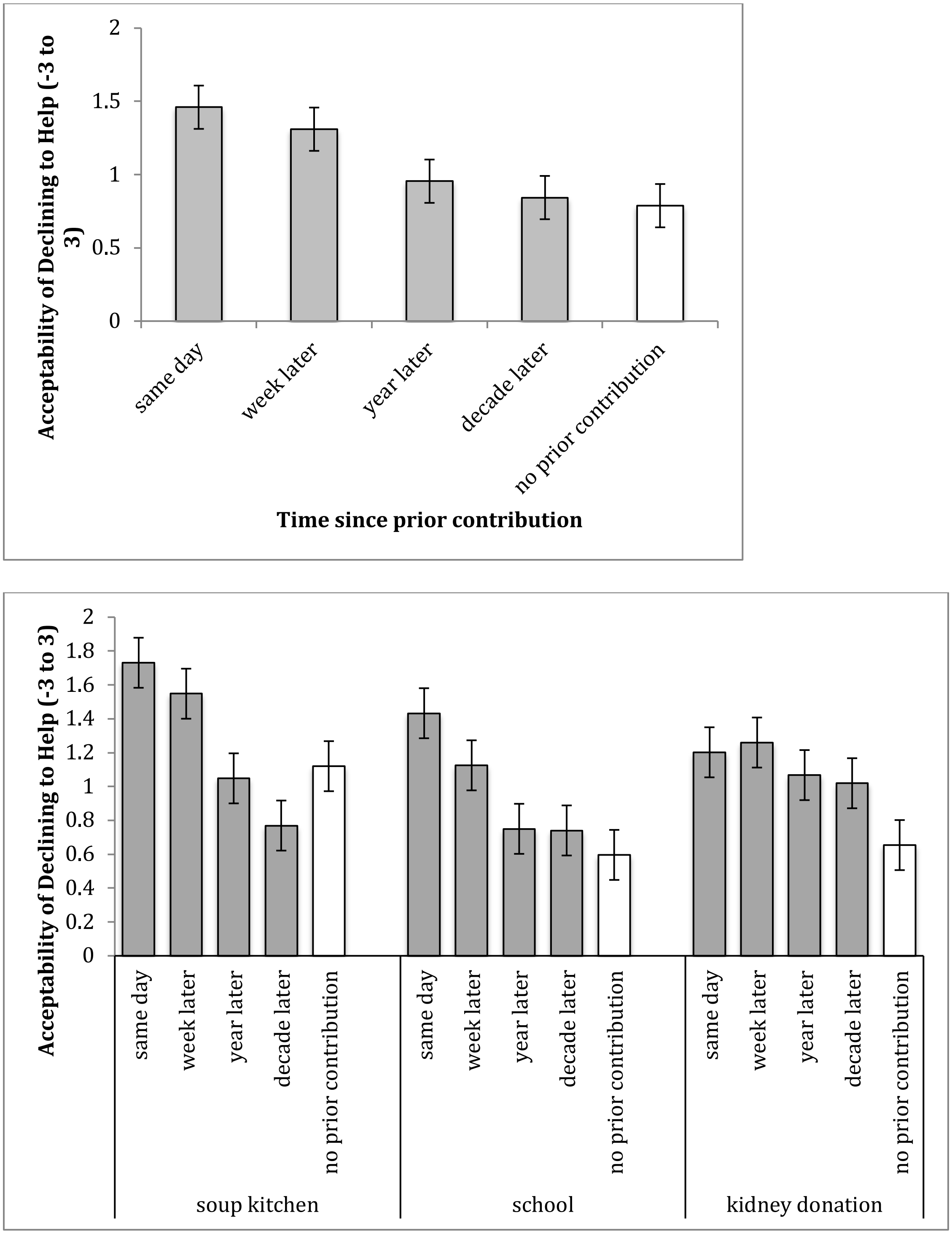


Figure 5.6. The temporal generalizability of moral credit from prior contributions (a) and temporal generalizability for each individual scenario set (b). Error bars represent one standard error.

Study 5 Discussion

Moral credit appears to be temporally sensitive, in that the moral acceptability of declining to help a cause to which one had previously contributed diminishes over time. Some moral credit appears to be more enduring — a kidney donation continued to give moral credit a decade after the donation. However, moral credit for having helped in a soup kitchen or an after school program disappeared entirely. It could be that donating an organ gives indefinite credit because the recipient will spend the rest of their life without that organ, which could be viewed as continuing to make a donation. Testing whether a donation that regenerates, such as a liver lobe, or blood, produces less enduring credit could explore such a notion.

Less clear, normatively, is how long moral credit should extend for acts that are not accompanied by a perceived indefinite cost. At first glance, it might seem that if people are morally required to do a substantial, albeit limited, amount of good, then they could do all of their allotted good early on in their lives and then be required to do no more. But this seems counterintuitive, and also counter to the findings regarding participants' reactions to particular scenarios.

One way of reconciling participants' reactions with normatively plausible principles would be to offer normative reasons for adopting a time-sensitive principle. Various moral theories have the resources to do this. For example, on a simple virtue theory, according to which one ought to act as a virtuous person would, what is crucial is to possess and cultivate virtuous dispositions, and to be oriented toward the good (e.g., Zagzebski, 2010). If one never again acted well when in apparently relevant circumstances, then this would cast doubt on the idea that one in fact has the relevant dispositions. Even on a Kant-inspired deontological theory that directs us to promote the good of others (though not necessarily in a maximizing way), it seems that this should always be an action-guiding maxim (Kant, 1785/2002). It would cast doubt on one's having adopted a maxim of beneficence if one never appeared to act on it after a certain point. Thus, there are resources in moral theory, together with empirical facts about humans' finite capacities and situational factors, that go some way toward reconciling participants' reactions with plausible normative principles.

Moral credit does appear, as our model suggests, to decay over time. One implication of this is that it might be possible to change the timing of a gift of a given size to alter how much credit is received, over time. For example, one could volunteer once for ten hours, or for 2 hours per week for 5 weeks. The latter, more diffuse form of beneficence may grant more credit, if the recency of helping outweighs the decreased credit for amount of good done, analogously to how monthly payments result in increased gym attendance compared to equivalent annual fees (Gourville & Soman, 1998). Study 6 thus explored whether smaller amounts of good, done over longer periods of time, give more enduring moral credit than a single large amount of good done.

Study 6: Temporal Diffusion of Donations

For this study, we explored temporal diffusion in a domain where it is frequently encountered: monthly vs. one-time charitable donations. If recency matters, it could be the case that the recency of continued monthly donations outweighs them being smaller than a single lump sum given some number of months ago.

Study 6 Materials

In Study 6 the soup kitchen and after school helping scenario sets were adapted to explore how the moral credit from monthly contributions compares to single contributions both when first made (e.g., how does the moral credit from signing up to make monthly donations compare to the moral credit for donating a larger amount immediately), and several months after the initial request (e.g., how acceptable is it to decline helping a soup kitchen if eight months ago one donated \$240 to it, vs. signed up to make monthly donations of \$20/month for a year). The new moral act inquired about was related to the first (e.g., making an additional donation to the soup kitchen). Subjects saw all levels of both scenario sets. See Appendix E for scenarios.

Study 6 Results

Ninety-four participants were recruited for partial course credit from an undergraduate participant pool (8 excluded for failing attention check, 70.9% female, Mage = 20.2, SD = 2.41). A 2 (after school help vs. soup kitchen) x 2 (single vs. monthly donation) x 2 (time: same day vs. months later) repeated measures ANOVA revealed a predicted main effect of time $F(1, 85) = 13.7$, $p < .001$, $\eta_p^2 = .14$ (same day mean = 2.26, SD = 1.13; months later mean = 2.06, SD = 1.29). Regarding our central question as to whether monthly donations give more enduring credit than a single lump-sum donation, a significant donation type by time interaction indicated that monthly donations, 8-9 months after commencing, give more moral credit than having donated a lump sum, $F(1, 85) = 5.05$, $p = .027$, $\eta_p^2 = .06$ (same day single mean = 2.31, SD = 1.04; months-later single mean = 1.99, SD = 1.33l same day monthly mean = 2.20, SD = 1.21; months-later monthly mean = 2.12, SD = 1.25), MDES *d* = .26, suggesting a differential decay of credit for the one-time and monthly donations, see Figure 7.

There was no main effect of scenario *F*(1, 85) = .57, *p* = .45, η_p^2 = .01 (soup kitchen mean = 2.20, SD = 1.21; school mean = 2.11, SD = 1.23), nor was there a main effect of type of donation, *F*(1, 85) = .028, *p* = .87, $\eta_p^2 < .01$ (monthly mean = 2.16, SD = 1.23; single mean = 2.15, SD = 1.21). The interaction of scenario and time was not significant, *F*(1, 85) = 2.15, *p* = .15, η_p^2 = .02. There was, however, a significant interaction of scenario and type of donation, *F*(1, 85) = 8.25, *p* = .005, η_p^2 = .09 (the monthly donation trended towards more moral credit than the single donation in the soup kitchen scenario, *F*(1, 85) = 2.13 *p* = .13, η_p^2 = .03, and less moral credit than single donation in the school scenario, *F*(1, 85) = 2.57, *p* = .11, η_p^2 = .03). The 3-way interaction between scenario, time, and donation type was not significant *F*(1, 85) = .92, *p* = .34, η_p^2 = .01).

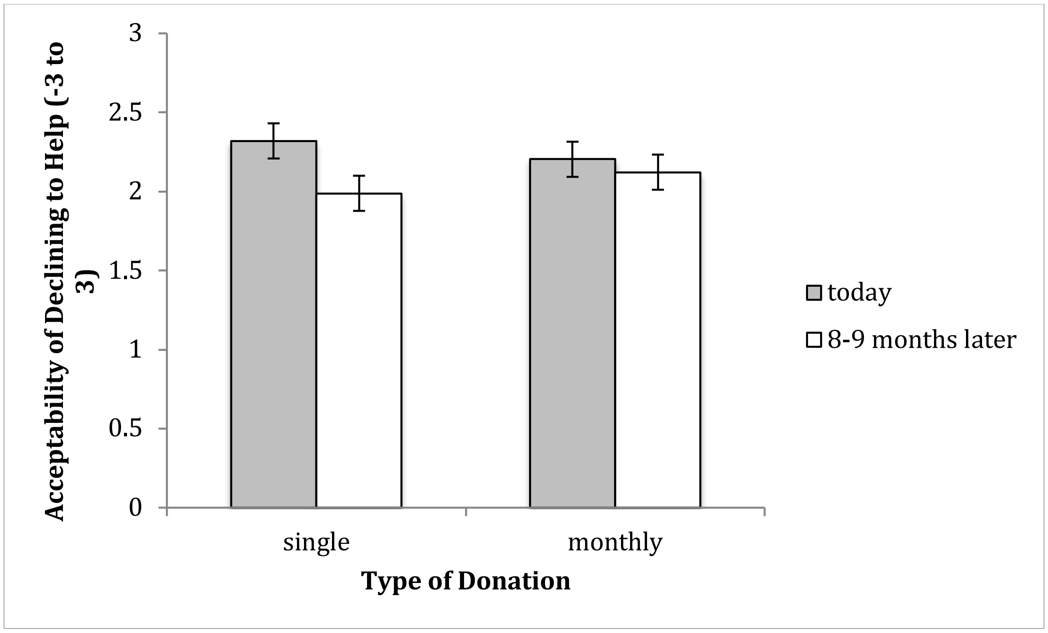


Figure 5.7. The temporal generalizability of moral credit from single-instance and monthly recurring donations which total the same amounts. Error bars represent one standard error.

Study 6 Discussion

In Study 6 we find that spreading out the single donation into monthly installments spreads the moral credit from the action, making it more enduring. Monthly donations are generally offered with the thought that they lower the barriers to giving and encourage more gifts. But such an effect could be offset if, by replacing one-time larger donations, they give people more enduring moral credit, and so inhibit future giving.

It might seem that given any plausible moral principle of beneficence, it should make no difference to one's moral credit whether one donates a lump sum at the beginning of the year or spreads out the same donation over twelve months. But it is possible that subjects infer that someone who donates in monthly installments does not have as much disposable wealth as someone who donates in a lump sum. So, we may expect more total contribution from the person who has more to give, and demand more from that person nine months later, not (only) because time has gone by, but because we believe that they are in a position to add to their initial contribution at a lower cost. It might also be that participants infer from the fact that a person makes regular contributions to a good cause that she has made it a goal to be beneficent, and this itself might be taken as a sign that she is virtuous and thereby deserves greater moral credit. Given these possibilities, it would be premature to conclude from Study 6 that participants' conceptions of moral credit are untethered to normative morality.

Study 7

We take the prior studies to indicate participants' normative beliefs regarding the requiredness of beneficence, as joint evaluation lets participants decide whether a difference between different variants of the same scenario *should* matter. Equally important is whether those differences actually do matter when evaluating single versions of each scenario. It could be that participants believe that helping a month ago should grant less moral credit than helping a week ago, but, assign them equivalent credit when not directly comparing actions. In order to explore whether our model of lay theory of moral credit describes how moral credit is assigned to individual actions, we presented participants with one level of each scenario set from Studies 2-6. Participants were presented with one scenario from each of the scenario sets used in each study. Four hundred seventeen participants were recruited for partial course credit from an undergraduate participant pool (107 excluded for failing attention check, 74.8% female, Meage = 20.2, SD = 3.35).

All analyses were conducted using two-level hierarchical linear mixed-effects models, with subjects specified as random intercepts, and scenario set and the factors of interest specified as fixed factors, using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2014). Model fit was assessed using likelihood ratio tests. Sensitivity power analyses were conducted for Study 7 analyses (using alpha = .05 and beta = 80%) to determine the minimum detectable effect size (MDES; unstandardized) using the simr package in R (Green & MacLeod, 2016) and 1,000 resamples.

Study 7 Results

Study 2 Scenarios. Study 2 examined whether prior contribution and future cost both matter. The mixed model yielded a significant difference in acceptability of declining to help by cost, $\chi^2(1)=72.8$, $p < 0.001$), $\beta = .956$, SE = .120, (mean high cost = .93, SD = 1.84; mean low cost = -.05, SD = 1.92), MDES = .30, and by prior contribution, $\chi^2(1)=39.3$, $p < 0.001$), $\beta = .700$, SE = .111, (mean prior contribution = .77, SD = 1.87; mean no prior contribution - .04, SD = 1.96), MDES = .30, see Figure 8. There was no significant interaction between prior contribution and future cost, $\chi^2(1)= .02$, p = 0.90, $\beta = .028$, SE = .214.

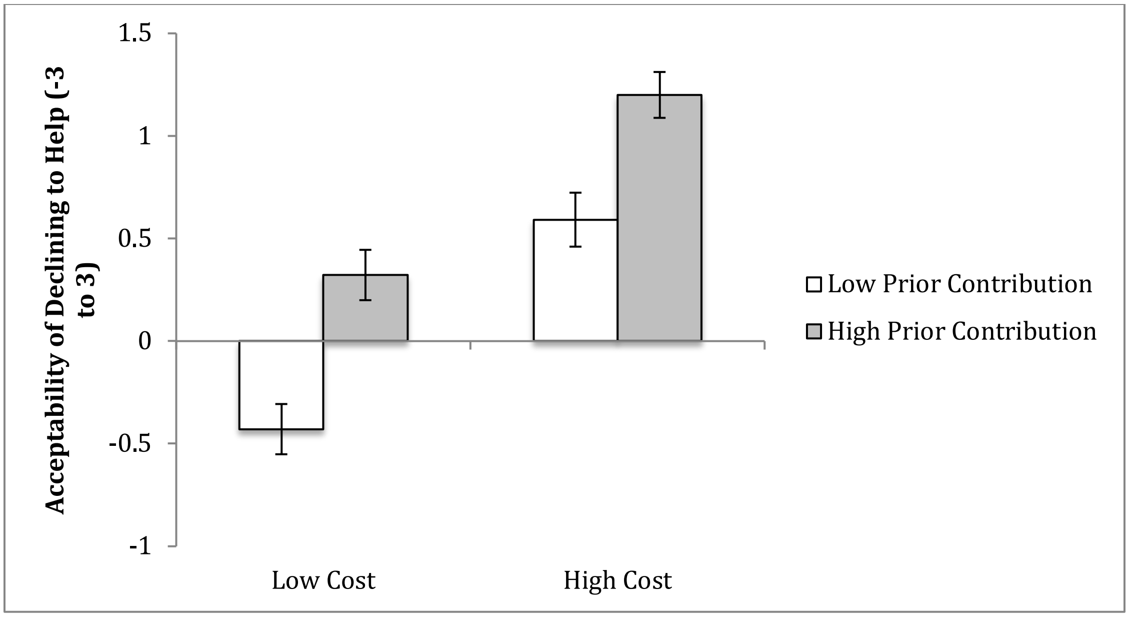


Figure 5.8. The effects of cost of further action and prior contribution on the acceptability of declining to help when seeing one level of each scenario set. Error bars represent one standard error.

Study 3 Scenarios. Unlike in our within-subjects study, participants did not think amount of effort mattered when moral luck (whether the action was successful or not) varied. The mixed model yielded a significant difference in acceptability of declining to help by luck, $\chi^2(1)$= 7.16, p = 0.007, β = .341, SE = .127, (mean unsuccessful = .36, SD = 1.92; mean successful = .68, SD = 2.04), MDES = .35, but not by effort, $\chi^2(1)$= .32, p = .57, β = .350, SE = .127, see Figure 9, MDES = .35 (mean high effort = .61, SD = 1.96; mean low effort = .43, SD = 2.00). There was no significant interaction between luck and effort, $\chi^2(1)$=1.48, p = 0.22, β = .22, SE = .31.

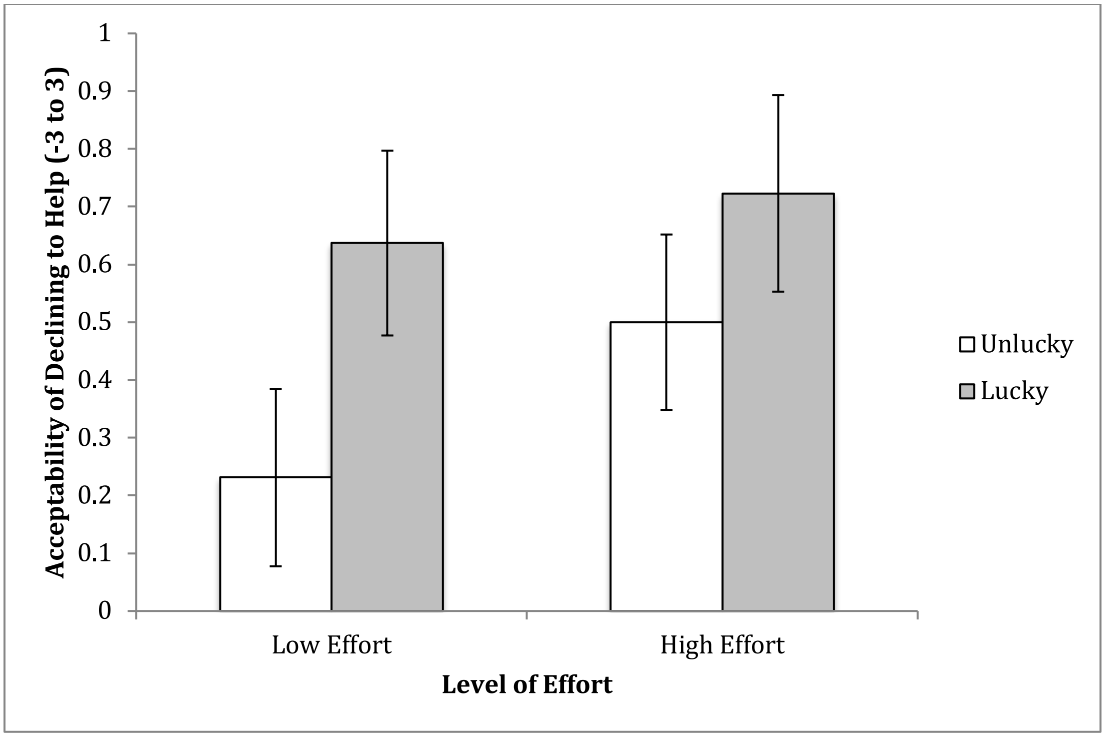


Figure 5.9. The effect of effort and luck on the acceptability of declining to help when seeing one level of each scenario set. Error bars represent one standard error.

Study 4 Scenarios. In Study 4, participants' normative position was that moral credit from prior beneficence should have limited domain generalizability. A mixed model with the interaction of domain (coded as continuous) and prior helping did not fit the data better than a model only containing their main effects, inconsistent with findings from our within-subject design, $\chi^2(1)= 2.35$, p = 0.13, β = .17, SE = .11, MDES = .31, (mean help same domain = 1.32, SD = 1.60; mean no help same domain = .62, SD = 1.74; mean help similar domain = 1.50, SD = 1.50; mean no help similar domain = .69, SD = 1.78; mean help different domain = 1.64, SD = 1.52; mean no help different domain = 1.15, SD = 1.72) see Figure 10. We thus did not observe the limited domain generalizability of moral credit found in the within-subjects design.

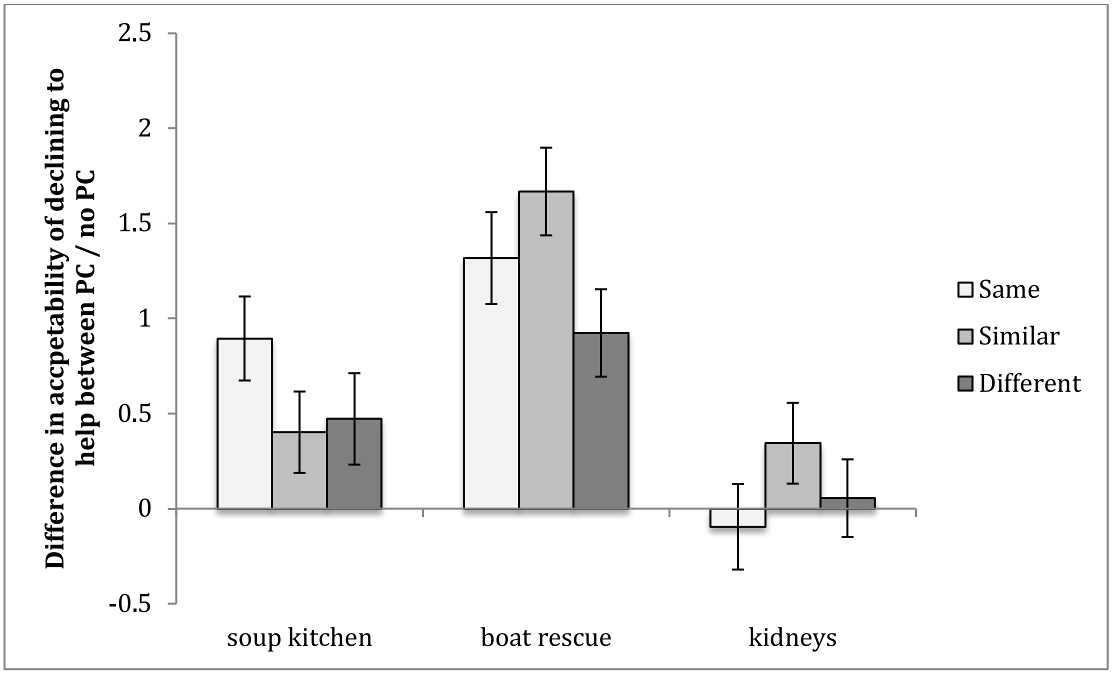
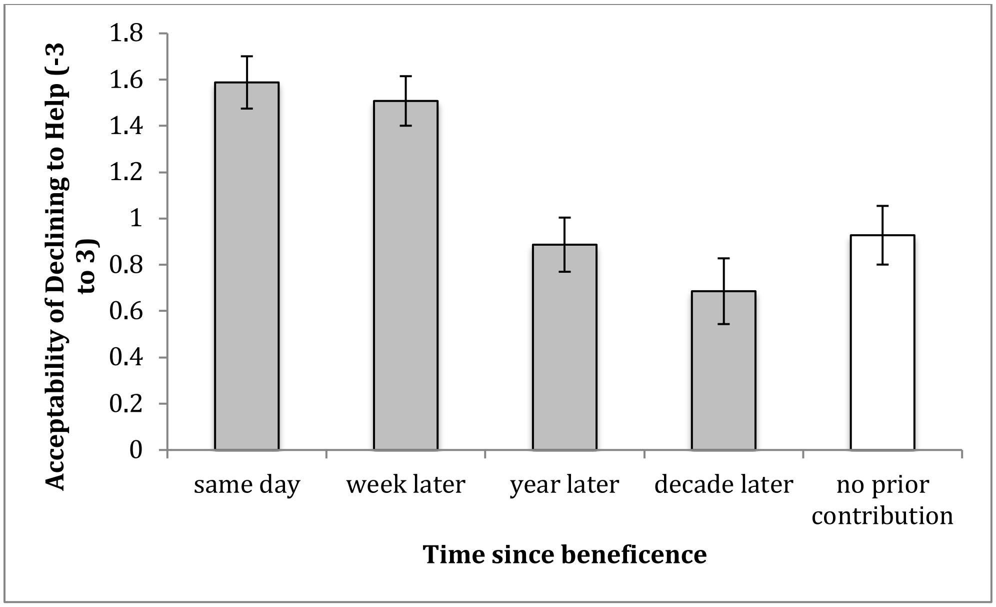


Figure 5.10. The domain generalizability of moral credit from prior contributions as expressed through difference in acceptability of declining helping for actors who had helped, and actors who had not helped, for participants seeing one level of each scenario set. Error bars represent average of one standard error for scores from which the difference score is derived.

Study 5 Scenarios. As in the within-subjects design, the mixed model yielded a significant decrease in the acceptability of declining to help as time since prior beneficence increased (coded as continuous), ($\chi^2(1)$= 49.3, p < 0.001), β = .342, SE = .048, (mean today = 1.59, SD = 1.53; mean week = 1.51, SD = 1.45; mean year = .89, SD = 1.71, mean decade = .69, SD = 1.85), MDES = .13 see Figure 11a. There was a significant interaction between time and scenario, $\chi^2(2)$= 8.99, p = 0.01, such that kidney donation gave more enduring credit than soup kitchen volunteering or after-school helping, see Figure 11b.



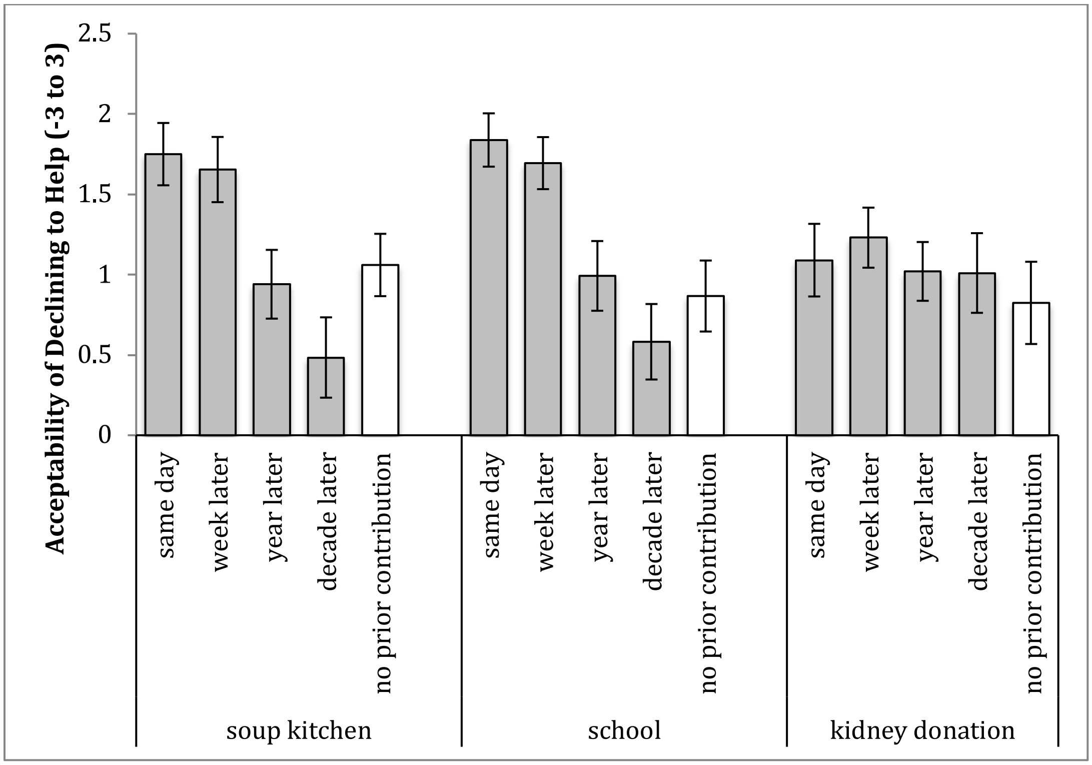


Figure 5.11. The temporal generalizability of moral credit from prior contributions (a) and temporal generalizability for each individual scenario set (b), when presenting participants with one level of each scenario set. Error bars represent one standard error.

Study 6 Scenarios. In Study 6 we were interested in whether signing up for a monthly donation equivalent in total value to a single lump sum donation could grant more enduring moral credit. The mixed model yielded a significant difference in acceptability of declining to help by type of donation, such that monthly donations gave more credit as a main effect, $\chi^2(1)= 13.09$, $p < 0.001$, $\beta = .332$, SE = .112. There was no main effect of time since donation, $\chi^2(1)= 2.42$, $p = .12$, $\beta = .005$, SE = .112. There was, importantly, a significant interaction between luck and effort, $\chi^2(1)=3.97$, $p = 0.046$), $\beta = .36$, SE = .18, (mean monthly same day = 2.18, SD = 1.09; mean monthly months later = 2.14, SD = 1.26; mean single same day = 1.97, SD = 1.40; mean single months later = 1.73, SD = 1.31), MDES = .50, suggesting that the observed main effect difference was driven by a decline in acceptability of declining to help after having made a lump sum donation months earlier.

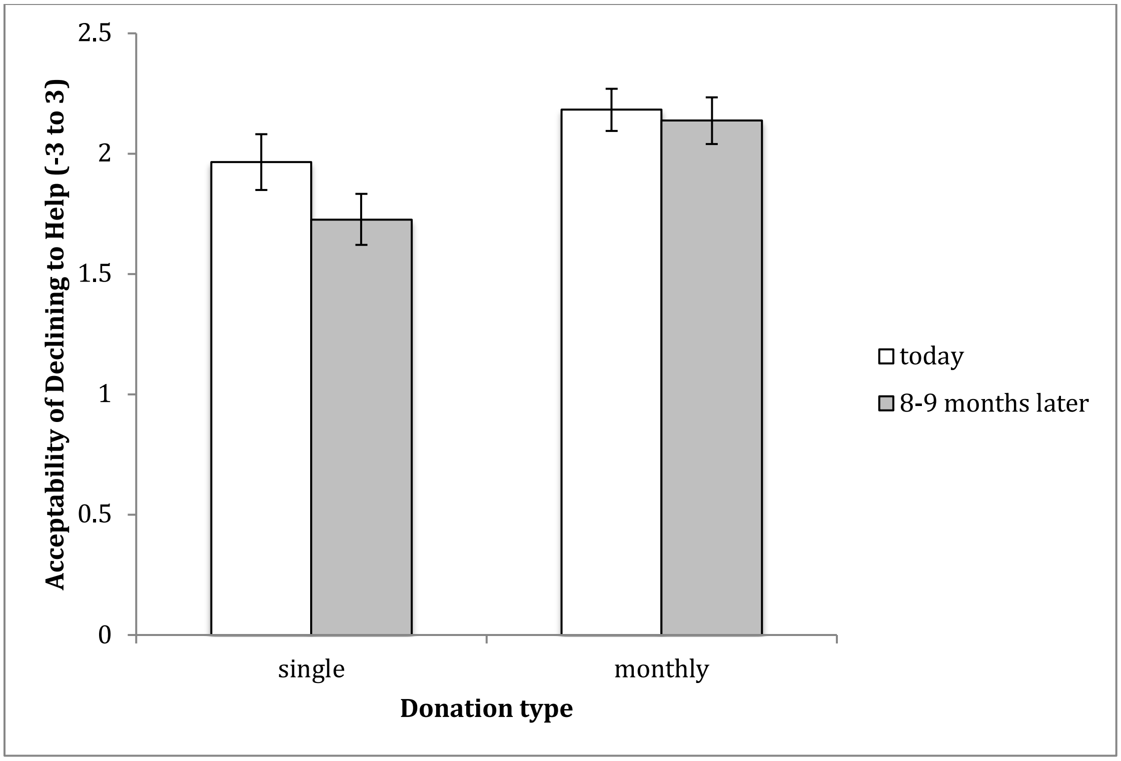


Figure 5.12. The temporal generalizability of moral credit from single-instance and monthly recurring donations which total the same amounts, when seeing one level of each scenario set. Error bars represent one standard error.

Study 7 Discussion

The lay theory of moral accounting we uncovered over Studies 1-6 is largely consistent with how people actually assign moral credit when not seeing side-by-side differences for scenario sets. Participants take into account the increased cost of future action, prior contribution, luck, recency, and temporal diffusion. There are, however, two differences between the lay theory of how moral credit should be assigned, and how credit appears to be assigned: the amount of effort put forth appears to not matter for actual judgment if effort was not ultimately successful, and moral credit in the case of actual judgment seems to generalize equally to unrelated domains.

General Discussion

In six studies we tested the core aspects of the Moral Accounting Model, which uncovers the lay theory of how prior beneficence reduces moral demand. People judge that after having helped there is less moral demand to continue helping (Study 1). Two paths contribute to decreased beneficence: The increased cost of continuing to help and moral credit (Study 2). Study 3 reveals that moral credit consists of prior effort and prior effect, has some domain specificity (Study 4), and is temporally sensitive (Study 5). This temporal specificity can cause a gift that is spread out over time to produce more lasting moral credit than a one-time donation of equal total size (Study 6). Study 7, which only presents one level of each scenario set, finds that the lay theory of which factors should matter, uncovered in Studies 1-6, is largely consistent with single-level judgments of actors. The emergent difference between lay theories of moral accounting and how participants actually judge others is in amount of effort not mattering to individual judgment, despite participants taking the normative stance that effort should matter.

Our model suggests that people have an intuitive solution to the problem of the demandingness of moral action. The findings suggest that, even with the global need unmet, people believe that it is at least possible to have done enough, morally speaking. This is not surprising in itself. At the same time, the findings show that a subtle collection of factors informs this judgment. Some are very plausibly morally relevant in themselves on a wide swath of moral theories, such as an increased cost to future helping once one has helped already, and a potential decrease in need from successful prior efforts. Others are arguably morally relevant in themselves as well, as long as one takes demands to be a function of something other than maximizing well-being, such as prior effort. And while some others, such as generalizability in domain and time, as well as diffusion, appear to be in tension with plausible moral principles, it is possible that background assumptions can defuse this tension.

Evidence for the limited generalizability of the moral credit suggests that mental moral credit accounting may function in much the same way as mental financial accounting, with moral credit not entirely fungible and time-sensitive.

While our work has identified a series of factors that are important in determining the level of perceived moral demand, exploring how these factors interact could further refine the model. For example, it could be that the temporal decay of moral credit is faster in domains that are more different from the original act. It could also be that bad luck in moral efforts does not diminish credit in other domains as much as in the same domain, if the significance of bad luck derives from the fact that that particular need is still unmet.

A diminished demand for moral action is clearly not the same as moral action becoming less admirable, and likely is even the reverse. Many ethical approaches distinguish between acts that are morally required and acts that are morally supererogatory, i.e., beyond what is required. It could well be the case that the factors we have identified as intuitive reasons for reducing the moral demands of beneficence also function to increase the praiseworthiness of a person who acts regardless.

Our results complement moral licensing findings, where previous good done excuses future moral lapses. People seem to regard prior contributions not just as an excuse for ethical indolence, but as a general principle that reduces the demands of beneficence. In other words, in some cases, doing good puts one in a position where one needs no excuse; one simply doesn’t violate the demands that would otherwise be in place. Our results also provide empirical evidence for helping being perceived as a finite resource outside of the self-regulation domain and methods typically employed in it: decreased moral demand is placed on others who have tried to do good, or have even incidentally done good without trying.

Empirical inquiries such as ours can also help inform normative debates regarding the factors that determine the moral demands of beneficence. Inasmuch as moral theory is sensitive to people’s intuitions, it should consider incorporating cost, prior effort, and prior good done as relevant to fixing what is morally required of us in the way of helping others in need. At least some of these factors have a good claim to be normatively relevant on their own. When it comes to others, there seems to be nothing inherently normatively significant in them, but it is possible to explain how they might be seen to be systematically, albeit contingently, correlated with normatively relevant features. Attention to isolating these factors can help reveal potentially morally relevant factors not previously distinguished clearly.

At the same time, our model can also contribute to increasing the impact of philanthropic efforts. Those in government or in the non-profit sector who are working to increase fundraising or volunteerism may find it easier to achieve their aims if they take our results on board: a better understanding of the factors that shape perception of how much aid to others one is morally required to give could contribute to the development of more effective solicitations.

Chapter 5 has been submitted for publication of the material as it may appear in the Journal of Experimental Social Psychology, 2019. Ryazanov, Arseny; Nelkin, Dana; Rickless, Samuel; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this material.

References

American Law Institute. (1985). Model penal code : official draft and explanatory notes : complete text of model penal code as adopted at the 1962 annual meeting of the American Law Institute at Washington, D.C., May 24, 1962. Philadelphia, Pa. :The Institute.

Arneson, R. J. (2004). Moral limits on the demands of beneficence? *The Ethics of Assistance*, 33-58.

Barnes, R. D., Ickes, W., & Kidd, R. F. (1979). Effects of the perceived intentionality and stability of another's dependency on helping behavior. *Personality and Social Psychology Bulletin*, *5*(3), 367-372.

Barak-Corren, N., Tsay, C., Cushman, F., & Bazerman, M. (2018). If you're going to do wrong, at least do it right: The surprising effect of considering two moral dilemmas at the same time. *Management Science*, *64*(4), 1528-1540.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beauchamp, T. (2016). The principle of beneficence in applied ethics, *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2016/entries/principle-beneficence/.

Bell, J., Grekul, J., Lamba, N., Minas, C., & Harrell, W. A. (1995). The impact of cost on student helping behavior. *The Journal of Social Psychology*, *135*(1), 49-56.

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, *41*(4), 540-558.

Camerer, C. F. (1997). Taxi drivers and beauty contests. *Engineering and Science*, *60*(1), 10-19.

Cascio, J., & Plant, E. A. (2015). Prospective moral licensing: Does anticipating doing good later allow you to be bad now? *Journal of Experimental Social Psychology*, *56*, 110-116.

Cullity, G. (2006). *The moral demands of affluence*. Oxford University Press.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS one*, *4*(8), e6699.

Dworkin, R. (2002). *Sovereign virtue: The theory and practice of equality*. Harvard University Press.

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality and Social Psychology Bulletin*, *36*(12), 1618-1634.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1-11.

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

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., & Norton, M. I. (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science*, *58*(1), 179-187.

Goodwin, G. P., & Landy, J. F. (2014). Valuing different human lives. *Journal of Experimental Psychology: General*, *143*(2), 778.

Gourville, J. T., & Soman, D. (1998). Payment depreciation: The behavioral effects of temporally separating payments from consumption. *Journal of Consumer Research*, *25*(2), 160-174.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498.

Heath, C., & Soll, J. B. (1996). Mental budgeting and consumer decisions. *Journal of Consumer Research*, *23*(1), 40-52.

Herman, B. (2001). The scope of moral requirement. *Philosophy & Public Affairs*, *30*(3), 227-256.

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: on the affective psychology of value. *Journal of Experimental Psychology: General*, *133*(1), 23.

Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*(3), 235-257.

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*(5), 701-713.

Kagan, S. (1989). *The limits of morality.* Oxford University Press.

Kant, I. (1785/2002). *Groundwork for the metaphysics of morals*. Translated by A. Zweig and edited by A. Zweig and T. E. Hill Jr. Oxford University Press.

Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study in the headline method. *Journal of Risk and Uncertainty*, *9*(1), 5-37.

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people? *Psychological science*, *21*(4), 494-498.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, *4*(5), 344-357.

Miller, R. W. (2004). Beneficence, duty and distance. *Philosophy & Public Affairs*, *32*(4), 357-383.

Moore, M. (1997). *Placing blame: A theory of the criminal law*, Oxford: Clarendon Press.

Murphy, L. B. (1993). The demands of beneficence. *Philosophy & Public Affairs*, 267-292.

Nagel, T. (1979). Moral luck. In *Mortal questions*, New York: Cambridge University Press.

Nagel, T. (1989). *The view from nowhere*. Oxford University Press.

Palmer, C. J., Paton, B., Ngo, T. T., Thomson, R. H., Hohwy, J., & Miller, S. M. (2013). Individual differences in moral behaviour: A role for response to risk and uncertainty? *Neuroethics*, *6*(1), 97-103.

Royzman, E., & Kumar, R. (2004). Is consequential luck morally inconsequential? Empirical psychology and the reassessment of moral luck. *Ratio*, *17*(3), 329-344.

Ryazanov, A. A., Knutzen, J., Rickless, S. C., Christenfeld, N. J., & Nelkin, D. K. (2018). Intuitive Probabilities and the Limitation of Moral Imagination. *Cognitive science*, *42*, 38-68.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners the paradox of moral self-regulation. *Psychological Science*, *20*(4), 523-528.

Scheffler, S. (1993). *Human morality*. Oxford University Press.

Scheffler, S. (1994). *The rejection of consequentialism*. Oxford University Press.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 229-243.

Singer, P. (2011). *Practical ethics*. Cambridge University Press.

Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*(1), 5-16.

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, *102*(2), 143-153.

Tanner, R. J., & Carlson, K. A. (2008). Unrealistically optimistic consumers: A selective hypothesis testing account for optimism in predictions of future behavior. *Journal of Consumer Research*, *35*(5), 810-822.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral decision making*, *12*(3), 183.

Touré-Tillery, M., & Fishbach, A. (2017). Too far to help: The effect of perceived distance on the expected impact and likelihood of charitable action. *Journal of personality and social psychology*, *112*(6), 860.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, S251-S278.

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

Vallentyne, P. (2006). Against maximizing act consequentialism. *Contemporary debates in moral theory*. 21-37.

Westfall, J. (2015). PANGEA: Power analysis for general anova designs. *Unpublished manuscript. Available at http://jakewestfall. org/publications/pangea. pdf.*

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

Zagzebski, L. (2010). Exemplarist virtue theory. *Metaphilosophy*, 41(1/2), 41–57.

Chapter 6: On the Limited Role of Efficiency in Charitable Giving

Arseny A. Ryazanov & Nicholas J. S. Christenfeld

University of California, San Diego

Corresponding Author: Arseny Ryazanov (aryazano@ucsd.edu)

Dept. of Psychology, UC San Diego

9500 Gilman Drive, La Jolla, CA 92093-0109

Other author contact info: nchristenfeld@ucsd.edu; same address as above

Abstract

Performance measurement is considered useful in guiding donations to charities. We investigated whether efficiency rates predominately guide donations relative to available alternatives, or influence donation amounts. Across 4 studies (N = 460) participants evaluating charity advertisements saw randomly assigned efficiency rates presented as background information. Participants could pledge a portion of a gift card, offered in return for participation, to their pick of presented charities. Participants were sensitive to relative, but not absolute, efficiency, giving more often to more relatively efficient charities, but generally did not pledge them more money. Even providing an explicit standard of efficiency did not create an absolute sensitivity to efficiency, suggesting that efficiency information, steers, rather than encourage, or discourage, donations overall.

Keywords: Efficiency, Donation, Performance Measurement, Giving,

On the Limited Role of Efficiency in Charitable Giving

Deciding whether, how much, and to whom to give is a daunting task, as there is an overwhelming number of charities vying for philanthropic donations. They vary in their mission, from saving frogs, to providing wished-for experiences to seriously ill children, their scope from regional to global, and their budgets from tiny to billions annually.

Charities also vary enormously in the efficiency with which they deliver the donations to the targets of the philanthropy, rather than spending the donations on the bureaucracy of the charity itself. The Kids Wish Network, for example, collected approximately 18.6 million dollars in 2012, yet used only 240,000 dollars directly on its cause of granting sick children's wishes (Taggart, 2017). In contrast, Direct Relief International, in 2016, spent all $1.104 billion of the money they raised from individual donations directly on providing humanitarian medical aid, with none going to administrative costs (Barrett, 2017).

Quantifiable performance measurement has become viewed as a particularly convincing way of appealing to donors (Cunningham & Ricks, 2004). Metrics that provide greater non-profit transparency and accountability should direct funding towards more deserving charities, and may even increase total giving by increasing donor confidence (Cunningham & Ricks, 2004). We investigated to what extent performance measurement motivates donation.

A simple way of conveying charity efficiency is through the percentage of money collected that is spent on the actual cause, as opposed to the overhead of the organization (Stork & Woodilla, 2008). While conceptually simple, determining efficiency becomes more complex in practice, because exactly what counts as overhead is somewhat ambiguous. Furthermore, the marginal efficiency rate can differ from the overall efficiency rate: the first dollar received likely goes all to overhead, and the last dollar much more to the end cause.

Proponents of efficiency rates suggest they can improve and increase philanthropy, believing that donors allocate personal resources by using price and quality like they do in private goods consumption (Callen, 1994). High efficiency rates should signify a higher quality product that would then solicit greater donations. Several online services provide potential donors with ratings of efficiency, such as letter grades similar to a school report card, that are predominantly based on financial efficiency (Lowell, Trelstad, & Meehan, 2005).

However, efficiency rates and ratings are not without their opponents. Preoccupation with appearing efficient can cause charities to handicap themselves by avoiding investments in infrastructure that could increase actual effectiveness, or funding the most efficient programs rather than the most effective ones (Gregory & Howard, 2009; Hager & Wing, 2004). Consequentially, some argue that comparing overheads between charities is meaningless (Bowman, 2006; Steinberg, 1986). Furthermore, misreporting efficiency is common in the nonprofit sector (Krishnan, Yetman, & Yetman, 2005). In one estimate, 75-85% of public charities misreport costs in order to appear more efficient (The Bedsworth, Gregory, & Howard, 2008). Although opponents of efficiency rates argue that overemphasis on efficiency decreases charity effectiveness, both sides presuppose that people rely on these financial indicators in philanthropic decisions (Gregory & Howard, 2009).

Evidence for high efficiency rates positively affecting donation is limited (Szper & Prakash, 2011). Modest correlations have been observed between efficiency and donations given to, and volunteering for, particular organizations (Callen, 1994; Bowman, 2006). However, other studies have found that efficiency ratings are largely ignored. Only one in five donors studied consulted charity watchdog websites before making donations in the past year (Cnaan, Jones, Dickin, & Salomon, 2011). Hope Consulting (2012) reported that only 3% of donors claimed even to have considered alternatives by effectiveness, the information that efficiency rates are supposed to convey. Preliminary laboratory studies show more of an effect: when choosing between two fictional charities, participants preferred to give to the option with a smaller overhead, though overhead was defined as percentage of the donation kept by the experimenter, rather than as the efficiency of the actual charity (U. Gneezy, Keenan, & Gneezy, 2014; see their field study for evidence of how covering overhead can encourage donation). Caviola, Faulmüller, Everett, Savulescu & Kahane (2014) likewise found that giving was sensitive to efficiency: participants asked to imagine donating to a particular fictional charity stated that they would hypothetically give more to a more efficient charity, even if the less efficient charity was actually more effective, and saved more lives.

Expanding on prior work, we explore two ways in which efficiency might influence donations to charity. The first is through sensitivity to the actual level of efficiency: charities that are highly efficient simply receive bigger donations. The second is through relative efficiency: charities that are more efficient than immediately-present alternatives garner the contributions, but the amount given is independent of absolute efficiency. Increased giving to more efficient charities could then be the result of more frequent, rather than more substantial, gifts.

Andreoni's theory of warm-glow giving suggests that donations are made, at least in part, as the result of donors wanting to feel a "warm-glow"(1990). But convincingly distinguishing between giving out of desire for that warmth, rather than warmth emerging as a natural consequence of giving, has proven difficult. Psychological studies on whether helping is done for egoistic reasons or is truly altruistic have been unable to reliably parse the two in real-world settings (Batson, 2014; Cialdini, 1991). Without settling the altruism debate, we can build on the notion that giving can be associated with a warm, affective component, and a colder deliberative component.

Previous studies suggest that there may be a general aversion to cold statistics in charitable donation. Jenni and Loewenstein (1997) found that appeals highlighting individual victims garnered larger donations than those emphasizing statistical victims, or large groups in need of aid. Informing donors of this single-victim bias did not cause increased donation to statistical victims, but instead decreased donation to single-victims (Small, Loewenstein, & Slovic, 2007). Small et al. (2007) suggest that this effect may exist due to a distinction, first proposed by Zajonc (1984), between a faster, more automatic affective system, and slower, more effortful deliberative system. Participants, informed of their bias, may have deliberated about identifiable victims just as they normally deliberate about statistical victims, and so the affective system did not engage. The deliberative information conveyed by efficiency rates may similarly cause people to donate less money to charities, even if the efficiency rates are high.

Not only do donors respond differently to affective and deliberative information, but they also seem to become less sensitive to deliberative information when provided with more affective bases for their decisions. In Hsee and Rottenstreich's (2004) study, donation amount was insensitive to number of pandas helped when potential donors were shown either a photo of one panda, or photos of four pandas they could help. When the pandas were instead represented by dots, the more abstract and calculated donation decision became sensitive to the number of pandas helped. Consistent with Small et al.'s (2007) findings that deliberation can suppress affective response, the effect was driven by decreased donation to a single panda represented by a dot, relative to one represented by a photo, with no difference between photos of four pandas or four dots representing pandas. Taken together, these studies suggest that donors not only react differently to deliberative and affective information, but that deliberative, numeric, information is sometimes overridden by more affective responses, and that donor sensitivity to efficiency should be explored in affective contexts as well.

Another perspective on the potential impact of efficiency information is based on general evaluability theory (Hsee, 1996; Hsee & Zhang, 2010; Loewenstien, Blount, & Bazerman, 1999). This theory suggests that evaluations made between simultaneously available alternatives can differ from evaluations of each alternative made separately. Applying evaluability theory to efficiency ratings, sensitivity to efficiency relative to available alternatives, but not to absolute level of efficiency, may reflect the difference between single and joint evaluation, or between evaluating a single charity's efficiency, and choosing from a selection of charities. Participants without a reference point of how efficient a charity should be would use efficiency to choose between available options when presented with a selection, but not base their donation on the actual efficiency value. Such an effect could be apparent in their choice of a charity to give to, and in their choice of how much to give.

In four studies, we explore the impact of efficiency rates on charitable giving. Study 1 explores the impact of relative rates on how people choose between charities, as well as the impact of the absolute rate on decisions about how much to give. That is, would a high efficiency rate make people give more, or would a better relative efficiency just attract more gifts? If choosing how much to give is the more affective component, and selecting a recipient is the more deliberative component, it might be that efficiency rates impact only those latter relative judgments. In Study 2, we expand the range of efficiencies investigated and remove affective information about the charities, in order to examine whether cold facts such as efficiency become more important in the absence of warm, affective appeals, consistent with earlier scale-sensitivity findings. Study 3 and 4 explore whether providing an explicit standard for charity efficiency causes an increase or decrease in donations when the presented charities clearly exceed or fall short of that standard. This might be expected if people care about charity efficiency, but do not know how to interpret the levels, consistent with general evaluability theory.

Study 1

Methods

Participants. One hundred thirty-eight undergraduates completed the study for partial course credit (1 participant excluded for not indicating donation amount; demographic info was not collected in Study 1 but was for all subsequent studies; participants in Study 1 were drawn from the same participant pool as in Studies 2 and 3).

Materials. A cover survey stated, as part of the instructions, that participants had been entered in a lottery to win a 50-dollar gift card in exchange for their participation. The survey then asked participants to evaluate the effectiveness of the advertisements for four charities, using a 7-point Likert scale. The charities included in this study were the American Red Cross (RC), Save the Children (StC), American Heart Association (AHA), and Susan G. Komen For the Cure (SgK). Participants viewed one actual published print advertisement per charity, selected for being predominantly visual and relying on an emotional appeal, rather than laying out a logical justification for donation. Advertisements were selected to contain simple taglines e.g. “a little from you can mean a lot to someone else” (American Red Cross), or “imagine a life without breast cancer” (Susan G. Komen). A table listing basic background information was provided with each advertisement, as context to help evaluate advertisements. The table presented the charities’ mission statements alongside one more column, which differed by condition. In the two experimental conditions, this column provided the efficiency rates of the charities. In the control condition, the year the charity was founded was displayed instead. An explanation of how efficiency rates are calculated was provided to those who saw them: *Efficiency is measured as cents per dollar spent directly on the cause rather than on raising more money. E.g. an efficiency of 48% would mean that 48 cents of every dollar raised went to the cause, with the remaining 52 cents constituting the overhead of* *the charity.* Participants in the high efficiency condition saw efficiency rates for the four charities of 93%, 88%, 84%, and 76% (mean = 85.25%). The low efficiency group was provided with efficiency rates that were 20% lower: 73%, 68%, 64%, and 56% (mean = 65.25%).

The chosen efficiencies were based on a previous survey of 1,007 American adults, which found that they estimate charities to be only 63.7 percent efficient, while simultaneously holding that charities should be 77.6 percent efficient (Grey Matter Research & Consulting, 2008). Thus, the low efficiency group was intended to be in line with typical expectations, and the high efficiency group was intended to exceed the typically desired level of efficiency.

Efficiencies were randomly assigned to charities between participants. The order the charities appeared in the table was randomized as well. In the control condition, the year each charity was founded was randomly assigned, to control for the amount of information presented, but with numbers unlikely to have any consistent impact on donation decisions. A page thanking participants for their advertisement evaluations asked them to consider pledging a portion of the gift card to any or all of the charities presented. The amount pledged would be donated if they won the gift card. The instructions also stated that previous participants had pledged to donate 5-50 dollars of the total amount to create a social norm of giving, without suggesting an appropriate donation amount.

Procedure. Upon arrival, participants were told that they had been recruited to evaluate charity advertisements. After being explained these instructions as a group (up to 4 participants), they were seated at separated computer workstations to complete the survey, which reiterated the gift card details and instructions. Participants were randomly assigned to condition: high efficiency, low efficiency, or control. Participants evaluated the four advertisements, presented in a randomized order, alongside the background information tables. Next, participants saw what appeared to be a concluding screen to the experiment, thanking them for their participation. This same page also asked participants if they would like to donate a portion of their potential winnings to one of the presented charities. The page again displayed the table of background information that contained the efficiency rates for the charities, or, for the control group, the year founded. Participants specified the recipient and amount they wanted to give. When indicating the recipient, participants had the option of selecting “any of the above”. Afterwards, participants answered two free response questions: *How did you pick the amount you chose to donate?* and *How did you pick which charity you would donate to?* to explore whether self-reports reflect how efficiency affects donation. Participants were debriefed as to the true purpose of the study. A rating of charity worthiness and an estimate of average charity efficiency are excluded from analyses. All procedures were approved by an IRB.

Results

In this experiment, we can look at the impact of both the absolute and the relative levels of efficiency on giving. If people only care about relative efficiency, participants would prefer to give to the more efficient of the charities they saw, but the average donation amount would not differ by whether participants were in the lower efficiency, higher efficiency, or control conditions. If people are also sensitive to the actual level of efficiency, then we ought to see markedly more giving when the efficiency rates are high than when low. Comparison to the control condition would indicate, in this case, whether high efficiency increased donations, or low efficiency suppressed it.

These data allow us first to explore whether people are more likely to donate to the higher ranked charities. Second, they let us examine whether the people who pick, for example, the most efficient of the charities presented, choose a bigger gift size than people who decide to give to the least efficient charity. Third, we can see whether people who are offered charities with high overall efficiency give more overall than people who are given ones with low efficiency. Ninety-seven percent of participants chose to make a donation.

First, we looked at the distributions of donations, in the two experimental groups, to the first, second, third, and fourth most efficient charity presented. Participants specifying a recipient charity in the efficiency conditions (14 participants gave money without preference to whom it went) showed a preference for better ranked charities $\chi^2(3, N = 76) = 26.9, p < .001$, Cramer's $V = 0.34$ (47.4%, 22.4%, 23.7%, 6.58% selected ranks 1-4 respectively; see Figure 1). Donation patterns did not differ between the Low Efficiency and High Efficiency conditions, $\chi^2(3, N = 76) = 1.18, p = .76$, Cramer's $V =$ 0.07, suggesting that participants in these two conditions were using the efficiency estimates similarly in guiding donations (48.4%, 22.6%, 19.4%, 9.68% selected ranks 1-4 respectively in the high condition; 46.7%, 22.2%, 26.7%, 4.44% in the low condition; see Figure 1). Participants in the control condition gave to charities irrespectively of assigned charity age $\chi^2(3, N = 40) = 1.01, p =.80$, Cramer's $V = 0.09$ (8 had no preference; 27.75%, 30.0%, 20.0%, 22.5% selected charities ranked 1-4 respectively by age). Thus, more relatively efficient charities were more often selected as the recipients of the donations.

Next we explored whether there was evidence for increased donation amounts to more efficient charities, the second process by which more efficient charities may solicit greater total donations. In the two conditions that saw efficiency information, donation amount did not differ by relative charity efficiency, $F(1, 74) = .003$, $p = .95$, $\eta^2 < .001$, indicating that efficiency rank did not play a significant role in determining amount given (means: rank 1 = $29.50, *SD* = 16.7; rank 2 = $28.94, *SD* = 18.6; rank 3 = $25.83, *SD* = 14.2; rank 4 = $33.00, *SD* = 23.3).

Finally, we examined whether seeing charities that, overall, were considerably more efficient elicited higher levels of donation. Average amount donated did not differ by condition, $F(2, 134) = 1.15$, $p = 0.32$, $\eta^2 = .017$. Participants pledged to donate an average of $27.06 of the 50-dollar gift card to charity: an average of $30.33 in the high efficiency condition (*SD* = 18.1), $26.94 in the lower efficiency condition (*SD* = 16.0), and $24.68 in the control condition (*SD* = 16.8; see Figure 2). The large variances suggest that participants were not simply defaulting to a donation of half of their potential earnings.

While more relatively efficient charities collected more money overall, this was the result of increased donation frequency, but not increased donation amount. Across the two experimental conditions, the first ranked charities were pledged, from the 90 participants, a total of $1065 dollars, while the fourth ranked charities were pledged $165.

The two open-ended self-report questions were coded for any mention of efficiency. While efficiency did not play a role in deciding amount, with a single participant reporting using efficiency information to decide an amount to give, 35.6% of participants reported using efficiency information to select a donation target, $\chi^2(1, N = 180) = 33.4, p <.001$, Cramer's $V = .43$.

Independent of efficiency effects, participants did find the charities differentially appealing, $F(3, 134) = 4.33$, p < .001, $\eta_p^2 = .036$, with Susan G. Komen getting the highest ratings. This charity also received the most donations (36.8%), though the magnitude of its average donation was not larger ($23.21, versus the average donation of $27.06).

Discussion

Donation amounts were not affected by whether low efficiency, high efficiency, or control information was presented. Nonetheless, participants preferred to donate to the best-ranked charity, albeit without adjusting their donation amount by the relative efficiency of the charity they chose. The results of Study 1 demonstrate sensitivity to relative charity rank rather than to absolute efficiency values. Furthermore, participants were deciding donation amount without sensitivity to the scale of efficiency, while choosing a recipient scale-sensitively to relative charity efficiency. This is consistent with amount donated being a more affective process, and selection of charity being a more deliberative process. These results also suggest that general evaluability theory does not entirely account for efficiency insensitivity, because in the joint-evaluation of charity efficiency, participant donation amount remained insensitive to relative efficiency.

In Study 2, we exaggerated low efficiency estimates to explore whether the insensitivity to absolute rates in donation amounts would persist even when the difference was much greater between the two conditions. One might expect the effect of relative rank to be slightly bigger in low efficiency condition in Study 1, as well as the following study, due to the relative difference between two smaller numbers being greater than between two larger numbers with the same absolute difference.

Given the previous findings of differences between affective and deliberative processes in donation, we also explored whether affective information, provided by the charity advertisements, was dampening sensitivity to efficiency. The nature of the selected ads, as is generally the case, was not to provide logical justification for the charity or analysis of its effectiveness, but instead to create an emotional appeal through bold, catchy graphics. The advertisement for Save the Children, for example, shows a portrait of a frontline health worker made from electrocardiograms of the children he saved. Study 2 also compared donation decisions made by participants exposed to advertisements with those made by participants not provided this affective information.

Study 2

Methods

Participants. One hundred two undergraduates completed the second study for partial course credit (81.3% female; *M* age = 20.8, *SD* = 2.40; 74% identified as Asian/Pacific Islander, 15% identified as Hispanic/Latino, 9% as White, and 3% as other; age not specified by 30 participants).

Materials. The high efficiency estimates remained 93%, 88%, 84%, 76% (*M* = 85.25%). Low efficiency estimates were made to be 60% lower (and 40% lower than the low efficiency group in Study 1: 33%, 28%, 24%, 16% (*M* = 25.25%). In the non-affective condition, print advertisements were replaced with a blank page with the name of the charity written in a plain font. Participants could still judge charities by their names and mission statements, but no longer by advertisement. The materials otherwise were unchanged from Study 1. The condition without efficiency information was eliminated.

Procedure. The same procedure was utilized as in Study 1, with the addition that participants were also randomly assigned to either an affective (containing print advertisements) or non-affective (containing blank pages with the names of the charities in place of advertisements) condition.

Design. The design of this study was a 2x2 factorial with efficiency – high (*M* = 85.25%) versus low (*M* = 25.25%), and affect – affective versus non-affective charity information – as the two between-subjects factors.

Results

As in Study 1, we first explored whether people are more likely to donate to the more relatively efficient of the charities they saw. We then tested for whether donation amount differed by relative efficiency. Third, we examined whether being presented charities with high overall efficiency attracted more overall giving than being presented ones with very low efficiency. With the addition of affective and non-affective conditions, we also tested for whether the removal of affective stimuli influenced any of these processes. Ninety-five percent of participants chose to make a donation.

Donors remained sensitive to relative efficiency: participants preferred to give to relatively more efficient charities, $\chi^2(3, N = 77) = 33.2, p <.001$, Cramer's $V = .38$, (excluding 20 selecting no preference; 53.2%, 18.2%, 20.8%, 7.79% selected ranks 1-4 respectively; see Figure 3a). There was no detectable difference in sensitivity to efficiency rank between the low and high conditions $\chi^2(3, N = 77) = 4.15, p =.25$, Cramer's $V = .13$, (40.7%, 22.2%, 22.2%, 14.8% selected ranks 1-4 respectively in the high condition; 60.0%, 16.0%, 20.0%, 4.00% in the low condition). Thus, relatively more efficient charities were again more often selected as the recipient of the donations. Although non-significant, the slightly increased sensitivity to relative efficiency in the lower condition could have been the result of greater relative difference between the lower numbers with the same absolute difference as the larger numbers.

Next, we explored whether this expanded range of efficiency influenced the amount donors were choosing to give. Once more, individual donations did not differ by the relative efficiency of the charity receiving the donation, *F*(1, 75) = .87, *p* = 0.36, η^2 = .01 (means: rank 1 = $25.37, *SD* = 15.2; rank 2 = $26.07, *SD* = 14.6; rank 3 = $17.50, *SD* = 15.2; rank 4 = $26.67, *SD* = 20.2; see Figure 4). Despite participants being sensitive to relative efficiency, and the spread of efficiencies between conditions being extreme, individual donation amounts were no smaller when participants were presented very low efficiencies than when efficiencies were high, *t*(100) = .927, *p* = .36, *d* = .18. Participants donated an average of $25.30 (*SD* = 16.0) in the low condition, and $21.48 (*SD* = 14.7) in the high condition. As before, the large variances in these values suggest that participants were not simply consistently pledging half of their potential earnings.

As a result of more frequent selection, but not increased donation amount, relatively efficient charities again collected more money overall. The first ranked charities were pledged a total of $1040 dollars, while the fourth ranked charities were pledged $160.

Next we explored whether the removal of the affective information, conveyed via print advertisement, influenced participant sensitivity to efficiency. Participants not seeing advertisements were more impacted by efficiency rank than those exposed to advertisements, $\chi^2(3, N = 77) = 7.94, p =.047$, Cramer's $V = .32$ (38.8%, 25.0%, 22.2%, 13.9% selected ranks 1-4 respectively in the affective condition; 65.8%, 12.2%, 19.5%, 2.44% in the non-affective condition; see Figure 3b). Participants in the affective condition did not display a strong preference for any of the advertisements, $F(3, 48) = 2.02, p = .11, \eta_p^2 = .038$, (RC = 4.38, $SE = .23$; AHA = 4.61, $SE = .27$; StC = 4.37, $SE = .17$; SgK = 4.72, $SE = .26$), though participants in the affective condition overall preferred certain charities, regardless of efficiency $\chi^2(3, N = 36) = 7.98, p = .046$, Cramer's $V = .272$, (RC = 30.6%; AHA = 13.9%; StC = 13.9%; SgK = 42.7%), while in the non-affective condition there was no detectable preference for any charity, $\chi^2(1, N = 41) = 3.36, p = .34$, Cramer's $V = .17$, (RC = 14.6%; AHA = 24.4%; StC = 26.9%; SgK = 34.1%), despite subjects still being provided with the name and mission statement of each charity. Average donation amounts did not vary between charities, F(4, 97) = .37, p = .83, $\eta^2 = .014$, (RC = \$22.06, $SD = 14.3$; AHA = \$22.33, SD = 17.0; StC = \$24.19, $SD = 14.6$; SgK = \$25.79, $SD = 16.5$; "any" = \$20.72, $SD = 18.5$).

We then explored whether individual donation amounts became more sensitive to efficiency in the absence of advertisements. Individual donation amounts did not differ by whether participants saw advertisements or not, $t(100) = .66, p = .51, d = .13$ (mean non-affective = \$24.26, $SD = 16.9$; mean affective = \$22.11, $SD = 15.8$). An interaction of affective condition and relative efficiency on individual donation amount, $F(3, 69) = 3.12, p = .03, \eta^2 = .16$, however, indicated that the removal of affective information resulted in participants basing their individual donation amounts on relative efficiency of the charity selected (means: rank 1 =\$21.07, $SD = 14.4$; rank 2 = \$25.00, $SD = 15.4$; rank 3 = \$25.62, $SD = 17.0$; rank 4 = \$22.00, $SD = 18.6$ in the affective condition; rank 1

=\$27.59, *SD* = 15.3; rank 2 = \$28.00, *SD* = 14.4; rank 3 = \$9.38, *SD* = 7.65; rank 4 = \$50 from single donation, in the non-affective condition). However, yet again, participant sensitivity to relative charity efficiency was not any greater among participants seeing low efficiencies than among those seeing high efficiencies, $t(39) = 1.15$, $p = .26$, $d = .36$ (mean selected rank high = 1.83, SD = 1.11, mean selected rank low = 1.48, SD = .78).

The two open-ended self-reports coded for mentions of efficiency showed that efficiency was more often mentioned when deciding recipient (31.4%) than amount (2.94%), $\chi^2(1, N = 204) = 27.0$, $p < .001$, Cramer's $V = .38$. There was no significant difference in mentions of efficiency when deciding recipient between the affective (38% and non-affective (25%) conditions, $\chi^2(1, N = 102) = 1.44$, $p = .23$, Cramer's $V = .14$.

Discussion

Study 2 investigated both the effects of a larger separation in charity efficiency, and of the removal of affective context on donations. When participants were presented with extremely low efficiencies, donation amount was not reduced. However, efficiency relative to available alternatives did influence charity selection. In the high efficiency condition, the 76% efficient charities garnered as few donations as the 16% efficient charities did in the low efficiency condition, as a result of both being presented as the least efficient charity. However, participants did use relative efficiency to compare the charities they were presented with, and more often chose the best-ranked charity.

In the absence of affective stimuli, participants became more sensitive to relative efficiency. However, that one group of charities was on average 25% efficient, while the other was 85% efficient, played no role in determining donation amount, even in this more deliberative decision context. Donation amount did become sensitive to relative efficiency in absence of the print ads, suggesting that more deliberative contexts may cause donors to rethink the amount they donate, but that these effects remain relative rather than absolute. Again, the participants' lack of sensitivity to relative efficiency in the affective condition suggests that donation insensitivity is not simply the result of a difference in joint/single evaluability. However, once advertisements were removed, the donation amounts became sensitive to relative charity efficiency, suggesting that, perhaps only in the absence of more emotional stimuli normally present in donation decisions can general evaluability theory entirely account for the observed patterns of behavior. People choose to donate smaller amounts when choosing less relatively efficient charities in the non-affective condition, although one might think that such gifts would need to be bigger to compensate for the decreased efficiency.

For the next study, we explored the possibility that participants in the previous study were not put off by the very low efficiencies because they were unaware of what efficiency they could reasonably expect from charities: one might only find a 25% efficient charity off-putting if one knew that charities in general are vastly better. By informing participants of average charity efficiency, and showing them charities that either failed to meet or exceeded this standard, we explored whether participants could be induced to care about poor efficiency, and conversely whether there might be an increase in giving when all presented options were more efficient than charities are in general.

Study 3

Methods

Participants. Ninety-one undergraduates completed the third study for partial course credit (67.0% female, mean age = 20.5, *SD* = 1.89; 1% identified as African-

American, 49% as Asian/Pacific Islander, 18% as Hispanic or Latino, 22% as White, 10% as other).

Materials. The materials were unchanged from the affective conditions of Study 2: half of the participants were exposed to the low condition, where charities were 25% efficient on average, and half were exposed to the high condition, where charities were 85% efficient on average. However, now all participants were also exposed to a page that preceded the advertisement evaluations, which stated that the average US charity is 70% efficient, alongside an explanation of charity efficiency. Thus half saw a set of efficiencies that exceeded this average, and half saw a set that fell far short of it.

Procedure. The same procedure was utilized as in Study 2. However, before beginning the advertisement evaluations, participants read a statement, which informed them that the average US charity is 70% efficient, alongside an explanation of efficiency, using the same wording as in the first two studies.

Results

As in Studies 1 and 2, we first explored whether participants more often chose to donate to more relatively efficient options. Next we examined whether there were any differences in donation amount attributable to relative efficiency. Lastly, we investigated whether an explicit standard of efficiency would induce sensitivity to absolute efficiency. Ninety-six percent of participants chose to make a donation.

As in the prior two studies, participants in both conditions showed a preference for relatively more efficient options, $\chi^2(1, N = 77) = 32.83, p < .001$, Cramer's $V = .38$, excluding the 10 people who donated with no preference for the recipient (54.5%, 11.7%, 20.8%, 13.0% selected ranks 1-4, respectively). There was no significant difference between participant sensitivity to efficiency between the low and high conditions; despite the former now being explicitly worse than average, and the later explicitly better than average, $\chi^2(3, N = 77) = 2.97, p =.40$, Cramer's $V = 0.20$, (46.3%, 14.6%, 22.0%, 17.1% selected ranks 1-4 respectively in the high condition; 63.9%, 8.33%, 19.4%, 8.33% in the low condition; see Figure 5). Again, the non-significant trend towards increased sensitivity to relative efficiency in the lower condition could have occurred due to greater relative difference between lower numbers with the same absolute difference as larger numbers.

Next we examined whether failing to meet or exceeding average efficiency affected donation amount itself. Individual donation amount did not differ by whether participants saw high or low efficiency rates, despite having clear information beforehand that the charities they were choosing from either exceeded or failed to meet average charity efficiency, $t(89) = 1.52$ $p = .13$, $d = .32$ (mean high = \$29.58, *SD* = 17.2; mean low = \$24.30, *SD* = 15.8; see Figure 6). Again, large variances in these values suggest that participants were not simply consistently pledging a specific portion of their potential winnings. Individual donation amount did not differ by relative efficiency rank, $F(1, 75) = .38, p = .54, \eta^2 = .005$, (means: rank 1 = \$26.55, *SD* = 15.5; rank 2 = \$26.11, *SD* = 18.5; rank 3 = \$26.88, *SD* = 16.6; rank 4 = \$31.00, *SD* = 17.6) providing further support that again, participants were choosing the amount they wanted to donate independently of efficiency information.

The first ranked charities were pledged a total of \$1115 dollars, while the fourth ranked charities were pledged \$310. As a result of more frequent selection, but not increased donation amount, relatively efficient charities again collected more money overall. Absolute levels of efficiency, despite now being comparable to an explicit standard, did not have a significant effect on overall donation.

The two open-ended self-reports were again coded for mentions of efficiency. While 8.79% reported using efficiency information to decide an amount to give, 39.6% of participants reported using efficiency information to select a donation target, $\chi^2(1, N = 182) = 21.9$, $p < .001$, Cramer's $V = .36$.

Discussion

The results of Study 3 demonstrated that being explicitly informed of how efficient charities are on average did not create an absolute sensitivity to charity efficiency in our study; participants still solely used relative efficiency to select a recipient. Surprisingly, participants gave just as much money to charities that fell well below what they were told was the national average efficiency, as did those exposed to charities that exceeded average efficiency: poor performance relative to the average charity did not dissuade overall donation, and conversely good performance did not boost overall donation. The results of Study 3 suggest that well performing charities have to compete just as hard as underperforming charities do for donors, rather than donors being satisfied with the charity having exceeded a particular level of efficiency. That donation amount remained insensitive to efficiency information despite an explicit reference point provides further evidence that insensitivity to charity efficiency is not entirely the result of an evaluability bias, or lack of reference point regarding what an acceptable efficiency level is. To verify that these results generalize beyond undergraduates, Study 4 replicates Study 3 with a sample of US adults.

Study 4

Methods

Participants. One hundred and twenty-nine US adults were recruited through Amazon's Mechanical Turk (58.1% female, mean age = 34.8, *SD* = 11.6; 8% identified as African-American, 7% as Asian/Pacific Islander, 9% as Hispanic or Latino, 1% as Native American, 76% as White).

Materials. The materials were unchanged from Study 3, other than the addition of a manipulation check.

Procedure. The same procedure was utilized as in Study 3. After rating ads and making a donation decision, participants were asked to recall the charities' average efficiency to ensure that they were accepting stipulated efficiency rates.

Results

Participants recalled the average efficiency of the charities they were presented with as a check of whether they had accepted stipulated efficiency rates. Participants reported different efficiency rates for the low and high conditions, $t(127) = 11.8$, $p < .001$, $d = 2.09$, (mean low = 40.1, *SD* = 23.0, mean high = 78.4, *SD* = 13.6).

As in prior studies we explored whether participants more often chose to donate to more relatively efficient options and whether there were any differences in donation amount attributable to efficiency. Ninety-three percent of participants chose to make a donation.

Participants in both conditions showed a preference for relatively more efficient options, $\chi^2(1, N = 107) = 21.1$, $p < .001$, Cramer's $V = .44$, excluding 14 people donating to "any", (39.3%, 31.8%, 16.8%, 12.2% selected ranks 1-4, respectively). There was again no significant difference in sensitivity to relative efficiency between the low and high conditions, $\chi^2(3, N = 107) = 2.47, p =.48$, Cramer's $V = .15$, (42.4%, 32.2%, 11.9%, 13.6% selected ranks 1-4 respectively in the high condition; 35.4%, 31.3%, 22.9%, 10.4% in the low condition; see Figure 7).

Failing to meet or exceeding average efficiency again did not influence donation amount in our sample of adults, despite clear information being provided that the charities either exceeded or failed to meet average charity efficiency, $t(127) = .97$ $p = .33$, $d = .17$ (mean low = $14.76, *SD* = 13.0, mean high = $17.18, *SD* = 14.9, see figure 8). Individual donation amount again did not differ by relative efficiency rank, $F(1, 106) = 1.71, p = .19, \eta^2 = .016$, providing further evidence that donation amount was being decided independently of efficiency information (means: rank 1 = $15.97, *SD* = 15.5; rank 2 = $15.74, *SD* = 10.9; rank 3 = $12.67, *SD* = 8.55; rank 4 = $25.38, *SD* = 14.4). The two open ended self-reports confirmed that while efficiency played a minimal role in deciding amount to give, (6.20% reporting using efficiency), it played a larger role in picking a target (28.7%), $\chi^2(1, N = 258) = 21.2, p < .001$, Cramer's $V = .30$.

Discussion

The results of Study 4 replicate our previous finding that being providing a standard of average charity efficiency did not create an absolute sensitivity to charity efficiency, generalizing this result to a sample of US adults. Participants nonetheless used relative efficiency to select a recipient. Poor performance relative to an average did not dissuade overall donation, nor did high performance boost overall donation. Competing on efficiency may be a zero-sum game: overall donations do not increase as efficiency grows.

General Discussion

The results of four studies provide evidence that efficiency estimates redirect donations to relatively more efficient charities, but do not increase or decrease donation overall. In all but the most abstracted situation, individual donation amounts were not influenced by relative efficiency either. Our results support the notion that deciding amount to give, and to whom, are distinct processes when donors are selecting between alternatives with varying efficiencies. Our results do not rule out any possible role for absolute efficiency in donation, but instead suggest that, consistent with general evaluability theory, any effect of absolute level of efficiency that may exist is much weaker than that of efficiency relative to available alternatives – even when all of the alternatives are explicitly and dramatically sub-par.

Our findings are also consistent with previous literature distinguishing between calculating, scale-sensitive processes and affective processes that are scale-insensitive (Hsee & Rottenstreich, 2004). Put into this framework, the scale-insensitivity in choosing an amount to donate suggests that this decision was predominantly affective. The sensitivity to scale in the form of relative efficiency when selecting a recipient may suggest that choosing a recipient of one's donation is predominantly deliberative, even in the presence of affective information. Deciding to give, and how much to give, seem not to be based on cold efficiency statistics, in the presence of relatively minor affective stimuli. These statistics, however, do guide the targeting of the donation, and do so to a much greater extent when people are not provided with an affective basis for that aspect of the decision. Further research may explore to what extent these proposed separate components of the donation decision can be independently manipulated.

These studies also demonstrate the importance of incorporating affective stimuli into research on economic decisions that are normally made in the presence of at least some degree of affective information. Had this study been run in the absence of advertisements and the affective information they convey, we would see only the donation patterns from the participants in the non-affective portion of Study 3, and incorrectly assume that donors base the amount they give to a particular charity on its relative efficiency – a finding which does not generalize to donation decisions made in the context of affective cues.

Our results provide some support for the use of efficiency ratings, albeit with strong reservations; while efficiency ratings may direct donations to relatively more efficient charities, this is more the result of framing than a response to absolute level of charity efficiency. Thus, impartial rankings can direct money to more efficient charities. However, outside of impartial rankings, inefficient charities may deceptively boost their appeal by portraying themselves as relatively efficient compared to even-less-efficient charities. Educating potential donors on what efficiency they should expect from charities does not seem to counteract participants relying on available alternatives, and does not decrease competition among well-performing charities. More importantly, our results do not support the use of efficiency ratings to increase overall donation amounts – participants did not donate more money overall when seeing high efficiencies than when exposed to low efficiencies, even when explicitly told that these efficiencies were better than average. The lack of overall benefit supports the arguments of those concerned that the negative consequences of emphasizing nonprofit efficiency outweigh the positives of its use.

Our studies do have several limitations, perhaps most notably that we employed windfall gains, rather than having participants decide whether to donate their own money. Future studies may explore the extent to which the observed effects generalize to donations made with own money, and could also use larger sample sizes to further explore the smaller effects observed in our studies.

"To give away money is an easy matter," Aristotle suggested, "and in any man's power. But to decide to whom to give it and how large and when, and for what purpose and how, is neither in every man's power nor an easy matter" (Williams, 1869). Efficiency rates may generally help decide to whom to give and for what purpose, but how large a gift is indeed a more complex process that seems to be outside the reach of cold statistics.

Chapter 6, in full, is a reprint of the material as it appears in Nonprofit & Voluntary Sector Quarterly, 2018. Ryazanov, Arseny; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this paper.

References

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal,* 100, 464-477.

Barrett, W. P.. (2017, December). The Largest U.S. Charities For 2017. Forbes. Retrieved from https://www.forbes.com/companies/direct-relief-international/

Batson, C. D. (2014). The altruism question: Towards a social-psychological answer. New York, NY: Psychology Press.

Bedsworth, W., Gregory, A. G., & Howard, D. (2008). Nonprofit overhead costs: Breaking the vicious cycle of misleading reporting, unrealistic expectations, and pressure to conform. Boston, MA: Bridgespan Group.

Bowman, W. (2006). Should donors care about overhead costs? Do they care?. *Nonprofit and Voluntary Sector Quarterly*, *35*(2), 288-310.

Callen, J. L. (1994). Money donations, volunteering and organizational efficiency. *Journal of Productivity Analysis*, *5*(3), 215-228.

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives?. *Judgment and decision making*, *9*(4), 303.

Cialdini, R. B. (1991). Altruism or egoism? That is (still) the question. *Psychological Inquiry*, *2*(2), 124-126.

Cnaan, R. A., Jones, K., Dickin, A. & Salomon, M. (2011). Nonprofit watchdogs: Do they serve the average donor? *Nonprofit Management and Leadership*, 21: 381–397. doi: 10.1002/nml.20032

Cunningham, K., & Ricks, M. (2004). Why measure. Nonprofits use metrics to show that they are efficient. But what if donors don't care. *Stanford Social Innovation Review*, *2*(1), 44-51.

Greymatter Research. (2008). Where'd My Money Go? Americans' Perceptions of the Financial Efficiency of Non-Profit Organizations. Phoenix, AZ: Grey Matter Research & Consulting.

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science*, *346*(6209), 632-635.

Gregory, A. G., & Howard, D. (2009). The nonprofit starvation cycle. *Stanford Social Innovation Review, Fall*, 49-53.

Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*(3), 235-257.

Hager, M. A., and Wing, K. (2004). The Quality of Financial Reporting by Nonprofits: Findings and Implications. Washington, DC: Center on Nonprofits and Philanthropy, Urban Institute.

Hundley, K., Taggart, K. (2017, October 2). America's 50 worst charities rake in nearly $1 billion for corporate fundraisers. Tampa Bay Times. Retrieved from http://www.tampabay.com/news/nation/americas-50-worst-charities-rake-in-nearly-1-billion-for-corporate/2339540

Hope Consulting. (2012). Money for Good: The US Market for Impact Investments and Charitable Gifts from Individual Donors and Investors. San Francisco, CA: Hope Consulting.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: a review and theoretical analysis. *Psychological bulletin*, *125*(5), 576.

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: on the affective psychology of value. *Journal of Experimental Psychology: General*, *133*(1), 23.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science*, *5*(4), 343-355.

Krishnan, R., Yetman, M. H., & Yetman, R. J. (2006). Expense misreporting in nonprofit organizations. *The Accounting Review*, *81*(2), 399-420.

Lowell, S., Trelstad, B., & Meehan, B. (2005). The ratings game. *Stanford Social Innovation Review*, *3*, 38-45.

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, *102*(2), 143-153.

Steinberg, R. (1986). Charitable giving as a mixed public/private good: implications for tax policy. *Public Finance Quarterly*, *14*(4), 415-431.

Stork, D., & Woodilla, J. (2008). Nonprofit organizations: An introduction to charity rating sources and cautions in their use. *International Journal of Applied Management and Technology*, *6*(4), 1.

Szper, R., & Prakash, A. (2011). Charity Watchdogs and the Limits of Information-Based Regulation. *Voluntas: International Journal of Voluntary & Nonprofit Organizations*, *22*(1).

The 50 Largest U.S. Charities. (2014, December). *Forbes*.

Williams, R. (1869) *The Nichomacean Ethics of Aristotle*. Page 56. London: Longmans, Green & Co.

Zajonc, R. B. (1984). On the primacy of affect. *The American psychologist*, *39*(2), 117-123.

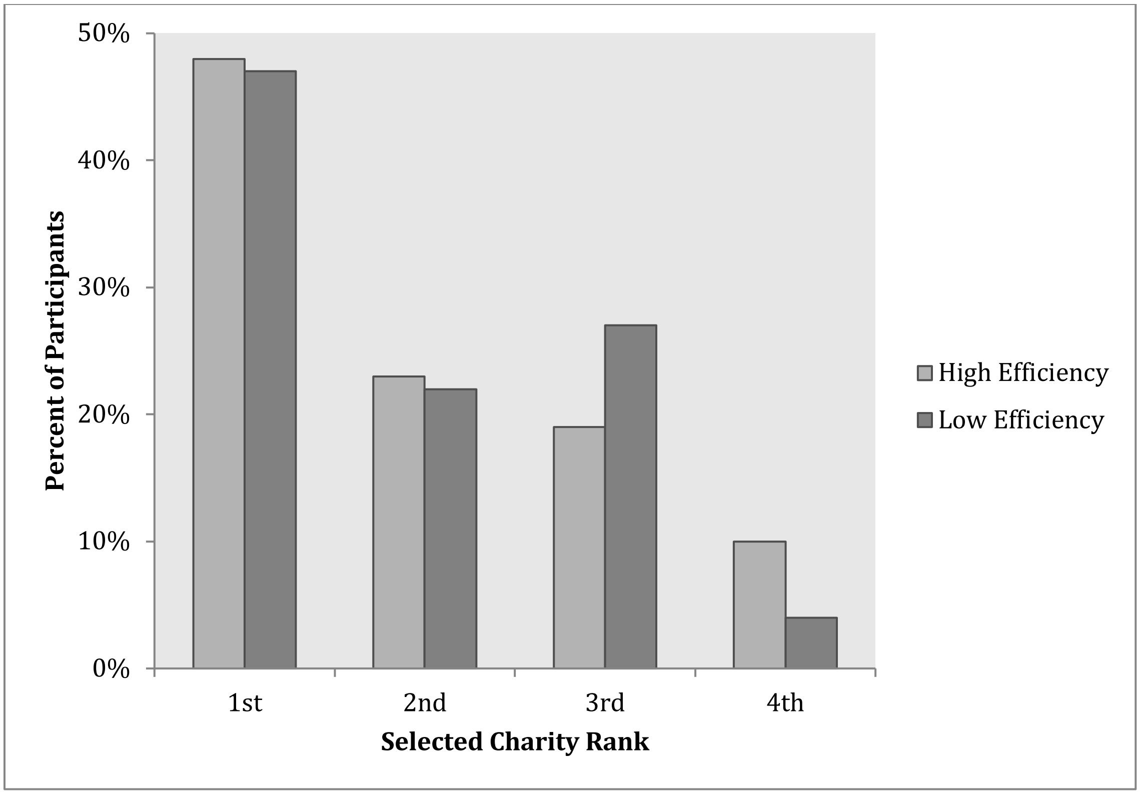


Figure 6.1. Distribution of selected charities' relative efficiencies (Study 1).

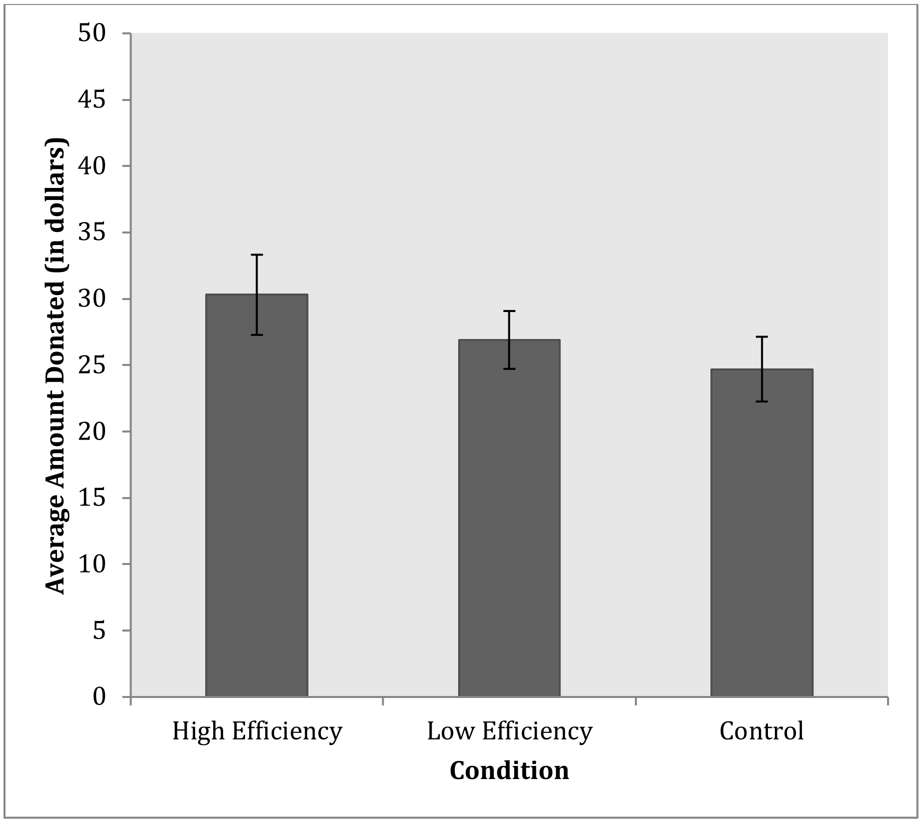


Figure 6.2. Average individual donation by efficiency condition (Study 1). Error bars represent standard errors.

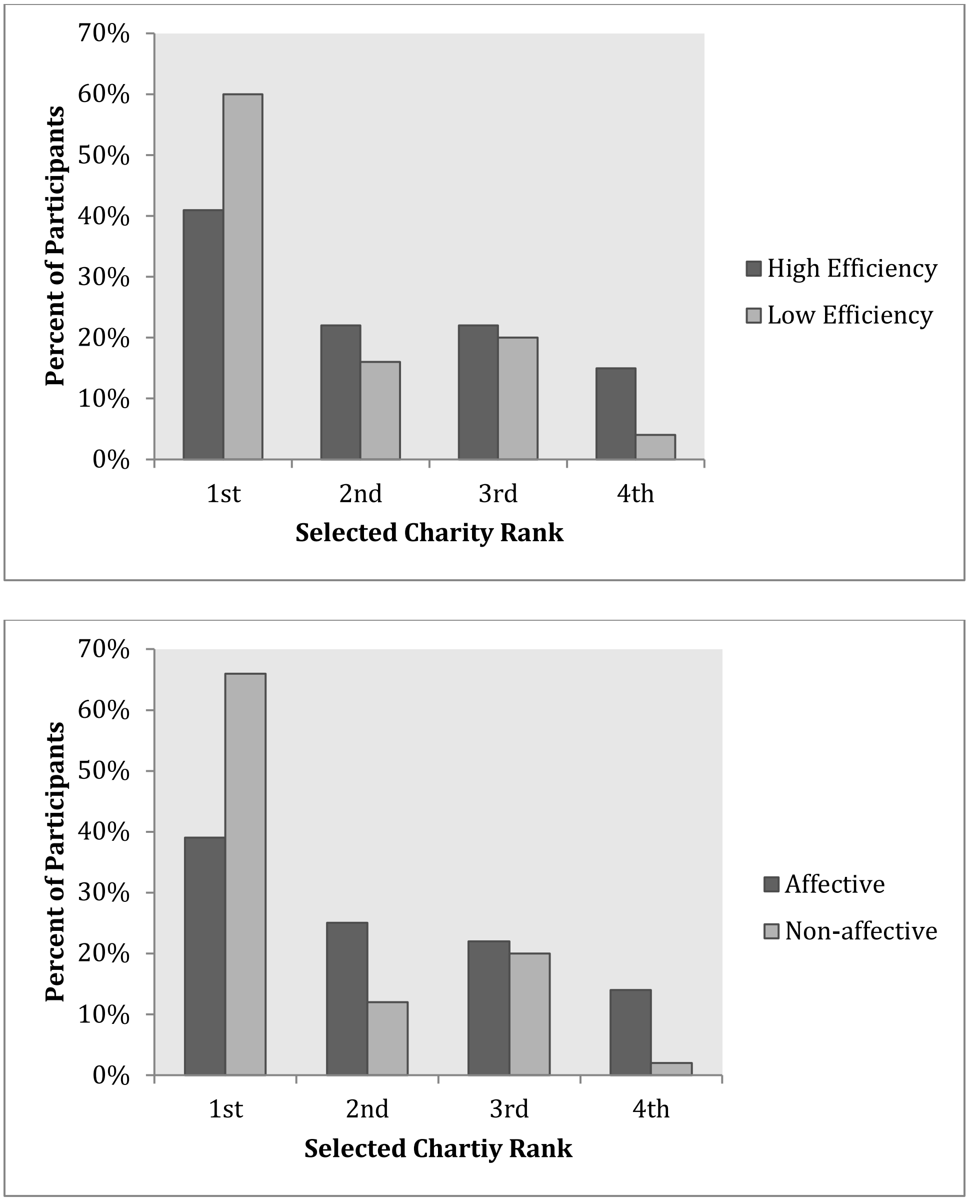


Figure 6.3: Distribution of chosen charities' relative efficiencies (a) over the expanded range of values and (b) between affective conditions (Study 2).

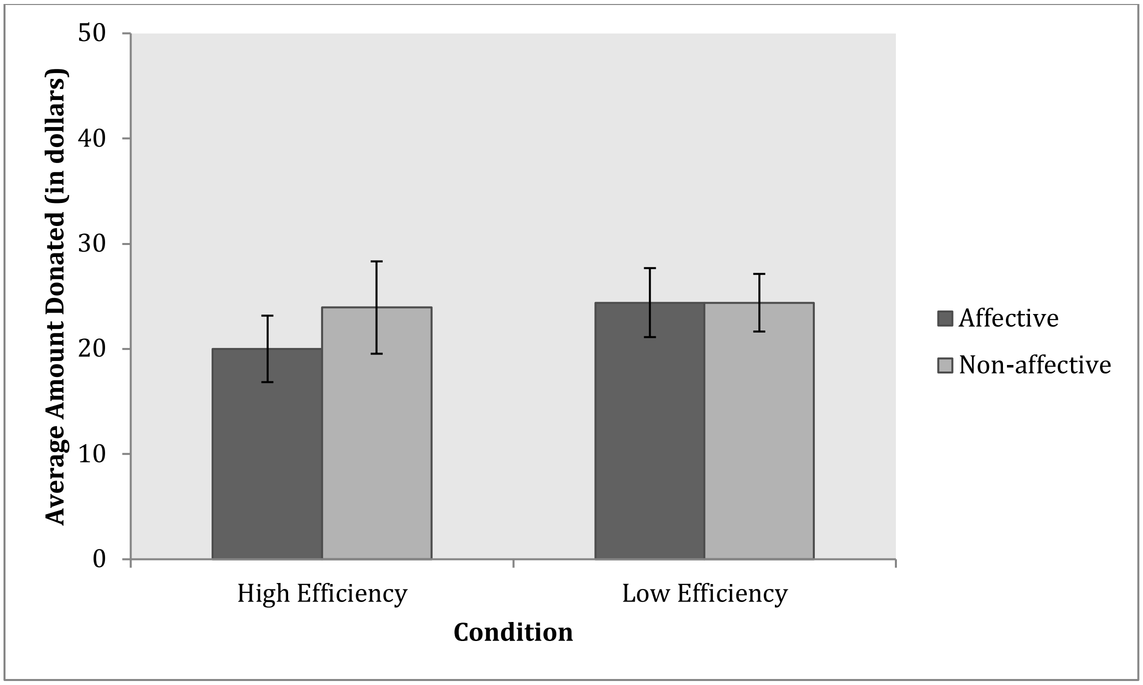


Figure 6.4. Average individual donations across overall efficiency and affective conditions (Study 2). Error bars represent standard errors.

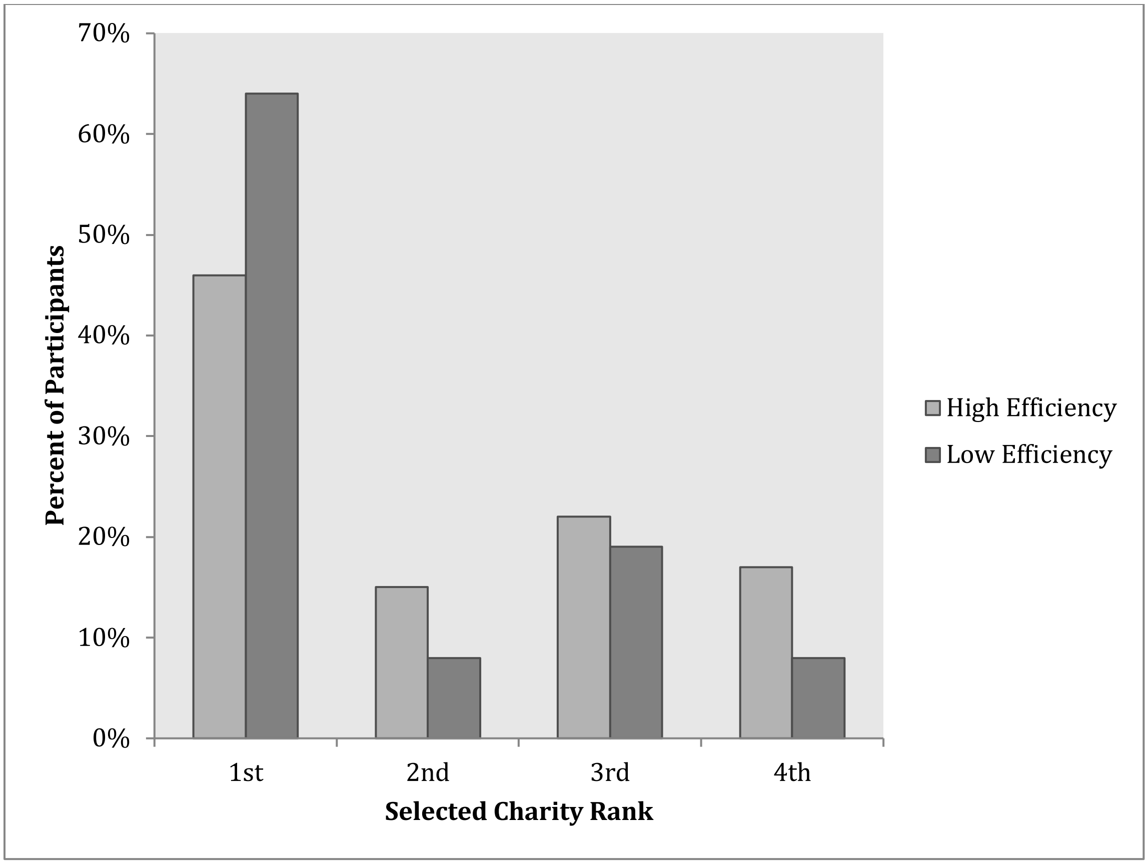


Figure 6.5. Distribution of selected charities' relative efficiencies when participants were informed of an explicit standard of efficiency (Study 3).

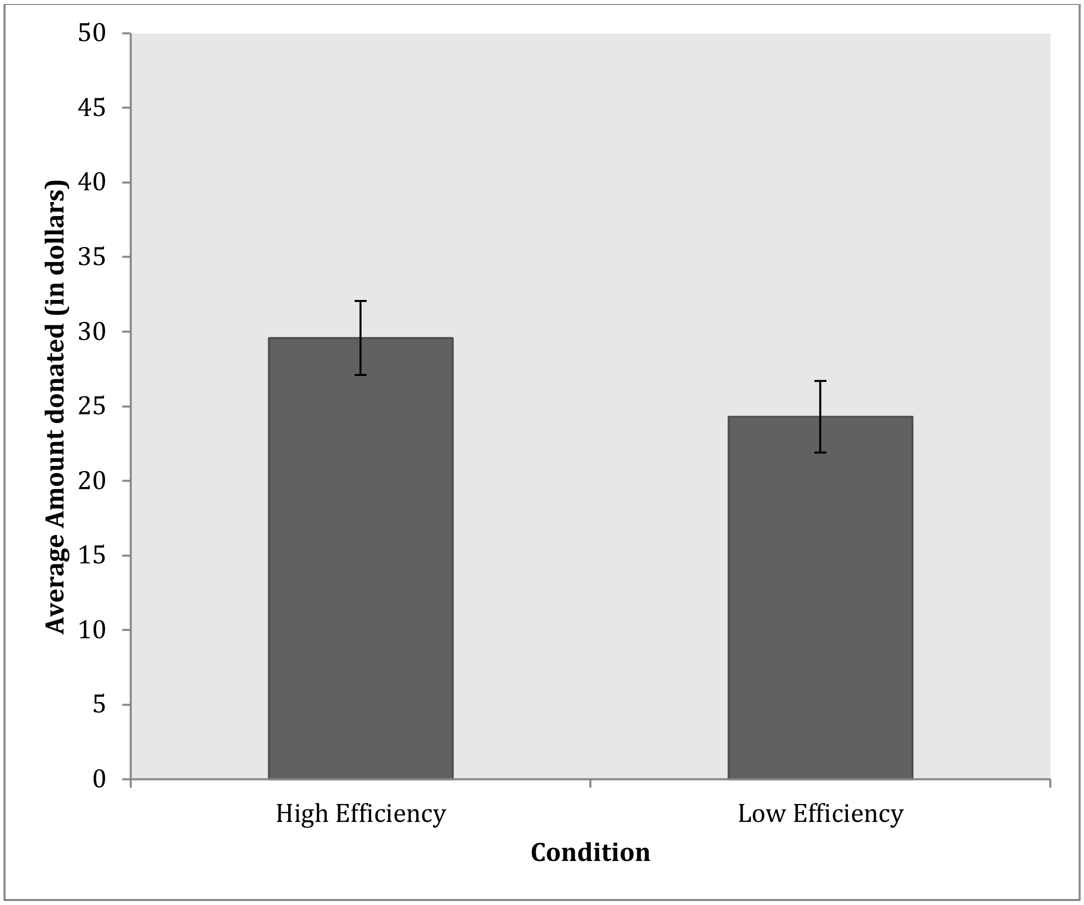


Figure 6.6. Average individual donation amount in high and low conditions when informed of an explicit standard of efficiency (Study 3). Error bars represent standard errors.

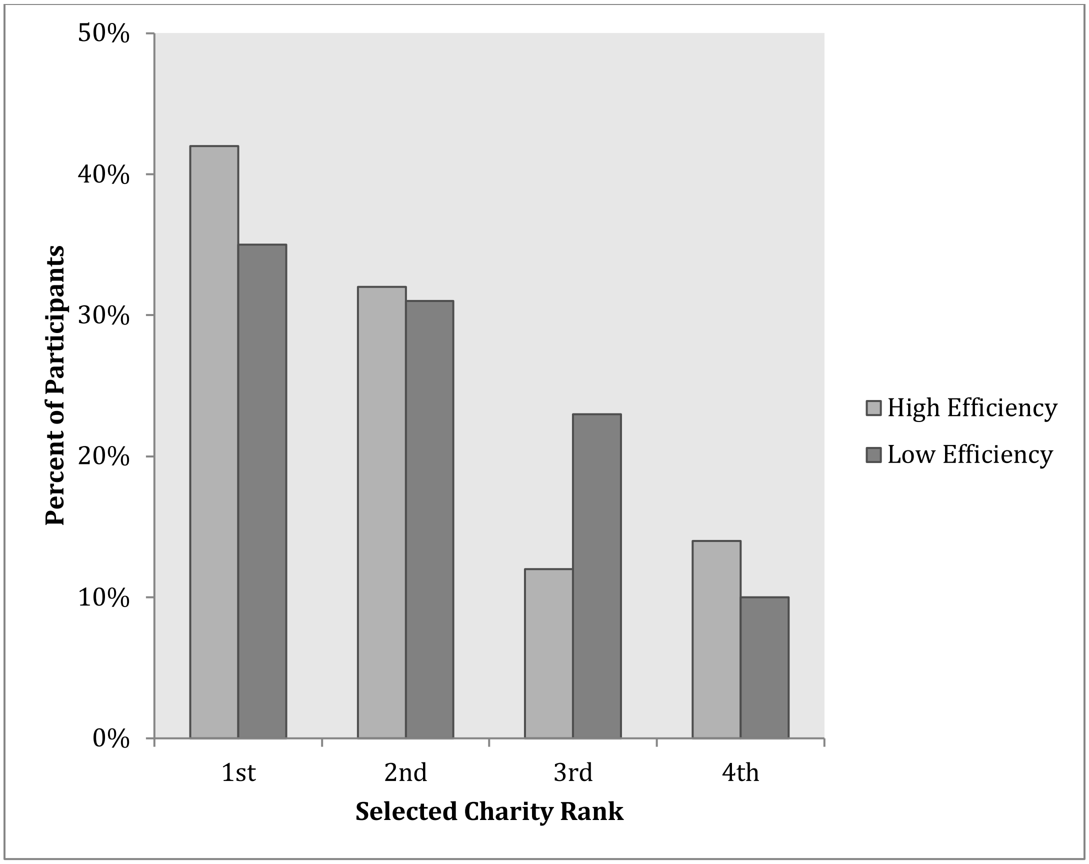


Figure 6.7. Distribution of selected charities' relative efficiencies when US adults were informed of an explicit standard of efficiency (Study 4).

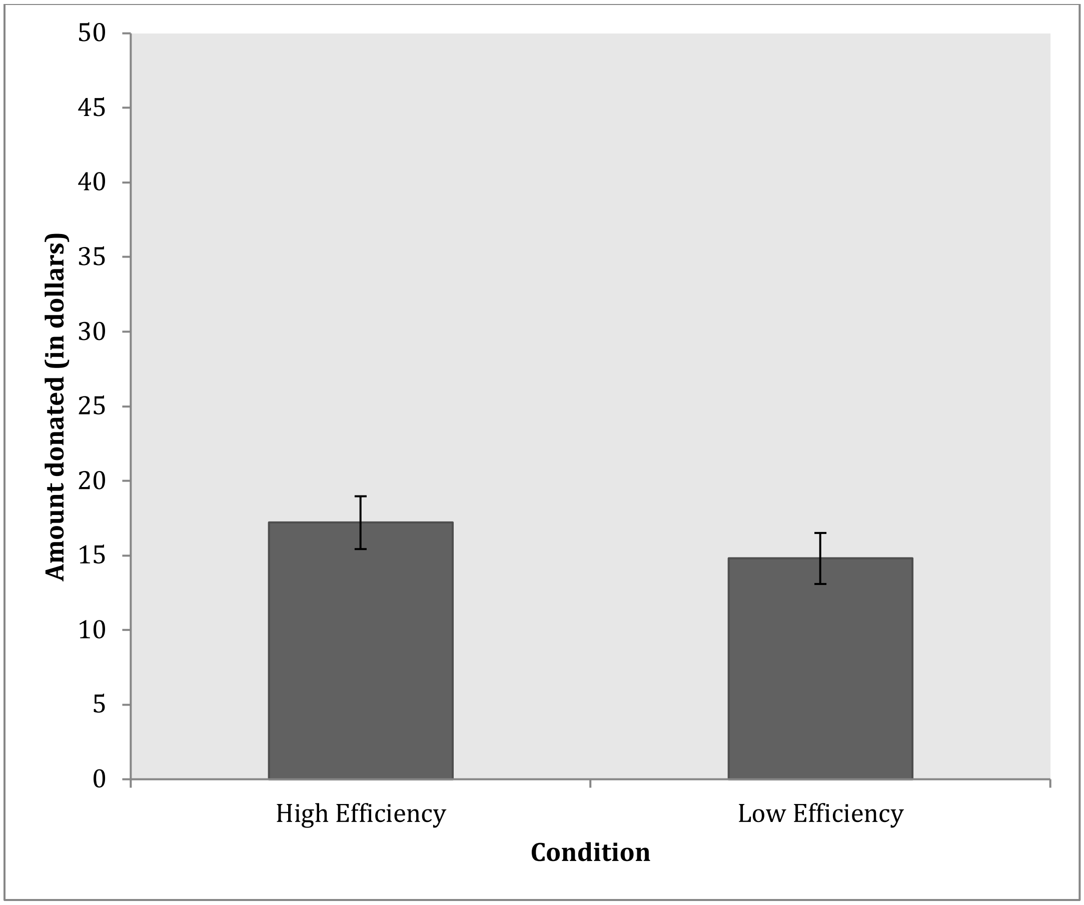


Figure 6.8. Average individual donation by US adult by condition when informed of an explicit standard of efficiency (Study 4). Error bars represent standard errors.

Chapter 7: The Strategic Value of Essentialism

Arseny A. Ryazanov & Nicholas J. S. Christenfeld

University of California, San Diego

Abstract

Psychological essentialism, or the belief that individuals contain an underlying essence that determines their category membership, has generally been regarded by social psychologists as harmful cognitive process that interferes with intergroup harmony at the group level. Fixed mindsets, or the belief that people's trait levels are determined and relatively unchangeable, have been regarded as a parallel impediment to self-improvement at the individual level. However, each of these domains contains findings that do not fit this narrative, suggesting that essentialized thinking is not a necessarily detrimental process at either the group or individual level, and that its effects may instead depend on motivation and context. This reconceptualization suggests shifting efforts beyond working towards broad decrements in essentialist thinking regarding groups and individuals, proposing that instead, in some instances, essentialism can be strategically employed, notably for the absolving of blame and identity formation.

Keywords: essentialism; mindset; morality; identity; true self

The Strategic Value of Essentialism

We are sometimes limited by the actual state of the world, and other times by constraints that exist only in our minds. When these limits are real, acting against them is futile—a short person cannot, try as he may, stretch his way to being tall. These limits can also be psychological—an obese person capable of losing weight may think of himself as unable to do so. Identifying which constraints are physical and which are psychological is a difficult task: how is the obese person to know whether his obesity is a matter of choice or an unalterable condition? Some people, but not all, are able to lose weight—and many more try than are ultimately unsuccessful. Is there a point at which has one has tried hard enough to be justified in resigning oneself to one's condition? This complex problem has been largely sidestepped in social psychology's examination of essentialism, or the perception of underlying immutable essences defining groups and individuals. Essentializing group and individual attributes results in perceiving fixed, unalterable differences between people and groups. As such, essentialism has been cast as an inequity-justifying process at the group level, (why bother funding exercise programs for obese people if they cannot change?) and self-handicapping cognitive process at the individual level (why bother exercising?).

We argue that any anti-essentialist position should also consider the increasing evidence of the benefits of essentialism. We outline parallels between essentialism regarding social groups, such as towards obese people in general, and mindsets at the individual level, such as the extent to which fitness is a perceived as a matter of effort. In reviewing the empirical findings that support the dominant position, that essentialist thinking is harmful, we emphasize a substantial body of research that suggests a more nuanced stance: essentialism is not always bad, and there are in fact many examples of it benefiting individuals and groups. These findings suggest that essentialism can be a strategy for decreasing moral responsibility over uncontrollable aspects of the self or others, and for identity formation.

Yet interventions that adopt a changeable view of oneself and groups show positive effects. We propose that these interventions work not by broadly de-essentializing individuals and groups, but through selectively essentializing the good aspects of oneself and one's group, while de-essentializing the negative.

Psychological Essentialism at Group and Individual Levels

Humans have an innate tendency to form natural categories in both non-social and social domains (Gelman, 2003; Rothbart & Taylor, 1992; Medin & Altran, 2004; Medin & Ortony, 1989). We sense a deep essence within people that accounts for shared characteristics of groups, as well as the identity of those belonging to the group. Essential categories are thought to share an underlying structure, be biologically based, and have well-defined boundaries, in contrast to nonessential groups (Rhodes & Gelman, 2009). Children essentialize innate potential by four years old, finding animals, materials, and social categories, such as gender, to be inductively potent (Heyman & Gelman, 2000a,b,c). Essentialism appears to be at least partially culturally imprinted and acquirable through language (Gelman, Taylor, & Nguyen, 2004; Rhodes, Leslie, & Tworek, 2012).

Two partially overlapping domains of research have emerged regarding essentializing aspects of people. The first, social essentialism, examines how social categories themselves are essentialized, such as race, gender, and sexual orientation. The second, trait essentialism, examines how people's qualities are essentialized, or how fixed one views people's attributes, such as intelligence, to be. Research on trait essentialism has largely been conducted using different terminology, that of implicit person theories, where growth mindsets reflect de-essentialized, malleable conceptions of individuals, and fixed mindsets reflect perceiving traits levels as essential, or unalterable. Our review will shift between these terminologies, but will also step back to draw broader connections between the two fields. Each section covers essentialism in various contexts, outlining initial findings that supported essentialism being harmful, as well as newer research on contexts where essentialism is not, and may in fact be beneficial.

Social Essentialism

Perceiving essences in the social domain, where these essences can serve to differentiate groups of people into categories, has been a contentious issue. Social psychologists generally reject the meaningfulness of group differences, viewing perceived differences as a precursor to racism and other forms of intolerance (Prentice & Miller, 2007; Haslam, Rothschild, & Ernst, 2000; Haslam & Whelan, 2008; Verkuyten, 2003; Hirschfeld, 1996; Yzerbyt, Rocher, & Schadron, 1997). As a result, much effort has been focused on reducing essentialized thinking regarding the groups people belong to.

This anti-essentialist position can be traced back to Allport (1954) theorizing a prejudiced personality emerging as a consequence of a general cognitive style that perceives social categories as rigid and rejects ambiguity in them. Subsequently, the role of perceived essences in intergroup behavior was largely neglected by psychologists until Rothbart and Taylor (1992) proposed that people could interpret social groups as natural kinds, undervaluing environmental and historical contributions to group differences. Rothbart and Taylor (1992) argued that psychological essentialism consisted of viewing social categories as unalterable, and as having inductive potential. By this time, social constructionists had grown to view perceived group differences as false cognitive constructions that supported an unjust status quo (Fuss, 1989; Stein, 1990; Burr, 1995).

However, holding essentialist beliefs about social categories proved to be quite natural: social essentialism was found to emerge early (Medin & Altran, 2004), to exist in every culture studied (Gil-White, 1999; Rhodes, Leslie, & Tworek, 2012), and to exist in both high and low status groups (Leyens et al., 2001). As empirical research on the relationship between social essentialism and prejudice emerged, essentialism was found to only weakly correlate with sexism and racism, and that individuals were inconsistent in the degree to which they essentialized various groups, (Haslam, Rothschild, & Ernst, 2002). Furthermore, the same individuals could essentialize an outgroup in one context, but not another (Morton, Hornsey, & Postmes, 2011).

This cognitive flexibility was demonstrated among minority and majority groups in the Netherlands, both of whom used essentialism as a conversational resource in discussions of multicultural issues (Verkuyten, 2003). While Dutch participants essentialized culture in discourses on how different cultures coexisting is inherently problematic, when discussing assimilation, they adopted de-essentialist arguments. Minority group participants essentialized culture to resist assimilationism, by claiming a right to their identity, but used a de-essentializing discourse when challenging the majority view that their group is negative and homogenous. Essentializing group differences thus appears to be a strategy for justifying potentially conflicting motivations:

identity legitimization and self-determination (Verkuyten, 2003). Identity legitimization decreases personal responsibility for a group's position, whereas self-determination, in opposition, is a resistance to being entirely defined by group membership.

Providing further evidence for essentialism being a motivated process, essentialism was associated with an increased tendency to categorize multiracial individuals as black only in individuals with negativity bias towards the minority group, a distortion in which negative entities weigh more heavily than positive entities (Ho, Roberts, & Gelman, 2015). The relationship's dependence on negativity bias suggests that the role of essentialism depends on the valence of the essence, rather than whether an essence was surmised. When the valence of the essence is positive or neutral, anti-racism efforts rely on essence to emphasize respect for group differences, such as by claiming a right to identity (Taylor, 1994). Despite initial theorizing of essentialism as contributing to a rigid cognitive style, essentialized thinking thus appears flexible and useful for identity formation, Such findings have been paralleled in research on essentializing other social groups.

Essentialism and Gender. Gender is arguably the most essentialized social categorization (Haslam, Rothschild, & Ernst, 2000), and essentialized conceptions of gender were found to contribute to sexism (Bem, 1993; Brescoll & LeFrance, 2004). Subsequent studies, however, have revealed the context-dependence of essentialism in sexism, like in the context of majority and minority group relations. When Morton, Postmes, Haslam, and Hornsey (2009) manipulated the stability of gender inequality, essentialism was associated with increased sexism only among men, and only when inequality between sexes was presented as changing. Subsequently, gender essentialism was found to be endorsed by both genders in response to a system-threat, which motivated participants to uphold their social system (Brescoll, Uhlmann, & Newman, 2013). These studies suggest that essentialism is not necessarily linked to sexism, and that, rather than being a stable cognition, essentialism is motivated by threats to the social order.

Essentialism and sexual orientation. Social essentialism has also been explored in the context of sexual orientation. While some components of essentialist thinking contribute to prejudice against homosexuals, others relate to acceptance (Haslam, Rothschild, & Ernst, 2002; Hegarty & Pratto, 2001). Antigay attitudes were associated with believing that sexual orientation is discrete, fundamental, and an informative category, while tolerance was associated with believing sexual orientation to be biologically based, immutable, and universal.

Here too, essentialism has recently been found to be a flexible resource, rather than a rigid way of thinking: Newman, Bloom, and Knobe's (2014) found that conservatives and liberals selectively essentialize a person's conflicting beliefs and feelings to fit with their own political stance. When a person's homosexual feelings conflicted with his beliefs that one should not act on these feelings, conservatives judged the person's beliefs the more essential part, whereas liberals interpreted his feelings as more essential. When another person's negative feelings toward homosexuals were inconsistent with his beliefs that homosexuality is perfectly acceptable, liberals essentialized his beliefs, and de-essentialized his feelings, while conservatives essentialized his feelings, and de-essentialized his beliefs. Non-prejudiced people appear, thus, to differ in which aspects of the person they essentialize, rather than in extent.

Furthermore, aspects of biological determinism relate to the acceptance of homosexuality by absolving people of personal responsibility for their orientation, similarly to how outgroups rely on essentialism to resist assimilation. Haslam and Levy (2006) caution that biological determinism could be used to medicalize homosexuality or to broaden its acceptance, since sexual orientation goes from being viewed as an immoral choice to an uncontrollable aspect of the self. Essentializing homosexuality legitimizes the social category (Haslam & Levy, 2006), although potentially at the cost of decreased perceived intragroup variability. It is possible that as acceptance of varied sexual orientations becomes increasingly mainstream, the value of essentializing them will diminish, as control and choice will become irrelevant. Such a process may underlie the lack of connection in Britain between immutability of homosexuality and homophobia (Hegarty, 2002).

Essentialism and mental health. Paralleling the role of essentialism regarding other social groups, essentializing psychiatric problems can both absolve sufferers of responsibility for their condition, but also perpetuate pessimism about improvement, resulting in inconsistent effects (Kvaale, Haslam, & Gottdiener, 2013). While biological and genetic explanations decrease acceptance of depressives and schizophrenics, they increase acceptance of alcoholics (Angermeyer, Matschinger, & Schomerus, 2013). Likewise, caretakers attribute less blame to schizophrenics when perceiving less control in the patients' delusions (Provencher & Fincham, 2000). Judges give shorter sentences when given a biomechanistic explanation for a psychopathic-diagnosed convict (Aspinwall, Brown, & Tabery, 2012). Biogenetic explanations may have two competing roles – decreasing personal responsibility at the cost of increasing perceived differentness (i.e. entitativity), providing yet another context in which essentialization decreases moral responsiblity .

Genetic essentialism. Social essentialism has also been explored in the context of genetic determinism, or the extent to which one views genetic contributions to behaviors, individuals, or groups resulting in them being immutable, homogenous, discrete, and natural (Dar-Nimrod & Heine, 2011). Ascribing an outcome to genetics can result in the "naturalistic fallacy", or viewing the current status as good, particularly when evaluating behaviors that may be controllable – such as criminal behaviors or obesity (Dar-Nimrod & Heine, 2011). Viewing obesity as genetic demotivates healthier eating and other potentially-effective lifestyle changes: an induction describing obesity as genetic lead participants to consume more food (Dar-Nimrod, Cheund, Ruby, & Heine, 2014).

While embracing genetic determinism is problematic, undervaluing the actual contributions of genetics and other external causes to behavior, or naïve environmentalism, can likewise be dysfunctional. If one were to adopt the stance that behavior is (almost) entirely controllable, obesity becomes a personally responsibility, regardless of whether one actually has much control over. In discussing trait essentialism we will outline how turning a behavior into a choice increases moral responsibility for the behavior, which can be problematic for behaviors that have biologically or environmentally-determined components. Perhaps a slightly-essentialist view of genetics in obesity could confer the benefits of motivating obese people to attempt losing weight, without obesity becoming a personal responsibility for those unable to do so.

Social essentialism contributes to identity formation

The discourse surrounding social essentialism contains examples of essentialism not only absolving blame for status or behavior, but it also contributing to identity legitimization and formation for minority groups. We now provide a theoretical basis for this process, as well as for its importance. The theoretical basis for how essentialism decreases blame will be described subsequently, in the context of trait essentialism, within which most research on moral responsibility is conducted.

Social identity research suggests that, because individual identity is construed in relation to cultural identity, one cannot form a functional individual identity in the absence of a clear cultural identity (Usborne & Taylor, 2010; Tajfel & Turner, 1986). Interventions focused on decreasing prejudice and stereotyping by rejecting categories may not ultimately help members of marginalized groups, whose underachievement stems from a poorly defined cultural identity (Taylor & de la Sablonnière, 2013). Essentializing a positive group identity instead may thus more effectively decrease achievement gaps between groups.

The need for cultural identity may be encompassed by a more general motivation to seek out social identities that reduce personal uncertainty (Hogg, Adelman, & Blagg, 2010). Personal uncertainty refers to a perceived instability in the self, world, or relationship between the two, and differs from informational uncertainty, which involves understanding that one has incomplete knowledge (Van den Bos, 2009). Uncertainty-identity theory posits that one of the most effective ways to reduce personal uncertainty is through group identification (Hogg, 2007).

When deciding which group to align with, Hogg, Sherman, Dierselhuis, Maitner, and Moffitt (2007) found that individuals prefer to identify with groups that are clearly defined or high in entitativity. As uncertainty increases, people turn to increasingly totalistic groups with comprehensive belief systems, uniform attitudes, and strongly enforced boundaries, such as cults (Baron, Crawley, & Paulina, 2003). Low-entitativity groups make for poor reductions in self-uncertainty because of their vague structure and indistinct boundaries (Hogg, 2007). Blurring boundaries between groups and rejecting the meaningfulness of membership may thus ironically push individuals towards intolerance and even fundamentalism by increasing self-uncertainty, causing people to seek out more extreme entitative groups. A strategic deployment of essentialism, given that people find personal uncertainty aversive, will replace dysfunctional identities with more functional ones and developing positive identities where they do not exist, rather than rejecting underlying essences defining people.

Uncertainty-identity theory also provides a theoretical framework for the findings regarding hostility emerging in social change, consistent with findings on gender essentialism contributing to sexism only when gender roles are perceived as changing. The minimal group paradigm, a classic demonstration of the biasing effects of group membership, fails to increase in-group bias if participant uncertainty is reduced beforehand (Hogg & Grieve, 1999). This suggests that in-group bias emerges not from social categorization itself, but from the uncertain context within which the categorization is made.

Our position does not oppose de-essentializing interventions in all contexts, particularly those with strong negativity biases between stratified groups, for example between Palestinians and Israelis, where describing group membership as malleable increased intergroup cooperation (Goldenberg et al., 2017). Our position does, however, limit the contexts in which such interventions will be successful. Beyond not necessarily helping minority groups, broadly de-essentializing group differences may not necessarily decrease negative attitudes towards these groups: intercultural training has been found to increase cultural essentialism alongside increasing openness to other cultures and cultural intelligence (Fischer, 2011).

Essentializing the good. People are predisposed not only to identify with groups, but also to essentialize most the idealized, or normatively good aspects of categories (Barsalou, 1985; Knobe, Prasada & Newman, 2013; Lynch, Coley & Medin, 2000; Hall, 1998). For example, De Frietas, Tobia, Newman, and Knobe (2017) found that identity of a nation is perceived as more enduring when it is improving by becoming more egalitarian, as opposed to when the nation is becoming more discriminatory. Likewise, Knobe, Prasada and Newman (2013) found that when people poorly fit groups, the normatively good aspects of that person's membership are most essentialized, rather than the aspects that make them a poor fit. People thus seem inclined to think that the essences defining categories are good, and as such should be receptive to replacing dysfunctional identities with positive ones. The impact of identity valence on attributions will be explored through findings that explore this topic in the context of individual trait levels.

Trait Essentialism

Essentialism contributes to perceptions of fixed differences not only between social groups, but between individuals, as well (Haslam, Bastian, Bain, & Kashima, 2006). While social essentialism research has moved from a dispositional conception of essentialism, to a state conception that is used conversationally, trait essentialism is generally explored as a disposition, rather than state, though the stability of essentialism is orthogonal to whether it concerns groups or individuals. Research concerning trait fixedness has been largely conducted through the lens of implicit person theories (Bastian & Haslam, 2006), so our discussion of trait essentialism adopts the terminology of this line of research.

Trait essentialism as implicit person theories. Implicit person theories can be categorized into one of two assumptions about any particular attribute (Dweck, 1999; Dweck & Leggett, 1988). The first is entity theory – that the attribute is a fixed, nonmalleable, trait-like entity, reflecting an essentialized view of the trait. The second is incremental theory – that the attribute is malleable and developable with effort, reflecting a de-essentialized concept of the trait.

Incrementalists focus on factors, such as effort, that could affect trait levels, whereas entity theorists believe trait levels to be relatively stable and unalterable (Dweck, Chiu, & Hong 1995; Dweck & Leggett, 1988). For example, a student holding an incremental theory of intelligence who earns a poor grade works to improve her performance, whereas the entity theorist assumes that the grade reflects an unalterable level of intelligence. While some people have a generalized theory for all attributes, others hold individual theories for different traits (Dweck, Chiu, & Hong 1995).

Implicit person theory research has gradually become unequivocal in its support for fostering incremental mindsets (Wheeler & Omair, 2016). Teaching that the self is malleable is thought to inspire people to tap undiscovered potential. Indeed, targeted interventions improve scholastic achievements (Levy & Dweck, 1999; Blackwell, Trzesniewski, & Dweck, 2007; Paunesku et al., 2015), social relationships (Dweck, 2012;

Kammrath & Dweck, 2006), and health (Yeager et al., 2014). Entity theories have, consequentially, become viewed as maladaptive justification of a self-limited status quo, akin to essentializing rigid social groups.

Essentialism and Moral Responsibility

Essentializing aspects of social groups shifted blame away from members of the group. Likewise, implicit person theories should affect judgments of moral responsibility, which often reflect judgments of controllability (Alicke, 2000; Weiner, 1995; Fincham & Schultz, 1981). Indeed, as in the context of mental health, manipulating an agent's capacity for choice affects blame ascriptions, suggesting that capacity for choice may be how folk theories of morality operationalize free will (Monroe, Dillon, & Malle, 2014). Recent culpability models have increasingly focused on how judgments of moral character, or a persons core, can affect these perceptions of control (Alicke, 2000; Malle, Gugliemo, & Monroe, 2014). Trait essentialism contributes to character judgments, which are critical to understanding blame attribution implicit person theories.

Traditionally, philosophers, other than Aristotle, and psychologists have focused the morality of specific acts, rather than examining how judgments of acts reflect broader evaluations of moral character (Pizarro & Tannanbaum, 2011). Virtue ethics instead suggests that the morality of an act depends not on outcomes or rule compliance, but on how the act reflects character (Pizarro & Tannenbaum, 2011). This better accounts for judgments where, despite two actions breaking the same rule and causing the same outcome, one is nevertheless consistently viewed as worse. For instance, college liberals are less likely to sacrifice a black than a white individual to save a group of people because sacrificing a black individual would reflect a racist character (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009). While morality has generally been examined through actions, a fuller depiction of morality emerges from examining how actions reflect character, or who one is at their core—their essence.

The relationship between implicit person theories and morality has likewise been largely examined via action-based paradigms, though character judgments may better account for attribution patterns. Despite initial excitement that implicit person theories regarding morality could account for inclination towards deontology or consequentialism (Chiu, Dweck, & Tong, 1997), subsequent research showed that incrementalists blame attributions do not neatly fit either act-based model of morality, particularly in moral judgments made across multiple failures. Incrementalists, while more self-forgiving after single failures, were harsher after multiple failures to improve (Niiya, Brook, & Crocker, 2010). These findings may reflect character judgments: incrementalists may excuse a single failure as resulting from external causes, while entity theorists believe the failure reflects an unchangeable trait. However, incrementalists expect themselves to be able to overcome the failure, and thus blame themselves more for continued failure. Incrementalists likewise exhibit greater negative affect than entity theorists do when seeing a person fail to improve after showing effort to do so (Plaks, Grant, & Dweck, 2005).

Growth mindset interventions allow people to reappraise attributes they had previously seen as fixed, or uncontrollable, as changeable. A person who adopts a growth mindset regarding his weight will perceive more control over it and in doing so become increasingly responsible for effecting change. Indeed, a growth mindset (or perhaps more appropriately, shrinkage mindset) regarding obesity can resemble the naïve environmentalism discussed in the context of genetic essentialism. Theories of moral judgment suggest that this expansion of perceived control would necessarily be accompanied by increased moral responsibility, and so extreme incrementalists would be harsher judges of behavior than entity theorists (Plaks, Levy, & Dweck, 2009).

Present implicit person theory research appears to be inconsistent with research relating perceived control to moral responsibility. Entity theorists, despite seeming to ascribe less personal control, make stronger negative moral judgments of people behaving undesirably (Molden & Dweck 2006). Entity theorists judge transgressor moral character as more negative, ignore mediators such as situational constraints, desire more punishment and revenge, and blame the wrongdoer more (Dweck, Hong, & Chiu, 1993; Erdley & Dweck, 1993; Gervey, Chiu, Hong, & Dweck, 1999; Heyman & Dweck, 1998; Yeager, Tirri, Trzesniewski, Nokelainen & Dweck 2011). Incremental theorists have been found to instead focus on reforming the transgressor.

While the incongruity between the moral psychological literature and implicit mindset literature has been noted (e.g. Plaks, Levy, & Dweck, 2009; Dweck & Molden, 2008), there exists little empirical support for any of the explanations offered for the disparity. One is that although entity theorists reject free will, they maintain the illusion of it to keep society functional (Chiu, Dweck, Tong, & Fu, 1997). Another is that determinism is ignored in the moral domain, and a third is that entity theorists think there exist rare situations in which effort triumphs, and hold people accountable for not having been the exception (Dweck & Molden, 2008). These explanations are inconsistent with findings regarding the importance of capacity for choice in blame attributions, and greater determinism resulting in less ascribed moral responsibility (Baumeister & Monroe, 2014).

Examining how implicit person theories relate to conceptions of the true self, or the degree to which a person is thought to be good or bad at their inner core, may better account for the discrepancy between implicit theory findings and free will findings. If incrementalists expect people to be able to improve because people are essentially good, and entity theorists are pessimistic about improvement because they think people are essentially bad, incremental theorist may make more forgiving character evaluations than entity theorists, despite perceiving more control. After exploring how improvability is based in an assumption of essential goodness, we will return to the contradiction that emerges from rejecting innate trait levels while assuming innate potential.

Trait essentialism contributes to identity: the true self. The true self distinguishes a person's deep essence from their more superficial, inauthentic attributes (Newman, De Freitas, & Knobe, 2015), and comprises the perceived innate and immutable characteristics that allow a person to live a meaningful life (Schlegel, Hicks, King, & Arndt, 2011). Thinking we have discoverable innate attributes facilitates psychological well-being by providing a sense of meaning, and seems most closely aligned with entity theories, in that the discoverable traits are perceived as unalterable (Schlegel, Vess, & Arndt, 2012). The alternative and more incremental theory, that we create ourselves, is associated with decreased perceptions of a meaningful life, both among college students and adults. Although the contradiction with implicit person theory findings has been noted by Schlegel and colleagues (2012), it remains unresolved.

This incongruity may be largely overlooked because people essentialize positive qualities more readily than negative ones, like they do for social groups. The asymmetry is early emergent: children expect both psychological and biological negative attributes, even a missing finger, to spontaneously improve over time (Lockhart, Chang & Story, 2002). Adults may have dampened expectations regarding fingers, but likewise essentialize positives more readily than negatives (Haslam, Bastian, & Bissett 2004; Newman, Bloom, & Knobe 2014). They find impulsive negative actions to be less blameworthy than deliberated ones, yet maintain that impulsive positive actions are just as praiseworthy as deliberated ones (Pizarro, Uhlmann, & Salovey, 2003). More essentialized traits are also judged as more important to defining identity, and people work hardest towards enhancing the traits they see as stable and important (Dunning, 1995). As a result, the true self, or core essence, is usually asymmetrically positive.

Essence Valence and Moral Responsibility. Since entity theorists are thought to generally perceive attributes as fixed, the same mechanism that makes a single failure diagnostic for entity theorists should also buttress them against the fears of failure after a single success. The current literature, however, focuses on entity theorists specifically overgeneralizing from negative instances. Wheeler and Omair (2016) have theorized that it may be that the entity mindset selectively essentializes negative qualities as stable, and positive qualities as transient, whereas the growth mindset selectively essentializes positive attributes, while viewing negative ones as transient. If this is the case, then the mechanism by which growth mindset interventions work is not by de-essentializing the person as a whole, but rather by selectively essentializing a positive core, or true self, while de-essentializing negative attributes. Indeed, parents' mindsets specific to failure, rather than intelligence, affect their children's intelligence mindsets (Haimovitz & Dweck, 2016).

Thus, like in the case of social essentialism, it may be that some of the difference between incrementalists and fixed theorists is not the extent to which they essentialize, but rather what they essentialize. Manipulating whether a participant sees an actor's true self as evil or good leads people to interpret the same actions as more or less blameworthy, respectively (Newman, de Freitas, & Knobe, 2015). If true self judgments influence what is externalized and internalized, developing a positive true self could be a strategic intervention that encourages well-being, achievement, and interpersonal cooperation in a way that avoids explicit over-ascriptions of control. For example, writing about how one could make a positive impact on the world leads to similar scholastic improvements to a growth mindset intervention (Paunesku et al., 2015), suggesting that mindset interventions work may implicitly rely on developing a positive true self.

More positive traits being more essentialized also suggests that de-essentializing interventions may work because the most readily abandoned traits are negative ones. If positive traits ultimately can become de-essentialized when one adopts an extreme incremental position, there could be a practical problem of how to stop short of de-essentializing properties that contribute to positive identities. Alternatively, if positive traits are so deeply essentialized that these interventions do not affect them, mindset interventions are actually more selective than currently theorized. Those with negative true selves may benefit from strategically essentializing a positive true self. Others may benefit from growth mindset interventions even if they already have a positive true self by further de-essentializing negative attributes. This hypothesis could be examined by exploring whether decreased confidence in perceived trait levels among incremental theorists is selective to negative inferences.

Goal disengagement. Incrementalism may not always be adaptive, yet the contexts in which malleability is explored are ones in which goal persistence is rewarded, or is itself an operationalization of success, rather than where goal-switching would be most expedient (Wheeler & Omiar, 2016). Thus, incrementalists may be less likely to give up on potentially futile goals. Disengaging from unattainable goals is associated with fewer symptoms of illness and greater well-being, better self-reported health, and more normal diurnal cortisol secretion (Miller & Wrosch, 2007, Wrosch, Miller, Scheier, and Brun de Pontet, 2007). Directly fostering, or essentializing, a positive identity independent of any particular goal may contribute to more functional goal disengagement than an incremental mindset by not implying that success is the result of sheer effort.

Teaching a Properly-Calibrated Psychological Flexibility

In its wide range of forms, at both the group and individual level, essentialism has been construed as a destructive force, with its benefits largely overlooked as exceptions to this general role. Rejecting essentialism outright may strip people and groups of identity, if not made specifically selective towards just negative attributes. It may also create illusions of control over aspects of the self that one has no control over, which could, through continuously failure to live up to perceived capabilities, foster a more negative true self. Some researchers are already calling for a closer inspection of the limitations of growth mindsets and anti-essentialism in the real world (Wheeler & Omair, 2016).

Extending McNulty and Fincham's (2012) notion of a properly calibrated psychological flexibility, future research should look beyond broad de-essentialization, and explore the specific contexts in which essentialization is adaptive. In certain cases, individuals may avoid pursuing fruitless endeavors by acknowledging their limitations in a particular area, and instead concentrating efforts in a domain they are better suited to. Likewise, groups could avoid being held accountable for aspects of themselves that may be beyond their control.

Developing a sensitivity to determining which domains one can and which one cannot control may help maintain a positive identity. And likewise a positive identity may make one more willing to goal switch in the face of insurmountable obstacles. Given the close relationship between self and group identity, it may be reasonable to strategically develop identity at both levels. People with unclear social category memberships will benefit more from a group level essentialization, whereas those with clear memberships, having a foundation in relation to which to construe self-identity, will respond most to self-identity essentialization,

Some interventions already strategically employ the relationship between essentialist thinking and positive identity formation: simply labeling an identity with a noun, such as “helper”, rather than describing its corresponding action “helping” encourages prosocial behavior in children (Bryan, Master, & Walton, 2014). Avoiding negative identities similarly decreases behaviors associated with these identities: asking people to not be “cheaters”, rather than to not “cheat” discouraged cheating more effectively (Bryan, Adams, & Monin, 2013). These studies have not examined how individuals labeled as "helpers" react to failure. Positive initial effects of identity striving on motivation could result in later negative effects in response to failure, particularly if identity depends on performance (Wicklund & Gollwitzer, 1981). Consequentially, it is important to distinguish the flexibility we are advocating for—the cultivation of a positive identity not reliant on any particular goal—from such identity striving.

Value-affirming interventions are one strategy for essentializing a goal-independent positive true self. Walton et al. (2015) tested the efficacy of two value-affirming interventions on increasing the GPAs of female engineering students: emphasizing social-belonging by providing a narrative within which to fit the adversity experienced by females in engineering, and affirmation-training that incorporated aspects of the self into a self-identity. While both raised female engineer GPAs, affirmation-training deepened identification with gender by effectively essentializing gender rather than rejecting gender categories, a process that would generally be expected to negatively affect performance. Essentializing a broad positive female group identity, as detected by increases in self-reported gender identification, allowed these female engineers to be less affected by the narrow threat of the "chilly climate" towards women in engineering. In another approach to positive identity formation, asking participants to think of the good they can contribute to the world by invoking a prosocial, self-transcendent purpose improved high school science and math GPAs (Yeager et al., 2014). Expanding beyond an anti-essentialist stance expands the possible routes to facilitate positive identity.

Towards situated freedom. When people over-essentialize attributes, at the group level or individual level, they dissolve the possibility of improving and transcending their current selves. Conversely, focusing on sheer transcendent will, ignoring the facts of the situation, can become equally problematic—both constitute acting in bad faith (Sartre, 1956, in Kernis & Goldman, 2006). An unrestrained freedom may leave people uncertain of who they are, hopelessly trying to change things they have no control over, and blaming themselves, and others, for failure to change. Optimal functioning emerges from accepting the ontological duality of situated freedom, rather than from organizing one's world as limitless.

Chapter 7, in full, is a reprint of the material as it appears in Social and Personality Psychology Compass, 2018. Ryazanov, Arseny; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this paper.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556.

Allport, G. W. (1954). *The nature of prejudice*. Reading, Mass.: Addison-Wesley, 1954

Angermeyer, M. C., Matschinger, H., & Schomerus, G. (2013). Attitudes towards psychiatric treatment and people with mental illness: changes over two decades. *The British journal of psychiatry: the journal of mental science*, *203*(2), 146-151.

Aspinwall, L. G., Brown, T. R., & Tabery, J. (2012). The double-edged sword: Does biomechanism increase or decrease judges' sentencing of psychopaths?. *Science*, *337*(6096), 846-849.

Baron, R. S., Crawley, K., & Paulina, D. (2003). 13 Aberrations of Power: Leadership in Totalist Groups. *Leadership and power: Identity processes in groups and organizations*, 169.

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, *42*(2), 228-235.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of experimental psychology: learning, memory, and cognition*, *11*(4), 629.

Baumeister, R. F., & Monroe, A. E. (2014). Chapter One-Recent Research on Free Will: Conceptualizations, Beliefs, and Processes. *Advances in Experimental Social Psychology*, *50*, 1-52.

Bem, S. L. (1993). *The lenses of gender: Transforming the debate on sexual inequality*. Yale University Press.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child development*, *78*(1), 246-263.

Brescoll, V., & LaFrance, M. (2004). The correlates and consequences of newspaper reports of research on sex differences. *Psychological Science*,*15*(8), 515-520.

Brescoll, V. L., Uhlmann, E. L., & Newman, G. E. (2013). The effects of system-justifying motives on endorsement of essentialist explanations for gender differences. *Journal of Personality and Social Psychology*, *105*(6), 891.

Bryan, C. J., Adams, G. S., & Monin, B. (2013). When cheating would make you a cheater: Implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General*, *142*(4), 1001.

Bryan, C. J., Master, A., & Walton, G. M. (2014). “Helping” Versus “Being a Helper”: Invoking the Self to Increase Helping in Young Children. *Child development*, *85*(5), 1836-1842.

Burr, V. (1995). *Social constructionism*. New York: Routledge.

Chiu, C. Y., Dweck, C. S., Tong, J. Y. Y., & Fu, J. H. Y. (1997). Implicit theories and conceptions of morality. *Journal of Personality and Social Psychology*, *73*(5), 923.

Dar-Nimrod, I., Cheung, B. Y., Ruby, M. B., & Heine, S. J. (2014). Can merely learning about obesity genes affect eating behavior?. *Appetite*, *81*, 269-276.

Dar-Nimrod, I., & Heine, S. J. (2011). Genetic essentialism: on the deceptive determinism of DNA. *Psychological bulletin*, *137*(5), 800.

De Freitas, J., Tobia, K. P., Newman, G. E., & Knobe, J. (2017). Normative judgments and individual essence. *Cognitive science*, *41*(S3), 382-402.

Dunning, D. (1995). Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives. *Personality and Social Psychology Bulletin*, *21*(12), 1297-1306.

Dweck, Carol S. 1999. *Self Theories: Their Role in Motivation, Personality, and Development*. Hove: Psychology Press.

Dweck, C. S. (2012). Mindsets and human nature: Promoting change in the Middle East, the schoolyard, the racial divide, and willpower. *American Psychologist*, *67*(8), 614.

Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological inquiry*, *6*(4), 267-285.

Dweck, C. S., Hong, Y. Y., & Chiu, C. Y. (1993). Implicit theories individual differences in the likelihood and meaning of dispositional inference. *Personality and Social Psychology Bulletin*, *19*(5), 644-656.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review*, *95*(2), 256.

Dweck, C. S., & Molden, D. C. (2008). 4 Self-Theories: The Construction of Free Will. *Are We Free? Psychology and Free Will*, 44.

Erdley, C. A., & Dweck, C. S. (1993). Children's implicit personality theories as predictors of their social judgments. *Child development*, *64*(3), 863-878.

Fincham, F. D., & Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology*,*20*(2), 113-120.

Fischer, R. (2011). Cross-cultural training effects on cultural essentialism beliefs and cultural intelligence. *International Journal of Intercultural Relations*, *35*(6), 767-775.

Fuss, D., (1989) *Essentially speaking: Feminism, nature and difference.* New York: Routledge.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press, USA.

Gelman, S. A., Taylor, M. G., & Nguyen, S. P. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs: I. Introduction. *Monographs of the society for research in child development*.

Gervey, B. M., Chiu, C. Y., Hong, Y. Y., & Dweck, C. S. (1999). Differential use of person information in decisions about guilt versus innocence: The role of implicit theories. *Personality and Social Psychology Bulletin*, *25*(1), 17-27.

Gil-White, F. J. (1999). How thick is blood? The plot thickens...: If ethnic actors are primordialists, what remains of the circumstantialist/primordialist controversy?. *Ethnic and racial studies*, *22*(5), 789-820.

Goldenberg, A., Endevelt, K., Ran, S., Dweck, C. S., Gross, J. J., & Halperin, E. (2017). Making Intergroup Contact More Fruitful: Enhancing Cooperation Between Palestinian and Jewish-Israeli Adolescents by Fostering Beliefs About Group Malleability. *Social Psychological and Personality Science*, *8*(1), 3-10.

Haimovitz, K., & Dweck, C. S. (2016). What predicts children’s fixed and growth intelligence mind-sets? Not their parents’ views of intelligence but their parents’ views of failure. *Psychological science*, 0956797616639727.

Hall, D. G. (1998). Continuity and the persistence of objects: When the whole is greater than the sum of the parts. *Cognitive Psychology*, *37*(1), 28-59.

Haslam, N., Bastian, B., Bain, P., & Kashima, Y. (2006). Psychological essentialism, implicit theories, and intergroup relations. *Group Processes & Intergroup Relations*, *9*(1), 63-76.

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*,*30*(12), 1661-1673.

Haslam, N., & Levy, S. R. (2006). Essentialist beliefs about homosexuality: Structure and implications for prejudice. *Personality and Social Psychology Bulletin*, *32*(4), 471-485.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, *39*(1), 113-127.

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice?. *British Journal of Social Psychology*, *41*(1), 87-100.

Haslam, N., & Whelan, J. (2008). Human natures: Psychological essentialism in thinking about differences between people. *Social and Personality Psychology Compass*, *2*(3), 1297-1312.

Hegarty, P. (2002). 'It's not a choice, it's the way we're built': Symbolic beliefs about sexual orientation in the US and Britain. *Journal of Community & Applied Social Psychology*, *12*(3), 153-166.

Hegarty, P., & Pratto, F. (2001). Sexual orientation beliefs: Their relationship to anti-gay attitudes and biological determinist arguments. *Journal of Homosexuality*, *41*(1), 121-135.

Hirschfeld, L. A. (1998). *Race in the making: Cognition, culture, and the child's construction of human kinds*. MIT Press.

Heyman, G. D., & Dweck, C. S. (1998). Children's thinking about traits: Implications for judgments of the self and others. *Child development*, *69*(2), 391-403.

Heyman, G. D., & Gelman, S. A. (2000). Preschool children's use of novel predicates to make inductive inferences about people. *Cognitive Development*, *15*(3), 263-280.

Heyman, G. D., & Gelman, S. A. (2000). Preschool children's use of trait labels to make inductive inferences. *Journal of experimental child psychology*, *77*(1), 1-19.

Heyman, G. D., & Gelman, S. A. (2000). Beliefs about the origins of human psychological traits. *Developmental Psychology*, *36*(5), 663.

Ho, A. K., Roberts, S. O., & Gelman, S. A. (2015). Essentialism and racial bias jointly contribute to the categorization of multiracial individuals. *Psychological science*, *26*(10), 1639-1645.

Hogg, M. A. (2007). Uncertainty–identity theory. *Advances in experimental social psychology*, *39*, 69-126.

Hogg, M. A., Adelman, J. R., & Blagg, R. D. (2010). Religion in the face of uncertainty: An uncertainty-identity theory account of religiousness. *Personality and Social Psychology Review*, *14*(1), 72-83.

Hogg, M. A., & Grieve, P. (1999). Social identity theory and the crisis of confidence in social psychology: A commentary, and some research on uncertainty reduction. *Asian Journal of Social Psychology*, *2*(1), 79-93.

Hogg, M. A., Sherman, D. K., Dierselhuis, J., Maitner, A. T., & Moffitt, G. (2007). Uncertainty, entitativity, and group identification. *Journal of experimental social psychology*, *43*(1), 135-142.

Kammrath, L. K., & Dweck, C. (2006). Voicing conflict: Preferred conflict strategies among incremental and entity theorists. *Personality and Social Psychology Bulletin*, *32*(11), 1497-1508.

Kernis, M. H., & Goldman, B. M. (2006). A multicomponent conceptualization of authenticity: Theory and research. *Advances in experimental social psychology*, *38*, 283-357.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242-257.

Kvaale, E. P., Haslam, N., & Gottdiener, W. H. (2013). The 'side effects' of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. *Clinical Psychology Review*, *33*(6), 782-794.

Levy, S. R., & Dweck, C. S. (1999). The impact of children's static versus dynamic conceptions of people on stereotype formation. *Child Development*,*70*(5), 1163-1180.

Leyens, J. P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M. P., Vaes, J., & Demoulin, S. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology*, *31*(4), 395-411.

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of traits: Protective optimism?. *Child Development*, *73*(5), 1408-1430.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition*, *28*(1), 41-50.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, *40*(2), 203-216.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive science*, *39*(1), 96-125.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

McNulty, J. K., & Fincham, F. D. (2012). Beyond positive psychology? Toward a contextual view of psychological processes and well-being. *American Psychologist*, *67*(2), 101.

Medin, D. L., & Atran, S. (2004). The native mind: biological categorization and reasoning in development and across cultures. *Psychological review*,*111*(4), 960.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. *Similarity and analogical reasoning*, *179*, 195.

Miller, G. E., & Wrosch, C. (2007). You've Gotta Know When to Fold'Em Goal Disengagement and Systemic Inflammation in Adolescence. *Psychological Science*, *18*(9), 773-777.

Molden, D. C., & Dweck, C. S. (2006). Finding" meaning" in psychology: a lay theories approach to self-regulation, social perception, and social development. *American Psychologist*, *61*(3), 192.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and cognition*, *27*, 100-108.

Morton, T. A., Hornsey, M. J., & Postmes, T. (2009). Shifting ground: The variable use of essentialism in contexts of inclusion and exclusion. *British Journal of Social Psychology*, *48*(1), 35-59.

Morton, T. A., Postmes, T., Haslam, S. A., & Hornsey, M. J. (2009). Theorizing gender in the face of social change: Is there anything essential about essentialism?. *Journal of Personality and Social Psychology*, *96*(3), 653.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96-125.

Niiya, Y., Brook, A. T., & Crocker, J. (2010). Contingent self-worth and self-handicapping: Do incremental theorists protect self-esteem?. *Self and Identity*, *9*(3), 276-297.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological science*, 0956797615571017.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. *The social psychology of morality: Exploring the causes of good and evil*, 91-108.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise the role of perceived metadesires. *Psychological Science*, *14*(3), 267-272.

Plaks, J. E., Grant, H., & Dweck, C. S. (2005). Violations of implicit theories and the sense of prediction and control: implications for motivated person perception. *Journal of personality and social psychology*, *88*(2), 245.

Plaks, J. E., Levy, S. R., & Dweck, C. S. (2009). Lay theories of personality: Cornerstones of meaning in social cognition. *Social and Personality Psychology Compass*, *3*(6), 1069-1081.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. *Current directions in psychological science*, *16*(4), 202-206.

Provencher, H. L., & Fincham, F. D. (2000). Attributions of causality, responsibility and blame for positive and negative symptom behaviours in caregivers of persons with schizophrenia. *Psychological medicine*, *30*(4), 899-910.

Rhodes, M., & Gelman, S. A. (2009). A developmental examination of the conceptual structure of animal, artifact, and human social categories across two cultural contexts. *Cognitive psychology*, 59(3), 244-274.

Rhodes, M., Leslie, S. J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*,*109*(34), 13526-13531.

Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? In G. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 11-36). London: Sage.

Sartre, J. P. (2004). From being and nothingness. In G. Marino (Ed.), *Basic writings of existentialism* (pp. 369–409). New York: The Modern Library. (Excerpts from Jean-Paul Sartre, Being and Nothingness, H. Barnes, Trans., 1956, New York: Philosophical Library).

Schlegel, R. J., Hicks, J. A., King, L. A., & Arndt, J. (2011). Feeling like you know who you are: Perceived true self-knowledge and meaning in life. *Personality and Social Psychology Bulletin*, 0146167211400424.

Schlegel, R. J., Vess, M., & Arndt, J. (2012). To discover or to create: Metaphors and the true self. *Journal of personality*, *80*(4), 969-993.

Stein, E. (Ed.). (1990). *Forms of desire: Sexual orientation and the social constructionist controversy*. Psychology Press.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In. Worshel, S.; Austin, W.(Eds.) *The psychology of intergroup relations*. Chicago: Nelson-Hall, 7-24.

Taylor, C. (1994). *Multiculturalism*. Princeton University Press.

Taylor, D. M., & De La Sablonnière, R. (2013). Why interventions in dysfunctional communities fail: The need for a truly collective approach. *Canadian Psychology/Psychologie canadienne*, *54*(1), 22.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*(6), 479.

Usborne, E., & Taylor, D. M. (2010). The role of cultural identity clarity for self-concept clarity, self-esteem, and subjective well-being. *Personality and Social Psychology Bulletin*, *36*(7), 883-897.

Van den Bos, K. (2009). Making sense of life: The existential self trying to deal with personal uncertainty. *Psychological Inquiry*, *20*(4), 197-217.

Verkuyten, M. (2003). Discourses about ethnic group (de-) essentialism: Oppressive and progressive aspects. *British Journal of Social Psychology*,*42*(3), 371-391.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology*, *107*(2), 468.

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

Wheeler, S. C., & Omair, A. (2016). Potential growth areas for implicit theories research. *Journal of Consumer Psychology*, *26*(1), 137-141.

Wicklund, R. A., & Gollwitzer, P. M. (1981). Symbolic self-completion, attempted influence, and self-deprecation. *Basic and Applied Social Psychology*, *2*(2), 89-114.

Wrosch, C., Miller, G. E., Scheier, M. F., & De Pontet, S. B. (2007). Giving up on unattainable goals: Benefits for health?. *Personality and Social Psychology Bulletin*, *33*(2), 251-265.

Yeager, D. S., Henderson, M. D., Paunesku, D., Walton, G. M., D'Mello, S., Spitzer, B. J., & Duckworth, A. L. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. *Journal of personality and social psychology*, *107*(4), 559.

Yeager, D. S., Johnson, R., Spitzer, B. J., Trzesniewski, K. H., Powers, J., & Dweck, C. S. (2014). The far-reaching effects of believing people can change: Implicit theories of personality shape stress, health, and achievement during adolescence. *Journal of Personality and Social Psychology*, *106*(6), 867.

Yeager, D. S., Trzesniewski, K. H., Tirri, K., Nokelainen, P., & Dweck, C. S. (2011). Adolescents' implicit theories predict desire for vengeance after peer conflicts: correlational and experimental evidence. *Developmental psychology*, *47*(4), 1090.

Yzerbyt, V. Y., Rocher, S., & Schadron, G. (1997) Stereotypes as explanations: A subjective essentialistic view of group perception. In *The Social Psychology of Stereotyping and Group Life*, Spears, R., Oakes, PJ., Ellemers, N., Haslam, S. A. Blackwell: Cambridge; 20-50.

Chapter 8: Incremental Mindsets and the Reduced Forgiveness of Chronic Failures

Arseny A. Ryazanov & Nicholas J. S. Christenfeld

Department of Psychology

University of California, San Diego

Correspondence concerning this article should be addressed to Arseny Ryazanov, Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109

Contact: aryazano@ucsd.edu

Abstract

Holding an incremental, rather than fixed, mindset confers wide-ranging benefits. Such benefits may, however, be accompanied by increased judgmental harshness of others' shortcomings. Across 3 studies (Studies 1, 2a, 2b; *N* = 416), after an induction of either an entity or incremental view of empathy, aggression, or motivation, participants were asked to imagine someone continually failing to show, or showing in abundance, the particular trait, and were then asked how blameworthy/praiseworthy each of these individuals was. Incremental-induced participants blamed a person showing consistently maladaptive levels of the trait more than did entity-induced participants. Increased blame was mediated by increased perceived control over behavior. Study 3 (*N* = 107) extended findings regarding lay theories of empathy to protagonists in short narratives. Study 4 (*N* = 184) attempted to reconcile our findings with previous research, showing that increased blame attribution by incremental theorists occurs for continual, but not single failures. Overall results suggest that the benefits of an incremental mindset may be partially offset by greater judgmental harshness of others.

Keywords: mindset, lay theories, morality, blame, empathy, aggression, motivation.

Incremental Mindsets and the Reduced Forgiveness of Chronic Failures

Messages of practically unlimited individual potential are ubiquitous, reinforcing the notion that with sufficient hours of practice, effort, desire, or even through sheer willpower, people can improve just about any aspect of themselves—whether they intend to lose weight, gain intelligence, or curb aggressive outbursts. There is a lay theory implicit in such messages—that we are not fixed, but changeable and improvable through persistent effort. But could such messages also, by emphasizing the efficacy of effort, convey that undesirable trait levels are a personal failing?

Implicit person theories capture the ways in which people organize and interpret their own and others' abilities. These theories are often categorized into one of two competing assumptions about a given attribute: an entity theory holds the attribute to be a fixed, nonmalleable, trait-like entity, while an incremental theory holds that the attribute is malleable and can be developed with effort (Dweck, 1999; Dweck & Leggett, 1988). A student with an entity theory of intelligence, for example, believes that she has a certain level of intelligence, high or low, and that there is little she can do to change it. A student with an incremental view of intelligence, on the other hand, believes that intelligence can be improved, for example through extra time spent studying. Because incrementalists view trait levels as changeable, they emphasize the behavioral and psychological mediators of traits, such as effort and situational constraints, rather than the underlying levels of the traits themselves (Dweck, Chiu, & Hong 1995). Incrementalists are more likely, then, to work to improve levels of the trait than entity theorists are. Indeed interventions that induce incremental mindsets have been shown to result in academic improvements (e.g., Yeager et al., 2016), increased willpower (Job, Dweck, & Walton, 2010), reduced aggression (Yeager, Trzesniewski, & Dweck, 2013; Yeager, Trzesniewski, Tirri, Nokelainen, & Dweck, 2011), and increased empathy (Schuman, Zaki, & Dweck, 2014), among other improvements (see Dweck, 2012 for overview).

Such incremental mindset interventions are often designed to foster incremental mindsets towards specific traits. For example, an intervention may induce a growth (i.e. incremental) mindset regarding empathy by describing empathic behavior as the result of deliberate effort and thus empathy being improvable, rather than fixed and unchangeable (Schuman, Zaki, & Dweck, 2014). Exposing participants to an incremental view of empathy causes them to subsequently expend more empathic effort towards others (Schuman, Zaki, & Dweck, 2014). Whether this increased perceived controllability is accompanied by an increased expectation regarding other’s levels of empathy remains to be explored. One’s judgments of another’s unempathic behavior may depend on whether one believes people have control over their level empathy in the first place.

Because mindsets have implications for people’s perceived capacity for change, generally improvement, they are likely to be connected to how people assign blame for shortcomings. Theories of moral responsibility would predict that because incremental inductions ascribe more control over traits and actions, they should result in incrementalists being harsher judges of poor behavior than entity theorists (Plaks, Levy, & Dweck, 2009; Molden & Dweck, 2008). According to these theories, a key component to moral judgment is whether the actor could have, or should have known to do otherwise (Pizarro & Tannenbaum, 2011). For example, Alicke’s (2000) Culpable Control Model suggests that extent of personal control is the primary factor in ascribing blame. The Path Model elaborates upon this position, proposing that judgments proceed through stages, from control to morality (Malle, Gugliemo, & Monroe, 2014). Indeed, manipulating capacity for choice affects blame attributed by participants (Monroe, Dillon, & Malle, 2014).

The predictions of theories of moral responsibility seem to contradict findings regarding implicit person theories and moral judgment, where, despite perceiving increased control, incrementalists are nevertheless found to be more forgiving (Plaks, Levy, & Dweck, 2009; Molden & Dweck, 2006). Children with fixed theories of personality showed less empathy toward, and recommended more punishment for a new student behaving badly, and emphasized what a behavior revealed about a person's good or bad character, while incrementalists focused instead on mediating factors (Erdley & Dweck, 1993; Heyman & Dweck, 1998). Among adolescents, entity theorists desired more revenge than incremental theorists did, and exposure to an incremental induction reduced this desire (Yeager, Tirri, Trzesniewski, Nokelainen & Dweck, 2011). College students who were entity theorists regarding morality made more dispositional attributions for social transgressions and experienced greater negative affect in response to these transgressions than incremental theorists did (Miller, Burgoon, & Hall, 2007). However, the generalizability of these studies is limited in two ways. Firstly, these studies examine moral judgment in the context of global personality theories or lay theories of morality itself. Moral judgment could also, independently of these more global lay theories, depend on the theory of the specific trait along which one evaluates another's behavior. Secondly, and perhaps more importantly, these studies typically examine singular transgressions, rather than continual patterns of behavior.

Though research on implicit person theories' impact on blame attribution for behaviors across longer time spans is sparse, initial evidence suggests that incrementalists can be less forgiving than entity theorists when dealing with continual failure. Incrementalists become harsher towards themselves than entity theorists do in the face of multiple failures to improve (Molden & Dweck, 2006). When a difficult continuous task was tied to self-esteem, incremental theorists who continually performed poorly reported lower self-esteem than did entity theorists (Niiya, Brook, & Crocker, 2010). Repeated failure to improve by others, despite effort to do so, likewise resulted in greater anxiety among incremental theorists than entity theorists (Plaks, Grant, & Dweck, 2005). Since both incrementalists and entity theorists fit theory-violating information to their worldview, instead of adjusting it (Plaks, Grant, & Dweck, 2005; Plaks & Stecher, 2007; Xu & Plaks, 2015), it may be that incremental theorists are unable to reappraise an attribute as relatively uncontrollable when observing continual failure.

The entity theorists' view of limited potential for improvement may be accompanied by an acceptance of their own or another's limitations, be they real or imagined. Kammrath and Dweck (2005) found those with an entity theory regarding personality were more accepting of the faults of a dating partner following relationship transgressions, although at the cost of not working towards making changes that could improve the relationship. Subsequent research showed that incrementalist romantic partners, though initially more optimistic about their partners ability to change negative behaviors, were more likely to attribute failure to lack of effort and were more distrustful of partners exhibiting partial success at changing over a two-week period (Kammrath & Peetz, 2012).

The divergence between blame for singular transgressions and continual failures may be reconciled by an approach to moral psychology and philosophy known as virtue ethics, which suggests that judgment of specific acts can reflect what the acts reveal about the actor's character (Uhlmann, Pizarro, & Diermeier, 2015; Ryazanov & Christenfeld, 2018). It may be that, when evaluating transgressions, incrementalists are more likely to assume a positive character that is capable of improvement, while entity theorists see the action as more diagnostic of character. Indeed, entity theorists have been shown to infer traits from singular actions more readily than incremental theorists (e.g., Dweck, Hong, & Chiu, 1993; Miller, Burgoon, & Hall, 2007). If increased perceived control is accompanied by expectations of improvement across broader patterns of behavior, continual transgressions may provide evidence against an assumed positive character for incrementalists. For incrementalists, a single bad act might not be judged negatively, since it does not reflect being a bad person, and can readily be changed. However, a series of bad acts, betraying a failure to improve, might be judged more harshly. Manipulating whether participants see an actor's character as evil or good leads them to interpret the same action as more or less blameworthy, respectively (Newman, de Freitas, & Knobe 2015). For entity theorists, a single bad act and a series of such acts might both signal a bad person, albeit one with limited control and, thus, responsibility.

Alternatively, it could be that the relationship between mindset and blame is not mediated by perceived control. In the study of prejudice, changing the perceived controllability of stigmatized characteristics does not reduce prejudice regarding those characteristics, because here controllability justifies attitudes, rather than causing them (Hegarty, 2008). The previously discussed Culpable Control Model likewise acknowledges that sometimes control justifies negative attitudes towards an actor rather than causing them (Alicke, 2000). When participants were exposed to a car crash in inclement weather in which the driver was rushing home to hide cocaine, as opposed to an anniversary present, participants rated the former driver as having more control over the car crash (Alicke, 1992). If controllability judgments do not precede blame attributions in the context of implicit person theories, we would expect to see no clear relationship between controllability induced by growth mindset inductions and blame. Given the increasing popularity of growth mindset interventions (e.g., Dweck, 2012), whether such interventions can increase blame attribution over longer patterns of behavior is a pressing issue.

If incremental mindsets increase control, and control increases blame, then those induced to have an incremental mindset should be more prone to blame for failings. However, control may not always be a predecessor of blame, and incremental mindsets do seem generally to be combined with a view of the core character as good, so single failures may be seen as aberrations (Ryazanov & Christenfeld, 2018). However, continual failures may overcome that tendency. Across longer patterns of behavior, those who see poor behavior as controllable may ascribe more blame than those who see it as diagnostic of bad but uncontrollable character. In this case, inducing an incremental mindset could increase judgmental harshness. To explore whether implicit person theories can result in increased judgmental harshness of undesirable behavior, we explore the effects of inducing implicit theories of particular traits on judgments of patterns of behavior. We hypothesized that consistently undesirable behaviors would elicit more blame and moral judgment among those who view the trait as changeable, by increasing perceived control over the specific trait. Thus, people who have been induced to hold an incremental mindset about empathy, and are then asked to judge another's consistent failure in that trait, may find that person more blameworthy for their failure. And, conversely, people induced to hold an incremental mindset about empathy may find those showing an abundance of the trait to be more praiseworthy. We also explore whether such findings would generalize to other traits for which the benefits of an incremental mindset have been demonstrated. We test incremental inductions regarding aggression, where inductions of personality as incremental have shown reductions in aggression (e.g., Yeager, Trzesniewski, Tirri, Nokelainen, & Dweck, 2011). We also explore willpower, where inductions have shown, for example, more adaptive attention allocation on cognitive tasks (Schroder, Moran, Donnellan, & Moser, 2014).

Study 1 – Empathy

Study 1 Procedure

Two hundred and sixty two adults located within the US were recruited as participants via Amazon's Mechanical Turk (mean age = 36.2, SD = 12.7; 61.5% female; 213 participants were retained for analysis after excluding 49 for failing a basic attention check regarding the topic of the induction article). They were told they would be evaluating the appropriateness of reading materials for a high school audience. Participants were randomly assigned to an induction, adapted from Schumann, Zaki, and Dweck (2014), which involved reading a putative newspaper article that gave an overview of scientific research having concluded that empathy is either changeable (incremental) or relatively fixed (entity); all inductions are available as supplementary materials online. Participants were then asked to rate the appropriateness of the article for a high school audience on a 7-point scale. Next participants completed a six-item theories-of-empathy measure, which served as a manipulation check in Schuman, Zaki, and Dweck's (2014) study (e.g. *A person's level of empathy is something very basic about them, and it can't be changed much; 7-point scales anchored on strongly agree and strongly disagree*). Afterwards, participants were told about two people: *Imagine a person, Carol, who consistently behaves in a way that show a complete lack of empathy to the suffering of other people. Imagine another person, Jane, who consistently behaves in a way that shows an especially high level of empathy to the suffering of other people.* Participants were asked to respond to five statements using 7-point scales: *1) How much of the difference between these two people is under their personal control (none-all), 2) This difference in their behavior reflects aspects of these two people that they cannot do much to alter (not at all – completely), 3) How much of this difference is the result of moral choice? (none – all), 4) How blameworthy/praiseworthy is Carol for failing to be empathic (completely praiseworthy – completely blameworthy), 5) How blameworthy/praiseworthy is Jane for her high level of empathy? (completely praiseworthy – completely blameworthy).* Questions 1 and 2 (reverse-coded) were combined into a perceived-control composite measure for analysis.

Participants were then asked a series of questions regarding how they expected their own and other people's levels of empathy to change over time[1]. Participants

[1] Four questions regarding expectations of change in the abstract served as a lead-in to questions regarding expectations for the self and others. The results of these questions are reported in supplementary materials for all studies, but are consistent with manipulation check responses.

responded to the following questions using 7-point scales: *Do you expect your own level of empathy to change over the course of your life? (not at all - very much); How do you expect your level of empathy to change over the course of your life? (levels of empathy will decrease greatly – levels of empathy will increase greatly); Do you expect other people's levels of empathy to change over the course of their lives? (not at all – completely); How do you expect other people's levels of empathy to change over the course of their lives? (levels of empathy will decrease greatly – levels of empathy will increase greatly)*. The analyses of this set of questions will be deferred to a general analysis of the four studies. Afterwards, participants provided demographic information and were debriefed. All measures, manipulations, and exclusions in the study are disclosed. Final sample sizes were determined using study sample sizes in previous literature (e.g. Schuman, Zaki, & Dweck, 2014), which allowed us to have the same power to detect effects of incrementalism increasing blame that previous studies had to detect its positive effects. Data collection did not continue after data analysis.

Study 1 Results

The six-item theories of empathy measure indicated that the articles successfully influenced participant opinions on the entitativity of empathy, $t(211) = 9.04$, $p < .001$, $d = 1.24$ (entity-induced mean = 3.74, SD = 1.45; incremental-induced mean = 5.38, SD = 1.20; Cronbach's $\alpha = .96$). Next, we examined whether the participants who had been induced to adopt a more incremental view showed increased blame attributions to the person behaving unempathically. Indeed, as hypothesized, participants exposed to an entity view of empathy found the low-empathy person less blameworthy than those exposed to an incremental view did, $t(211) = 3.64$, $p < .001$, $d = .50$ (entity-induced mean = .72, SD = 1.37; incremental-induced mean = 1.38, SD =1.36).

The relationship between mindset and blame was mediated by perceived control (Cronbach's α for perceived control measure = .68). The regression of mindset induction on perceived control was statistically significant ($\beta = 1.64$, $t(211) = 9.05$, $p < .001$, SE = .18), as was the regression of perceived control on blame ($\beta = .27$, $t(211) = 4.54$, $p < .001$, SE = .059), see Figure 1. The standardized indirect effect was (.53)(.30) = .31. The significance of this indirect effect was tested using bootstrapping procedures: unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect fully mediated the relationship between induction and blame (mediated effect = .36, $p = .01$, 95% CI [.11, .62]; direct effect = .30, $p = .21$, 95% CI [-.17, .78]). These procedures will be followed for all subsequent mediation analyses.

Incremental-exposed participants also found the difference in behavior between the two people to reflect differences in moral character to a greater degree than entity-exposed participants did, $t(211) = 5.00$, $p < .001$, $d = .69$ (entity-induced mean = 3.77, SD = 1.65, incremental-induced mean = 4.83, SD = 1.43). Incremental participants also found the person showing a high level of empathy more praiseworthy, $t(211) = 3.23$, $p = .001$, $d = .44$ (entity-induced mean = -.51, SD = 1.26; incremental-induced mean = -1.10, SD = 1.42; lower numbers indicate greater praise).

Study 1 Discussion

Being induced to hold an incremental, as opposed to entity, view of empathy lead to increased judgmental harshness of an imagined other with a continually low level of empathy. Consistent with theories of moral responsibility, the increase in judgmental harshness was mediated by increased perceived control over behavior. Furthermore, the difference in behavior was moralized to a greater degree among incremental-induced participants. Incremental participants did find empathic behavior more praiseworthy than entity theorists did, however, suggesting some symmetry in praise and blame, contrary to previous findings that people generally ascribe positive qualities to character more readily than negative ones (e.g., Lockhart, Chang, & Story, 2002; Pizarro, Uhlmann, & Salovey, 2003).

In Experiment 2a and 2b, we explored whether increased judgmental harshness among incremental theorists would replicate with other traits, namely aggression and motivation. Aggression was picked as one of the traits, since its expression so readily has moral implications, and since prior studies have used similar article inductions to demonstrate increased forgiveness for singular transgressions (Yeager, Trzesniewski, Tirri, Nokelainen, & Dweck, 2011). Motivation provides an interesting dilemma for mindset interventions and inductions—the incremental view is that one can do better by working harder, but this assumes that one has control over how hard one is capable of working. What would it mean to not have control over how hard one is willing to work? In a sense, motivation forms a meta-trait for lay theory research, in that if it itself is unchangeable, then limits are placed upon just how much change can be accomplished in any other trait. Here too, article incremental inductions have shown benefits (Schroder, Moran, Donnellan, & Moser, 2014).

Study 2 – Aggression & Motivation

Study 2 Procedure

The procedure for Studies 2a and 2b paralleled that of Study 1, substituting aggression or motivation for empathy. For 2a, we recruited 116 adults (mean age = 37.3, SD = 13.4; 59.8% female; 102 retained for analysis after 14 participants were excluded for failing the attention check) via Amazon's Mechanical Turk. Participants were then randomly assigned to read an article describing one's level of aggression as changeable (incremental) or relatively fixed (entity), adapted from the empathy induction. Next participants completed a six-item theories of aggression measure, which paralleled the theories of empathy measure. Afterwards, participants read about two people: *Imagine a person, Thomas, who consistently behaves in a physically and verbally aggressive way towards people who do not deserve it. Imagine another person, Robert, who consistently does not respond with any aggression, even when repeatedly provoked*. Participants were then asked to respond the same five statements as in the first study, with aggression substituted for empathy. Afterwards participants responded to questions regarding expectations of change for the self and others, adapted to aggression.

For Study 2b, we recruited 136 adults (mean age = 33.7, SD = 10.9; 59.8% female; 102 participants retained for analysis after 34 were excluded for failing the attention check) via Amazon's Mechanical Turk. The manipulation and measures were the same as 2a, but with motivation substituted for aggression. Participants were asked to imagine two people: *Imagine a person, Jessica, who consistently fails to take action to improve her circumstances because she lacks the motivation to do anything constructive. Imagine another person, Rebecca, who consistently improves her circumstance because she is sufficiently motivated to do so*. After imagining both people, participants were asked to respond to five statements, now adapted to motivation. Afterwards participants responded to questions regarding expectations of change in motivation for the self and others. Sample sizes for Studies 2a and 2b were selected to be consistent with prior literature exploring the positive effects of similar inductions.

Study 2 Results

2a: The six-item theories of aggression measure indicated that the articles had successfully influenced participant opinions on the entitativity of aggression, $t(100) = 7.04$, $p < .001$, $d = 1.39$ (entity-induced mean = 3.82, SD = 1.33; incremental-induced mean = 5.54, SD = 1.10; Cronbach's α = .94). Participants exposed to an entity view of aggression found the person showing a high level of aggression to be less blameworthy than did those exposed to an incremental view, $t(100) = 2.34$, $p = .021$, $d = .46$ (entity-induced mean = 1.03, SD = 1.23; incremental-induced mean = 1.68, SD = 1.55). The relationship between mindset and blame was mediated by perceived control (Cronbach's α for perceived control measure = .77). The regression of mindset induction on perceived control was statistically significant ($\beta = 1.44$, $t(100) = 9.05$, $p < .001$, SE = .25), as was the regression of perceived control on blame ($\beta = .39$, $t(100) = 4.28$, $p <.001$, SE = .09), see Figure 2. The standardized indirect effect was (.51)(.39) = .39. The bootstrapped unstandardized indirect effect with 10,000 resamples was significant, thus perceived control fully mediated the relationship between induction and blame (mediated effect = .54, *p* < .001, 95% CI [.20, .91]; direct effect = .11, *p* = .70, 95% CI [-.45, .69]).

Incremental-exposed participants also moralized the difference in behavior to a greater degree than entity-exposed participants, *t*(100) = 2.01, *p* = .047, *d* = .40 (entity-induced mean = 4.24, SD = 1.49; incremental-induced mean = 4.85, SD = 1.59). Incremental participants did not praise the positive behavior significantly more, *t*(100) = 1.61, *p* = .11, *d* = .32 (entity-induced mean = -.64, SD = 1.38; incremental-induced mean = -1.15, SD = 1.83).

2b: The six-item theories of motivation measure indicated that the articles had successfully influenced participant opinions on the entitativity of motivation, *t*(99) = 8.31, *p* < .001, *d* = 1.65 (entity-induced mean = 3.98, SD = 1.25; incremental-induced mean = 5.72, SD = .82; Cronbach's α = .92; 1 blank response). Participants exposed to an entity view of motivation attributed less blame to a person showing a low level of the trait than did those exposed to an incremental view, *t*(99) = 2.78, *p* < .01, *d* = .55 (entity-induced mean = .92, SD = 1.11; incremental-induced mean = 1.53, SD = 1.10; 1 blank response). The relationship between mindset and blame was mediated by perceived control (Cronbach's α for perceived control measure = .58). The regression of mindset induction on perceived control was statistically significant (*β* = 1.14, *t*(99) = 4.94, *p* < .001, SE = .23), as was the regression of perceived control on blame (*β* = .29, *t*(99) = 3.51, *p* = .001, SE = .08), see Figure 3. The standardized indirect effect was (.44)(.33) = .36. The significance of this indirect effect was tested using bootstrapping procedures with 10,000 resamples. The bootstrapped unstandardized indirect effect was statistically significant, thus perceived control fully mediated the relationship between induction and blame (mediated effect = .27, *p* = .01, 95% CI [.05, .56]; direct effect = .34, *p* = .19, 95% CI [-.17, .85]).

Incremental-exposed participants attributed the difference in behavior to a difference in moral character to a greater degree than did entity-exposed participants, *t*(99) = 2.87, *p* < .01, *d* = .57 (entity-induced mean = 3.44, SD: 1.61; incremental-induced mean = 4.43, SD = 1.86; 1 blank response). Incremental-exposed participants did praise the positive behavior more, too, *t*(99) = 2.01, *p* = .047, *d* = .40 (entity-induced mean = -.63, SD = 1.76; incremental-induced mean = -1.36, SD = 1.89; 1 blank response).

Study 2 Discussion

As with empathy, being exposed to an incremental, as opposed to entity, view of aggression lead to increased judgmental harshness of an imagined other with a consistently high level of aggression. Again, the increase in judgmental harshness was mediated by increased perceived control over behavior. Furthermore, the difference between the person acting aggressively and the person showing a low level of aggression was viewed as a moral choice to a greater degree by incremental-exposed participants. Study 2a did not find significant differences as a result of induction in praiseworthiness of showing a low level of aggression, suggesting that there may indeed sometimes be an asymmetry in praise and blame.

Exposure to an incremental view of motivation lead to increased judgmental harshness of an imagined other with a consistently low level of motivation. Again, the increase in judgmental harshness was mediated by increased perceived control over behavior. Furthermore, the difference in behavior was viewed as a moral difference to a greater degree by incremental-exposed participants. Unlike Study 2a, Study 2b did find differences in praiseworthiness of showing high motivation as a result of induction. Thus, holding an incremental mindset regarding motivation was associated with increased judgmental harshness of failures along that trait.

The next experiment explores whether the moral judgments would generalize to people described not with traits terms, but rather with a recounting of their behaviors. Accordingly, the procedure was repeated with detailed vignettes that described two people, one behaving empathically and one unempathically.

Study 3 – Empathy Stories

Study 3 Procedure

The procedure from Study 1 was repeated, with 117 college students participating for partial course credit (mean age = 20.2, SD = 1.88; 78.5% female; 107 retained for analysis after excluding participants failing a basic attention check about the topic of the article). In Study 1, participants had been asked to imagine a person with a consistently high level of empathy, and a person with a consistently low level of empathy. In Study 3, participants were instead given detailed vignettes describing two managers. One had responded with a series of highly empathetic behaviors when her subordinate's daughter was involved in a car accident, such as giving the employee her own sick days. The other manager had responded with a consistent lack of empathy to a similar event, by, for example, not seeing the need for the employee to visit her daughter in the hospital because the daughter was unconscious, see supplementary materials for vignettes. Sample size was selected to be consistent with prior studies exploring the positive effects of mindset inductions.

Study 3 Results

The six-item theories of empathy measure indicated that the articles had again influenced participant opinions regarding the entitativity of empathy, $t(105) = 4.53$, $p < .001$, $d = .88$ (entity-induced mean = 3.75, SD = 1.31; incremental-induced mean = 4.77, SD = 1.00; Cronbach's α = .91). As in the case of imagined scenarios from the first experiment, participants exposed to an entity view of empathy were more forgiving of a person showing a low level of empathy than those exposed to an incremental view, $t(105) = 2.18$, $p = .03$, $d = .42$ (entity-induced mean = .79, SD = .93; incremental-induced mean = 1.19, SD = .93). The relationship between mindset and blame was again mediated by perceived control (Cronbach's α for perceived control measure = .16). While the internal reliability of this measure was low in this study, each of the two items independently significantly differed by induction type ($p = .011$, and $p = .002$), and the results remain consistent with prior studies. The regression of mindset induction on perceived control was statistically significant ($\beta = .74$, $t(105) = 4.00$, $p < .001$, SE = .19), as was the regression of perceived control on blame ($\beta = .32$, $t(105) = 3.79$, $p < .001$, SE = .08), see Figure 4. The standardized indirect effect was (.36)(.35) = .36. The bootstrapped unstandardized indirect effect with 10,000 resamples was statistically significant, thus perceived control fully mediated the relationship between induction and blame (mediated effect = .21, $p < .001$, 95% CI [.07, .41]; direct effect = .18, $p = .19$, 95% CI [-.20, .55]).

Incremental-exposed participants moralized the difference in behavior marginally more than entity-exposed participants did, $t(105) = 1.76$, $p = .08$, $d = .34$ (entity-induced mean = 4.36 SD: 1.44, incremental-induced mean = 4.85, SD = 1.46). Unlike in Study 1, incremental-induced participants did not praise the positive behavior more than entity-induced participants did, *t*(105) = 1.58, *p* = .12, *d* = .31 (entity-induced mean = -1.74, SD = 1.26; incremental-induced mean = -1.24, SD = 1.90). If anything, the results trended towards entity-induced participants praising more.

Study 3 Discussion

Exposure to the incremental, rather than entity, view of empathy lead to increased judgmental harshness of protagonists in a vignette showing a low level of empathy. Thus, our findings from the first experiment were replicated and extended to judging others from more detailed accounts of behavior. While the difference in moralizing behavior was marginal, it was consistent with the pattern observed from the first experiment. Unlike in that experiment, a significant relationship between induction and praise was not observed.

Study 4 –The Consistency of Failure

The results of the previous studies all suggest that entity theorists can be more forgiving than incremental theorists when actors continually fail to improve along particular traits. In order to reconcile our results with the literature showing more incremental theorists to generally be more forgiving, we directly compare the consistent failures we have explored thus far with single instances of failure. It could be that incremental theorists more readily excuse singular failures under the assumption that actors will work to improve themselves, but become increasingly judgmental relative to entity theorists when the failure is consistent and thus signals a lack of effort, or decision not to improve oneself. Study 4 explores whether this is the case.

One could make the argument that the interventions used in the preceding studies cause people to adopt an "impatient" understanding of incrementalism. Indeed, there is some evidence that changeability correlates with positive outcomes in empathy incremental mindsets, but that controllability correlates with negative outcomes (Tullett & Plaks, 2016). Our findings extend beyond empathy, to at least motivation and aggression, and we also use the same empathy induction mindset researchers have used to demonstrate benefits of incremental mindsets. It could, however, be that, by having included control measures, we suggested to participants that control is an important factor in attributing blame. To verify that increased incremental-induced blame was not an artifact of suggesting that control and character-judgments are factors to consider when attributing blame, Study 4 excludes all questions regarding character and controllability. The absence of these questions allows for a purer test of whether incremental inductions themselves are interpreted as conveying controllability and, as a consequence, increased blame. We also incorporated a measure of punishment. There could be many reasons to punish including deterrence, "just desserts", and incapacitation (Carlsmith, Darley, & Robinson, 2002). It could be, for example, that, though they view offenders as less blameworthy, entity theorists feel a heightened need to keep the offender out of society, because of a pessimism regarding their ability to improve.

Study 4 Procedure

The procedure for Study 4 paralleled that of Study 1, but no longer included a praiseworthy individual, and split participants into either seeing consistent or single failure. As in previous studies, half of participants were exposed to an incremental article, and half to an entity article. Half of participants now saw Carol's empathic failure as consistent "*Imagine a person, Carol, who consistently behaves in a way that shows a complete lack of empathy towards the suffering of other people*". The other half saw her failure as a single instance "*Imagine a person, Carol, who behaves in a way that shows a complete lack of empathy towards another person a single time*". Since a praiseworthy individual was no longer included, the seven-point blame scale was now anchored on not blameworthy (1) and completely blameworthy (7). Two hundred and fourteen adults were recruited via Amazon's Mechanical Turk (184 retained for analysis after excluding participants failing an attention check; mean age = 36.2, SD = 13.1; 58.2% female). Sample size was selected so that both single and continual failure conditions had similar power to prior studies that explored the positive effects of incremental mindset inductions.

Study 4 Results

The six-item theories of empathy measure (Cronbach's α = .99) indicated that the articles had successfully influenced participant opinions on the entitativity of empathy, $t(182) = 11.4$, $p < .001$, $d = 1.69$ (entity-induced mean = 2.93, SD = 1.38; incremental-induced mean = 5.21, SD = 1.32).

The primary purpose of Study 4 was to explore whether our results depend on the failure being chronic, and so a single failure would replicate previous studies showing that incremental theorists are more forgiving. A significant interaction between failure (single or consistent) and induction type confirmed that while incremental theorists ascribed more blame to consistent failures, this was not the case for single instances of failure, $F(1, 180) = 6.62$, $p = .011$, $r = .25$, see Figure 5. The simple effect of entity theorists being more forgiving of consistent failures replicated the effect of increased blame by incrementalists we found in our previous studies, *t*(86) = 3.29, *p* = .002, *d* = .70 (entity-induced mean = 3.56, SD = 1.93; incremental-induced mean = 4.76, SD = 1.46). Contrary to prior research showing that incrementalists blame transgressors less for their failures (e.g., Miller et al., 2007), however, we did not find a simple effect of incrementalists being any more forgiving of single failures than entity theorists were, *t*(94) = .13, *p* = .90, *d* = .03 (entity-induced mean = 4.15, SD = 1.59, incremental-induced mean = 4.11, SD = 1.50). There was no main effect of induction type on blame attribution, *F*(1, 180) = .001, *p* = .92, *r* = .01, though there was a predictable main effect of failure type on blame—chronic failure is worse—*F*(1, 180) = 5.79, *p* = .017, *r* = .17.

Next we examined whether people's judgments of punishment showed the same effects as their blame attribution did. Punishment was overall correlated with blame attribution, *F*(1, 182) = 62.4, *p* < .001, *r* = .51. However, punishment did not neatly track blame attribution differences between inductions, in that we did not observe an interaction between failure (consistent or single) and induction, *F*(1, 180) = .200, *p* = .66, *r* = .11 (entity-induced consistent mean = 3.00, SD = 1.69; entity-induced single mean = 2.79, SD = 1.40; incremental-induced consistent mean = 3.02, SD = 1.50; incremental-induced single mean = 2.61, SD = 1.37). There was no main effect of induction, *F*(1, 182) = .084, *p* = .77, *r* = .02, nor an effect of consistency of failure, *F*(1,182) = 1.91, *p* = .17, *r* = .10, though consistent failure trended towards being rated as deserving more severe punishment.

Study 4 Discussion

Study 4 is a partial reconciliation of our results with the general finding in the implicit person theory literature that incremental theorists attribute less blame than entity theorists do. Though we find no evidence for incrementalists being more forgiving for failures seen as single instances, we replicate our prior finding that failures that are consistent become more blameworthy for incrementalists than entity theorists, and find that this effect does not extend to single failures. Thus, our effect of greater blame ascribed by those induced to hold an incremental mindset seems to depend on the failure being consistent. It may be that incrementalists assume a positive character capable of improvement when observing single failures, but grow to view consistent failure as reflecting a negative character that is capable of improvement, but has chosen not to do so, and attribute more blame to the actor accordingly. While it could be the case that our participants endorsed an “impatient” incrementalism rather than a more functional “patient” incrementalism, this would appear to be a problem that extends beyond our methodology and to the study of inducing growth mindsets more generally, as questions regarding control were not asked in this study.

It is surprising that our single-failure condition, which uses the same induction previously used to demonstrate the benefits of incrementalism (Schumann et al., 2014), does not increase forgiveness of other’s empathic failures, despite many studies showing this effect more generally (e.g., Yeager et al., 2011). Though our own theorizing does not entirely account for the lack of effect, neither can implicit person theories as currently specified, and future theorizing must reconcile the observed lack of increased forgiveness by incrementalists for single failures with theory suggesting they should be generally more forgiving. Nonetheless, we do confirm that increased judgmental harshness emerges when there is continual failure. Our findings regarding punishment are also unclear: it appears that though entity theorists blame consistent failures less than incrementalists do, the decrease in blame does not correspond to a decrease in punishment severity. This may well be due to the various functions of punishment, some of which depend on moral blame, and some of which depend on predictions of the improvability of behavior.

General Results

In all the studies, after evaluating the two actors, participants had been asked their expectations of change in trait levels for themselves and others. This permits an analysis of whether being induced to hold an incremental mindset not only creates a possibility of change, but an expectation of change, and, more specifically, improvement.

A meta-analytic approach (Schwarzer, Carpenter, & Rücker, 2015) standardizing effect sizes across sample sizes for the five studies (N = 669) yielded a significant effect of the induction on expectations of change for the self, standardized mean difference (*smd)* = .722, p < .001, 95% CI [.56, .88] (entity-induced mean = 3.64, SD = 1.77, incremental-induced mean: 5.05, SD = 1.66; scale anchored at 1 and 7), as well as for others, *smd* = .892, p < .001, 95% CI [.74, 1.05] (entity-induced mean = 3.57, SD = 1.57; incremental-induced mean = 5.01 SD = 1.35; all meta-analytic results for self/other are also present in the individual studies, across the three traits measured). These findings show that, across the five studies, being exposed to an incremental induction not only increased perceptions of one's own changeability, but also fostered an expectation that this change will occur. The expected change did not reflect greater variability, but instead reflected an asymmetrical expectation of improvement—incremental induced participants expected their own change to be in a more positive direction, *smd* = .436, p < .001, 95% CI [.28, .59] (entity-induced mean = .582, SD = 1.07; incremental-induced mean = 1.16, SD = 1.30, scale anchored on -3 and 3), and others to change in a more positive direction as well, *smd* = .500, *p* <.001, 95% CI [.34, .65] (entity-induced mean = .293, SD = 1.02; incremental-induced mean = 1.01, SD = 1.13; scale anchored on -3 and 3). This is consistent with people generally seeing negative attributes as the more changeable aspects of the self, and incrementalists nonetheless maintaining a fixed positivity towards which all should be striving. Failure to make progress towards improvement becomes more blameworthy than not being capable of making the progress in the first place. The effects of incremental inductions appear to be accompanied by an assumed expectation, or norm of improvement, which, when not met, may signal character to a greater degree to incrementalists than to entity theorists, who may not hold the same expectation.

Consistent with already-presented findings for each individual trait across the four studies exploring consistent praise and blame, a meta-analytic approach shows that entity-exposed participants overall attributed less blame to the person showing undesirable trait levels, *smd* = .477, *p* < .001, 95% CI [.30, .65], (entity-induced mean = .84, SD = 1.21; incremental-induced mean = 1.42, SD = 1.27), and less praise to the person showing desirable trait levels than incremental-exposed participants did, *smd* = .255, *p* = .004, 95% CI [.08, .43] (entity-induced mean = -.81, SD = 1.46; incremental-induced mean = -1.19, SD = 1.69), see Figure 6.

Given the numerous studies demonstrating broad benefits of an incremental mindset, it may be that those at the extreme of incrementalism are most prone to blame others for their failures, and that blame does not linearly relate to incrementality, particularly among more restrained incrementalists. Exploratory analyses utilized Akaike Information Criteria to compare models (Hurvich & Tsai, 1989). These analyses supported an exponential model with both induction and self-reported incrementality entered as factors for data from the first four studies (which examined consistent failures only) fitting the relationship between self-reported mindset and blame better than the linear model (AICc 1239 vs. 1701, respectively), suggesting that data more parsimoniously fit an exponential model than a linear one. This is consistent with a stronger effect at the incremental end of the scale: there may be a threshold beyond which unrestrained incrementalism begins especially to foster blame attributions.

General Discussion

Across empathy, aggression, and motivation, adopting an incremental theory of the trait was associated with greater blame of those showing a continually maladaptive level of the trait. Consistent with theories of moral judgment, greater blame was mediated by increased perceived control over behavior. Thus, the same lay theory mechanism that leads people to exert more effort in the face of challenges can lead people to ascribe more blame to others who are continually unable to surmount these challenges.

The seeming inconsistency between implicit theory findings and theories of moral responsibility regarding moral judgment may be resolved when lay theories regarding the specific attribute are measured, rather than the fixedness of morality or personality more generally, and, more importantly, when chronic behaviors are compared to acute ones. Growth mindset interventions often target specific traits rather than global meaning systems through which one interprets the self and others—our results confirm the necessity of not only exploring morality in relation to implicit theory of morality or personality, but also morality in relation to implicit theory of the specific trait itself.

Consistent failures reveal a negative character that has failed to improve oneself, and result in greater blame attributions than entity theorists would make, an effect that is not apparent with a single failure.

There is a certain irony to the finding that teaching an incremental view of empathy can result in decreased empathy towards others behaving unempathically. Perhaps the most difficult form of empathy involves empathizing with others not feeling empathy—our results suggest that incrementalism can move people away from being able to do so. Results on aggression extend this relationship to another domain, suggesting that incrementalists' increased judgmental harshness of undesirable trait levels is not unique to empathy. The malleability of motivation offers an interesting possibility, since a key aspect of the general incremental position, or even more broadly the self-help stance, is that with enough effort, great things can be achieved. But it is possible to see the amount of effort that people are capable of exerting as fixed. Thus, one could be incremental about various traits, thinking that with enough effort trait levels could be perfected, but maintain an entity view of effort itself, and so regard a failure to perfect a particular trait as not in itself blameworthy. Perhaps this mixed view would decrease blame attributions towards those with certain undesirable trait levels and maintain one's own motivation to improve.

There are certainly significant benefits to an incremental mindset. However, the same increased perceived control that causes one to work to improve trait levels also contributes to increased judgmental harshness of others' failures, particularly at the extremes of incrementalism. Increased blame ascriptions for those with maladaptive trait levels could have societal consequences: The same mechanism that suggests that those in underprivileged communities can overcome adversity through effort and determination could deleteriously lead to the avoidance of addressing the structural inequity that contributes to collective disadvantage, by shifting control away from situational factors, to the victim of the situation. Focusing on individual resilience and perseverance implies increased personal control over outcomes, and this may be accompanied by decreased perceived societal responsibility for providing a supportive environment in the first place. In domains such as economics, the propensity to ignore base rates of failure and instead focus on examples of success has been labeled the survivorship bias (Brown, Goetzmann, Ibbotson, & Ross, 1992). Positing people as nearly-infinitely perfectible, the outer limit of an incremental theory, may, by focusing on survivors that overcome the odds, amplify a neglect of what those odds actually are. Further exploration may yield a teachable psychological flexibility in implicit theory, which confers the benefits of the theory to the self without detriment to treatment of others. Future research should also explore growth mindsets in older populations, who, after exposure to a culture that does not entirely fit their worldview, may develop system-justifying associated meanings for their lay theories that, at the surface level, conveyed infinite potential and inspired greater effort. It could also be that, rather than truly blaming the actor more, incrementalists are misattributing or displacing their own frustration at witnessing a theory-violating pattern of behavior (Plaks, et al., 2005; Plaks & Stecher, 2007; Xu & Plaks, 2015). This would not be inconsistent with our approach, but could be an additional factor contributing to incrementalist condemnation that could be explored in further studies.

As has increasingly been found throughout positive psychology, one-dimensional representations of adaptive traits often obscure the reality that these same mechanisms can be coopted for detrimental processes (McNulty & Fincham, 2012). Thus, teaching an adaptive flexibility, one that views certain traits as uncontrollable in certain situations, may prove most advantageous. A more situated incrementalism may confer benefits to the self, without attributing greater blame to others, and perhaps excuse oneself and others of blame over aspects of the self that that are in actuality beyond individual control. This worldview could resemble Helen Keller's perspective, who, after overcoming the limitations of both deafness and blindness to become one of America's most celebrated figures, said, "I had once believed that we are all masters of our fate — that we could mould our lives into any form we pleased... But as I went more and more about the country I learned that I had spoken with assurance on a subject I knew little about. I forgot that I owed my success partly to the advantages of my birth and environment. Now, however, I learned that the power to rise in the world is not within the reach of everyone" (Dreier, 2012).

Open Practices

The experiments in this article earned Open Materials and Open Data badges for transparent practices.

Chapter 8, in full, is a reprint of the material as it appears in the Journal of Experimental Social Psychology, 2018. Ryazanov, Arseny; Christenfeld, Nicholas. The dissertation/thesis author was the primary investigator and author of this paper.

References

Alicke, M. D. (1992). Culpable causation. *Journal of personality and social psychology*, *63*(3), 368.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556.

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. Review of Financial Studies, 5(4), 553-580.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of personality and social psychology*, *83*(2), 284.

Dreier, P. (2012). *The 100 Greatest Americans of the 20th Century: A Social Justice Hall of Fame*. Nation Books.

Dweck, C. S. (1999). *Self Theories: Their Role in Motivation, Personality, and Development*. Hove: Psychology Press.

Dweck, C. S. (2012). Mindsets and human nature: Promoting change in the Middle East, the schoolyard, the racial divide, and willpower. *American Psychologist*, *67*(8), 614.

Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological inquiry*, *6*(4), 267-285.

Dweck, C. S., yi Hong, Y., & yue Chiu, C. (1993). Implicit theories individual differences in the likelihood and meaning of dispositional inference. *Personality and Social Psychology Bulletin*, *19*(5), 644-656.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review*, *95*(2), 256.

Erdley, C. A., & Dweck, C. S. (1993). Children's implicit personality theories as predictors of their social judgments. *Child development*, *64*(3), 863-878.

Hegarty, P., & Golden, A. M. (2008). Attributional beliefs about the controllability of stigmatized traits: Antecedents or justifications of prejudice?. *Journal of Applied Social Psychology*, *38*(4), 1023-1044.

Heyman, G. D., & Dweck, C. S. (1998). Children's thinking about traits: Implications for judgments of the self and others. *Child development*, *69*(2), 391-403.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307.

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego depletion—Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological science*, *21*(11), 1686-1693.

Kammrath, L. K., & Dweck, C. (2006). Voicing conflict: Preferred conflict strategies among incremental and entity theorists. *Personality and Social Psychology Bulletin*, *32*(11), 1497-1508.

Kammrath, L. K., & Peetz, J. (2012). You promised you'd change: How incremental and entity theorists react to a romantic partner's promised change attempts. *Journal of Experimental Social Psychology*, *48*(2), 570-574.

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children’s beliefs about the stability of traits: Protective optimism?. *Child Development*, *73*(5), 1408-1430.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

McNulty, J. K., & Fincham, F. D. (2012). Beyond positive psychology? Toward a contextual view of psychological processes and well-being. *American Psychologist*, *67*(2), 101.

Miller, C. H., Burgoon, J. K., & Hall, J. R. (2007). The effects of implicit theories of moral character on affective reactions to moral transgressions. *Social Cognition*, *25*(6), 819.

Molden, D. C., & Dweck, C. S. (2006). Finding" meaning" in psychology: a lay theories approach to self-regulation, social perception, and social development. *American Psychologist*, *61*(3), 192.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to earth: People’s psychological concept of free will and its role in moral judgment. *Consciousness and cognition*, *27*, 100-108.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive science*, *39*(1), 96-125.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. *The social psychology of morality: Exploring the causes of good and evil*, 91-108.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise the role of perceived metadesires. *Psychological Science*, *14*(3), 267-272.

Plaks, J. E., Grant, H., & Dweck, C. S. (2005). Violations of implicit theories and the sense of prediction and control: implications for motivated person perception. *Journal of personality and social psychology*, *88*(2), 245.

Plaks, J. E., Levy, S. R., & Dweck, C. S. (2009). Lay theories of personality: Cornerstones of meaning in social cognition. *Social and Personality Psychology Compass*, *3*(6), 1069-1081.

Plaks, J. E., & Stecher, K. (2007). Unexpected improvement, decline, and stasis: a prediction confidence perspective on achievement success and failure. *Journal of personality and social psychology*, *93*(4), 667.

Ryazanov, A. A. and Christenfeld, N. J. S., The strategic value of essentialism, *Social and Personality Psychology Compass,* 2017, e12370.

Schroder, H. S., Moran, T. P., Donnellan, M. B., & Moser, J. S. (2014). Mindset induction effects on cognitive control: A neurobehavioral investigation. *Biological psychology*, *103*, 27-37.

Schumann, K., Zaki, J., & Dweck, C. S. (2014). Addressing the empathy deficit: beliefs about the malleability of empathy predict effortful responses when empathy is challenging. *Journal of personality and social psychology*, *107*(3), 475.

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. Springer.

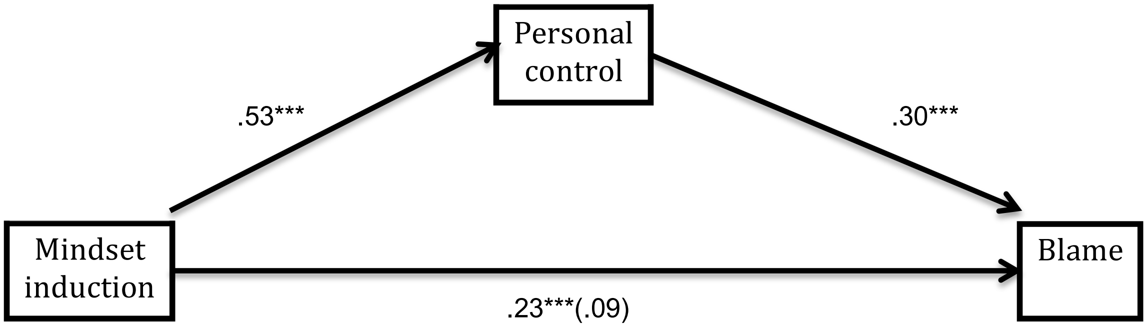
Tullett, A. M., & Plaks, J. E. (2016). Testing the Link Between Empathy and Lay Theories of Happiness. *Personality and Social Psychology Bulletin*, *42*(11), 1505-1521.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72-81.

Xu, X., & Plaks, J. E. (2015). The neural correlates of implicit theory violation. *Social neuroscience*, *10*(4), 431-447.

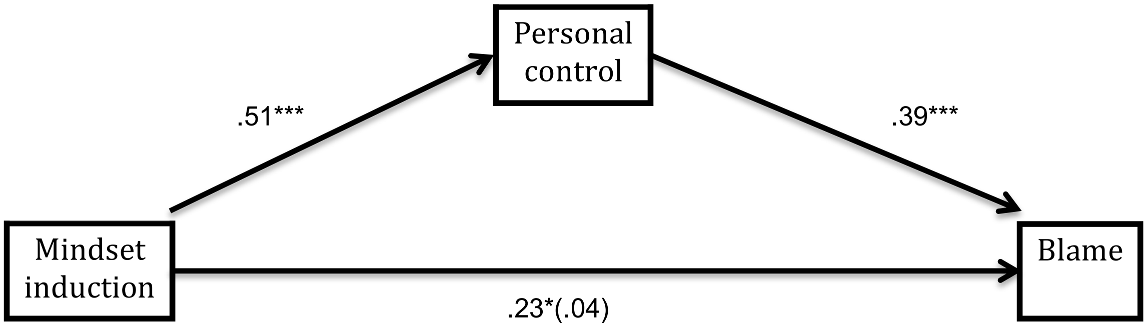
Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., ... & Trott, J. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of educational psychology*, *108*(3), 374.

Yeager, D. S., Trzesniewski, K. H., Tirri, K., Nokelainen, P., & Dweck, C. S. (2011). Adolescents' implicit theories predict desire for vengeance after peer conflicts: correlational and experimental evidence. *Developmental psychology*, *47*(4), 1090.



Empathy Theories

Figure 8.1. Standardized regression coefficients for the relationship between mindset induction and rated blameworthiness of a person showing a consistently low level of empathy as mediated by perceived control. The standardized regression coefficient between mindset induction and rated blameworthiness, controlling for perceived control, is in parentheses.
****p* < .001.



Aggression Theories

Figure 8.2. Standardized regression coefficients for the relationship between mindset induction and rated blameworthiness of a person showing a consistently high level of aggression as mediated by perceived control. The standardized regression coefficient between mindset induction and rated blameworthiness, controlling for perceived control, is in parentheses.
****p* < .001, **p* < .05

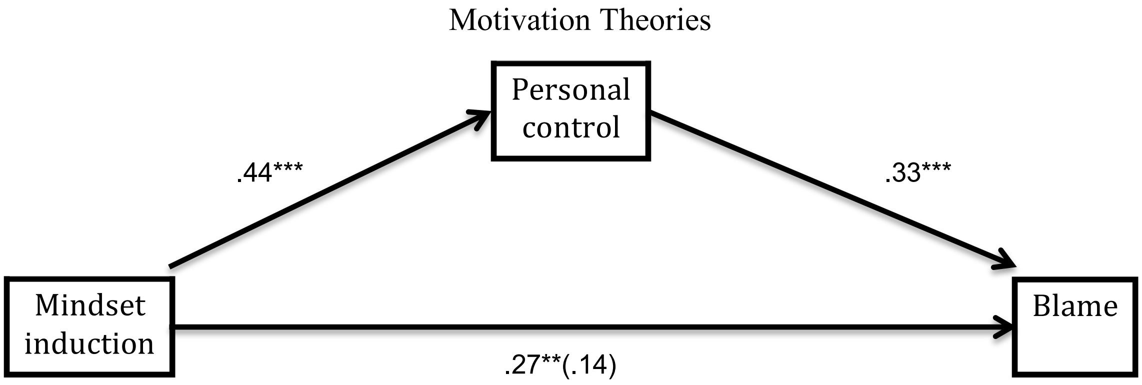


Figure 8.3. Standardized regression coefficients for the relationship between mindset induction and rated blameworthiness of a person showing a consistently low level of motivation as mediated by perceived control. The standardized regression coefficient between mindset induction and rated blameworthiness, controlling for perceived control, is in parentheses.
***p < .001, **p < .01.

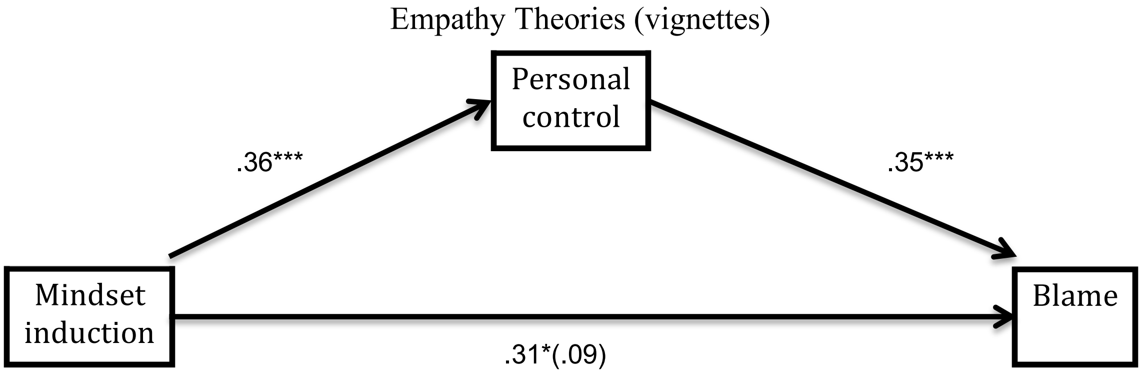


Figure 8.4. Standardized regression coefficients for the relationship between mindset induction and rated blameworthiness of a person showing a consistently low level of empathy in a detailed vignette as mediated by perceived control. The standardized regression coefficient between mindset induction and rated blameworthiness, controlling for perceived control, is in parentheses.
***p < .001.

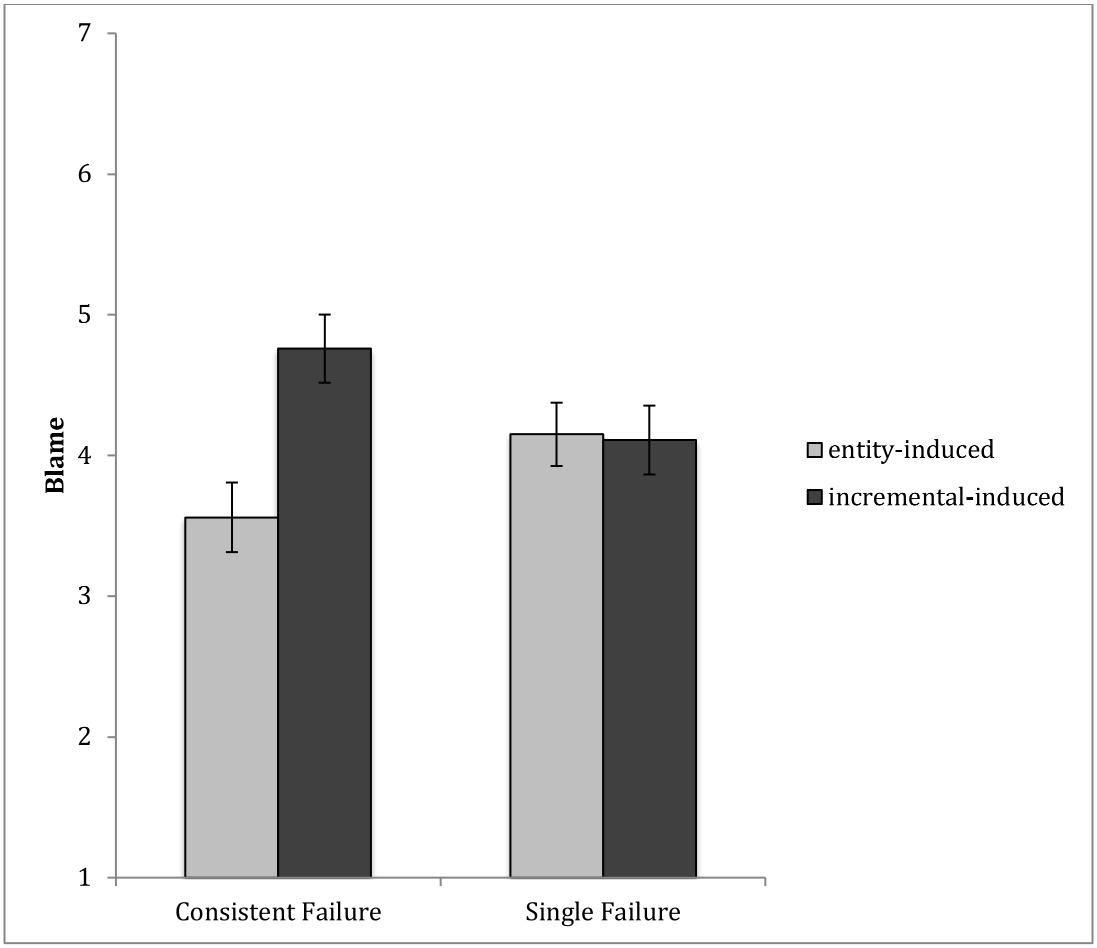


Figure 8.5. Interactive effect of mindset induction and failure type on blameworthiness of a person showing empathic failure. Responses were along a scale from 1 to 7, anchored at not at all blameworthy and completely blameworthy, respectively.

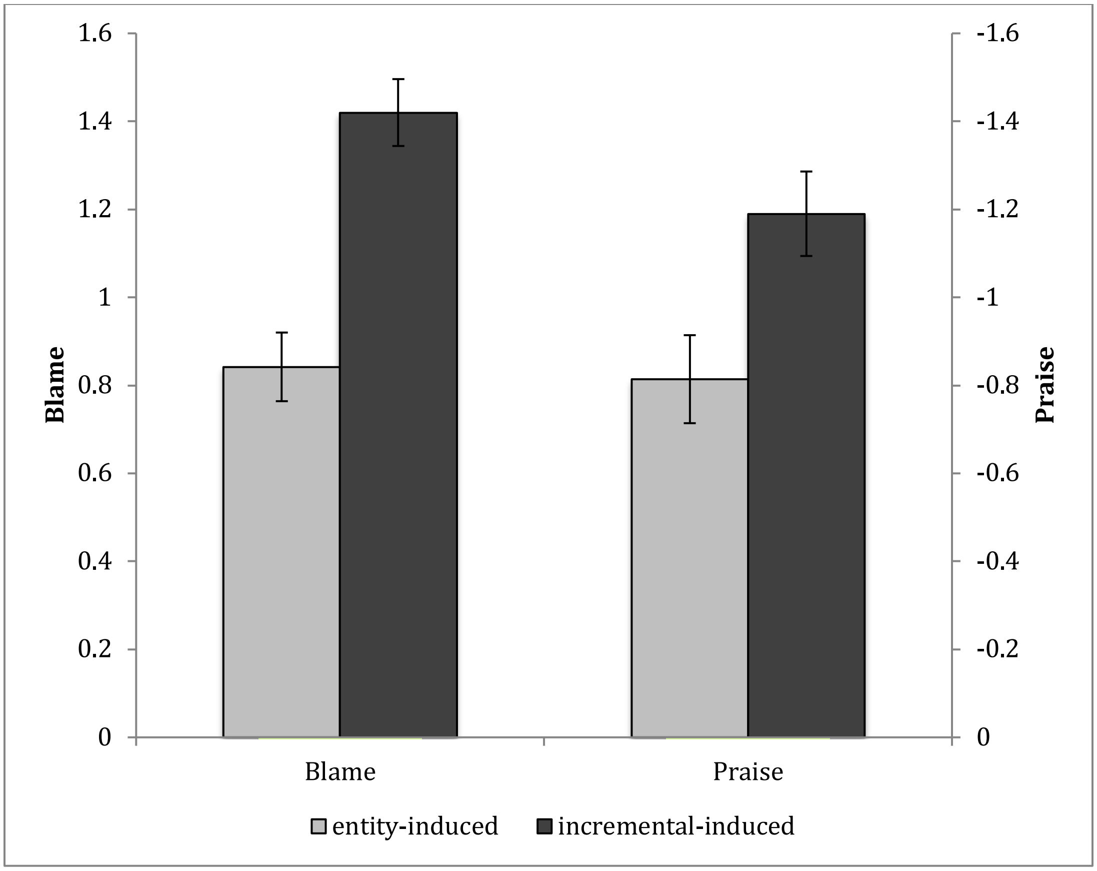


Figure 8.6. Relationship across studies between mindset induction and rated blameworthiness of a person consistently showing undesirable trait levels (blame) and rated praiseworthiness of a person consistently showing desirable trait levels (praise). Both responses were along a scale from -3 to 3, anchored at completely praiseworthy and completely blameworthy, respectively.

Conclusion

This dissertation examined three mechanisms underlying aspects of moral cognition: The role of expected value calculation in judgments of actions that harm some to benefit many, how mental accounting of prior good deeds excuses further beneficence, and how changeability beliefs influence responsibility attribution. Chapter 1 identified how intuitive probabilities can override stated outcomes in moral dilemmas, where harm to few will result in benefit to many. Chapters 2 and 3 found that participants were sensitive to expected value, probability, and where shifts in probability occur in moral dilemmas, providing evidence against affective and deliberative paths being competing cognitive processes. Additionally, a discrepancy between single evaluations and joint evaluations of moral dilemmas was identified, such that participants deviated further from expected value under joint evaluation than under single evaluation. Chapter 4 provided additional evidence for the integration of affective and deliberative paths by demonstrating how incidental affect can shift risk preferences in moral decisions. Chapters 1-4 thus revealed the affective and deliberative paths in moral reasoning to be more interconnected than current dual-system models suggest, and that moral judgments are sensitive to expected value, outcomes likelihoods, where shifts in likelihood occur, decision-mode, and incidental affect.

These findings have philosophical implications: People appear to be neither pure utilitarians, willing to cause harm as long as the action results in a net benefit, nor absolutist deontologists, refusing to cause any harm for the greater good. Instead, participant responses are consistent with threshold deontology—a certain amount of good must be achieved in order to justify imposing harm or a risk of harm. Our findings also raise the issue of how risks of harm relate to causing definite harm, a topic of interest to those subscribing to deontology, which provides less clear guidance regarding imposing risks of harm than it does for avoiding causing certain harm.

Such research will also help inform a variety of applied decisions regarding how to trade off risks of harm and probabilistic benefits in a variety of domains, from public health policy, to how autonomous vehicle collision avoidance systems should be programmed to decide between possible outcomes. Additionally, these findings identify that choosing between plans can lead to decision-making less consistent with expected value than choosing whether to carry out a single plan. Which of these decision-modes is normatively correct cannot be empirically derived. However, our studies do suggest that, when deciding between certainly harming few or a causing greater harm, as expressed through a lower likelihood of harm but to a greater number people, choosing between plans may lead to suboptimal decisions with respect to expected value. Overall, our research indicates that findings from single-evaluation studies of moral preferences with certain outcomes may not generalize to real-world ethical preferences, where outcomes are generally not certain to occur, nor is the range of possible actions necessarily constrained to a single option.

Chapter 5 moved from the judgment of actions in which harm is done for the greater good, to examining the cognitive representation of good deeds. The proposed Moral Accounting Model outlined how, in addition to the increased cost of further action, moral credit from having already done good contributes to the acceptability of declining subsequent opportunities for beneficence. Moral credit was found to be sensitive to effort, effect, domain, and time since beneficence. Because of the steep depreciation of moral credit, monthly donations resulted in more enduring moral credit than equivalent one-time donations, suggesting that monthly donation schemes may decrease the amount of good that would otherwise done by the people enrolled in them. Our model thus reveals an intuitive solution to the philosophical challenge of how much good one ought to do in the world. Chapter 6 further examined a specific aspect of beneficence: the extent to which donors care about the effectiveness of their donations. Findings suggest that participants have limited concern for efficiency, or the proportion of donations going to the cause rather than administrative overhead, a metric often used as a proxy for effectiveness.

The final section of the dissertation turned from judgments of actions to judgments of people. Chapters 7 and 8 examined how blame attributions of others relate to how changeable one views people to generally be. Chapter 7 proposed a theoretical account of how the widely-documented benefits of changeable mindsets and the rejection of essences underlying groups and individuals may, ironically, rely on a perceived deeper positive essence. In contexts of continual failure, however, believing in changeability could instead result in increased blame of those unable to improve their circumstance, by implying that they are capable of doing if only they tried harder. Chapter 8 provided empirical evidence for this perspective by demonstrating that those exposed to incremental (changeable) inductions judged continuous failures more harshly than those exposed to an entity (fixed) induction. Such research cautions against the increasingly widespread adoption of mindset interventions by schools and businesses around the world. Since belief in changeability is accompanied by increased perceived responsibility for achievement, implementing such interventions in communities and populations facing actual barriers to achievement may not only leave the root causes of underachievement unaddressed, but can also shift blame away from structural barriers, and towards members of those communities.

Our exploration of three aspects of moral cognition underscores the close relationship between basic cognition, lay philosophy, and moral judgment of specific actions and people. Understanding the cognitive processes underlying moral judgment thus helps inform practical ethical decisions, such as how to program autonomous vehicles, how to encourage beneficence, and how to best address achievement gaps. Though our research does not reveal which positions are normatively correct, a closer examination of their cognitive bases aids in building consensus on which components ought to matter, and which disregarded, in developing normative frameworks for addressing the ethical challenges we face.

Appendix

Chapter 3 Supplementary Studies

Chapter 3 Supplementary Study 1

To verify that the observed lack of effect on the saving side observed in Study 1was not the result of framing the saving likelihood as a decrease in the likelihood of dying, rather than an increase in the likelihood of survival, a separate study compared the framing the benefit as an increase in the likelihood of the eight surviving (0-25% increase, 75-100% increase), to a decrease in the likelihood of eight surviving (25-0% decrease, 100-75% decrease).

Three hundred and forty six participants were recruited via Amazon’s Mechanical Turk (290 passed an attention check; 55.8% female; mean age = 35.0, SD = 12.9). Because results do not differ between the full set of participants and those passing the attention check, results are reported for all participants. A 2x2 ANOVA was used to analyze the results. There was no main effect of framing, $F(1, 342) = .80$, $p = .37$, $r = .08$, nor a main effect of location of shift, $F(1, 342) = .35$, $p = .55$, $r = .02$. The critical test for whether “survive” framing would result in location sensitivity was the interaction of framing and shift, which was not significant, $F(1, 342) = 2.44$, $p = .12$, $r = .05$ (mean decrease likelihood of dying from 25-0% = .53, SD = .326; mean increase in likelihood of surviving 0-25% = .77, SD = 2.27; mean decrease likelihood of dying 100-75% = .874, SD = 3.14; mean increase likelihood of surviving 75-100% = 0.00, SD = 3.46; $p = .36$ for participants passing attention check). The results suggest that, regardless of framing, participants are insensitive to location shifts for likelihoods of benefit.

Chapter 3 Supplementary Study 3a I

One hundred and thirteen participants were recruited via Amazon's Mechanical Turk service (100 passed an attention check; 55.8% female; mean age = 35.0, SD = 12.9). Because results do not differ between the full set of participants and those passing the attention check, results are reported for all participants. Participants were exposed to the 0-95 vs 95-100 joint evaluation in study 3a, without any preceding single evaluations. There was no preference between the 0-95 plan, and the 95-100 plan, $t(112) = .84$, $p = .40$, $d = .08$ (mean = -.26, SD = 3.25). Thus, participants were indifferent between a plan that raised the likelihood of four bystanders dying from 0-95% and a plan that raised the likelihood of a different four bystanders dying from 95-100%, in order to save two individuals, consistent with results obtained in Study 3a.

Chapter 3 Supplementary Study 3a II

One hundred and seven participants were recruited via Amazon's Mechanical Turk service (100 passed an attention check; 62.6% female; mean age = 34.5, SD = 11.4). Because results do not differ between the full set of participants and those passing the attention check, results are reported for all participants. Participants were exposed to the 0-50 vs 50-100 joint evaluation in study 3a, without any preceding single evaluations. There was a strong preference for the 0-50 plan over the 50-100 plan, $t(106) = 13.9$, $p < .001$, $d = 2.34$ (mean = -2.89, SD = 2.15). Thus, participants had a strong preference for a plan that raised the likelihood of four bystanders dying from 0-50% over a plan that raised the likelihood of a different four bystanders dying from 50-100%, in order to save two individuals, consistent with results obtained in Study 3a.

Chapter 5 Study 2 Scenarios

Soup Kitchen

High Cost High Contribution

Gene has just put in a ten hour shift at the local soup kitchen feeding the homeless and is on his way home. He is approached by a man seeking someone to help for one hour at the soup kitchen. The help does not require any prior experience or expertise.

Low Cost Low Contribution

John has had a quiet day relaxing at home. He steps outside his home and is approached by a man seeking someone to help for one hour at the soup kitchen feeding the homeless. The help does not require any prior experience or expertise.

Low Cost High Contribution

Geoff has just made a financial donation to the local soup kitchen feeding the homeless. His donation is sufficient to hire someone to work a ten hour shift at the kitchen. He is approached by a man seeking someone to help for one hour at the soup kitchen. The help does not require any prior experience or expertise.

High Cost Low Contribution

Gabe has just put in a ten hour shift as a chef at the restaurant where he works, and is on his way home. He is approached by a man seeking someone to help for one hour at a soup kitchen feeding the homeless. The help does not require any prior experience or expertise.

Kidney Donation

High Cost High Contribution

Grace has donated one of her kidneys to save a distant relative’s life. A scientist who is on the verge of finding an immunization for malaria – a discovery that could save hundreds of thousands of lives – will die if he does not get a rapid kidney transplant, and the only viable donor is Grace. A person can live normally with one kidney, but with none they must spend the rest of their life on painful, time-consuming dialysis, and mobility is severely limited.

Low Cost Low Contribution

Jackie has not done organ donation, and still has both of her kidneys. A scientist who is on the verge of finding an immunization for malaria – a discovery that could save hundreds of thousands of lives – will die if he does not get a rapid kidney transplant, and the only viable donor is Jackie. A person can live normally with one kidney, but with none they must spend the rest of their life on painful, time-consuming dialysis, and mobility is severely limited.

Low Cost High Contribution

Gloria has made a large financial contribution that funded one kidney transplant surgery for someone who could not afford the life-saving procedure. Gloria has not done organ donation herself, and still has both of her kidneys. A scientist who is on the verge of finding an immunization for malaria – a discovery that could save hundreds of thousands of lives – will die if he does not get a rapid kidney transplant, and the only viable donor is Gloria. A person can live normally with one kidney, but with none they must spend the rest of their life on painful, time-consuming dialysis, and mobility is severely limited.

High Cost Low Contribution
Gwen recently discovered that she was born with one kidney. A scientist who is on the verge of finding an immunization for malaria – a discovery that could save hundreds of thousands of lives – will die if he does not get a rapid kidney transplant, and the only viable donor is Gwen. A person can live normally with one kidney, but with none they must spend the rest of their life on painful, time-consuming dialysis, and mobility is severely limited.

Boat Rescues
High Cost High Contribution
Thousands are homeless and stranded after Hurricane Katrina. Glenn has spent twelve hours today rescuing people using his boat and ferrying them to nearby shelters. He is very tired from his long day. He is approached by a volunteer who asks him to spend the next hour looking for people. By doing so, he could rescue about ten people. The help does not require any prior experience or expertise.

Low Cost Low Contribution
Thousands are homeless and stranded after Hurricane Katrina. Joel has spent twelve hours hanging out with his friends. He is approached by a volunteer who asks him to spend the next hour looking for people. By doing so, he could rescue about ten people. The help does not require any prior experience or expertise.

Low Cost High Contribution
Thousands are homeless and stranded after Hurricane Katrina. Greg has made a donation sufficient to cover the cost of a hiring a fisherman with a boat to spend twelve hours today rescuing people and ferrying them to nearby shelters. Greg is approached by a volunteer who asks him to spend the next hour on a similar boat looking for people. By doing so, he could rescue about ten people. The help does not require any prior experience or expertise.

High Cost Low Contribution
Thousands are homeless and stranded after Hurricane Katrina. Gil is a commercial fisherman who has just spent twelve hours fishing on his boat. He is very tired from his long day. He is approached by a volunteer who asks him to spend the next hour using his boat to look for people. By doing so, he could rescue about ten people. The help does not require any prior experience or expertise.

Chapter 5 Study 3 Scenarios

Soup Kitchen

Minimal Effort, No Good

Greg spends the last 5 minutes of his shift as a chef at a restaurant preparing leftover ingredients into enough food to feed 20 homeless people. He arranges for a soup kitchen volunteer to pick up the food. However, the soup kitchen's delivery truck breaks down and cannot pick up the food, so the food Greg has prepared does not get served. When he gets home, Greg finds a flyer soliciting monetary donations to the soup kitchen.

High Effort, No Good

Jerry spends the last 4 hours of his shift as a chef at a restaurant preparing leftover ingredients into enough food to feed 20 homeless people. He arranges for a soup kitchen volunteer to pick up the food. However, the soup kitchen's delivery truck breaks down and cannot pick up the food, so the food Jerry has prepared does not get served. When he gets home, Jerry finds a flyer soliciting monetary donations to the soup kitchen in his mailbox.

Minimal Effort, High Good

Gabe spends the last 5 minutes of his shift as a chef at a restaurant preparing leftover ingredients into enough food to feed 20 homeless people. He arranges for a soup kitchen volunteer to pick up the food. The food Gabe has prepared gets served. When he gets home, Gabe finds a flyer soliciting monetary donations to the soup kitchen in his mailbox.

High Effort, High Good

John spends the last 4 hours of his shift as a chef at a restaurant preparing leftover ingredients into enough food to feed 20 homeless people. He arranges for a soup kitchen volunteer to pick up the food. The food John has prepared gets served. When he gets home, John finds a flyer soliciting monetary donations to the soup kitchen in his mailbox.

Boat Rescue

Minimal Effort, No Good

Thousands are homeless and stranded after Hurricane Katrina. George is a commercial fisherman who has just spent ten hours fishing on his boat. He has spent the last 10 minutes of his fishing trip searching for stranded people. In that time he has been unable to rescue anyone. As he is docking his boat, George is approached by a volunteer who asks George to spend the next hour using his boat to look for people. By doing so, he would rescue about 5 people.

Minimal Effort, High Good

Thousands are homeless and stranded after Hurricane Katrina. Jason is a commercial fisherman who has just spent ten hours fishing on his boat. He has spent the last 3 hours of his fishing trip searching for stranded people. In that time he has been unable to rescue anyone. As he is docking his boat, Jason is approached by a volunteer who asks Jason to spend the next hour using his boat to look for people. By doing so, he would rescue about 5 people.

High Effort, Minimal Good

Thousands are homeless and stranded after Hurricane Katrina. Gil is a commercial fisherman who has just spent ten hours fishing on his boat. He has spent the last 10 minutes of his fishing trip searching for stranded people. In that time he has been able to rescue 10 people. As he is docking his boat, Gil is approached by a volunteer who asks Gil to spend the next hour using his boat to look for people. By doing so, he would rescue about 5 more people.

High Effort, High Good

Thousands are homeless and stranded after Hurricane Katrina. Jim is a commercial fisherman who has just spent ten hours fishing on his boat. He has spent the last 3 hours of his fishing trip searching for stranded people. In that time he has been able to rescue 10 people. As he is docking his boat, Jim is approached by a volunteer who asks Jim to spend the next hour using his boat to look for people. By doing so, Jim would rescue about 5 more people.

Chapter 5 Study 4 Scenarios

Soup Kitchen

High Prior Contribution Same Domain

Gene has just put in a ten hour shift at the local soup kitchen feeding the homeless and is on his way home. He is approached by a man raising funds for the soup kitchen.

Low Prior Contribution Same Domain

John has had a quiet day relaxing at home. He steps outside his home and is approached by a man seeking monetary donations for a soup kitchen feeding the homeless.

High Prior Contribution Similar Domain

Gary has just put in a ten hour shift at the local soup kitchen feeding the homeless and is on his way home. He is approached by a man raising funds for a new homeless shelter.

Low Prior Contribution Similar Domain

Jeff has had a quiet day relaxing at home. He steps outside his home and is approached by a man raising funds for a new homeless shelter.

High Prior Contribution Different Domain

Gerry has just put in a ten hour shift at the local soup kitchen feeding the homeless and is on his way home. He is approached by a man raising funds for an animal rescue center that would help wild animals recover from a recent oil spill.

Low Prior Contribution Different Domain

Jon has had a quiet day relaxing at home. He steps outside his home and is approached by a man raising funds for an animal rescue center that would help wild animals recover from a recent oil spill.

Kidney Donation

High Prior Contribution Same Domain

Grace has donated one of her kidneys to save a distant relative's life. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

Low Prior Contribution Same Domain

Jackie has not done organ donation, and still has both of her kidneys. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

High Prior Contribution Similar Domain

Gina has donated one of her kidneys to save a distant relative's life. She is approached by a volunteer from her local blood bank who informs her that the blood bank is raising money to fund a second mobile blood donation station that will greatly increase the blood bank's supply of blood.

Low Prior Contribution Similar Domain
Jennifer has not done organ donation, and still has both of her kidneys. She is approached by a volunteer from her local blood bank who informs her that the blood bank is raising money to fund a second mobile blood donation station that will greatly increase the blood bank's supply of blood.

High Prior Contribution Different Domain
Gloria has donated one of her kidneys to save a distant relative's life. She is approached by a local group that is raising awareness about the destruction of polar bear habitats and fundraising to conserve polar bear habitats.

Low Prior Contribution Different Domain
Jennifer has not done organ donation, and still has both of her kidneys. She is approached by a local group that is raising awareness about the destruction of polar bear habitats and fundraising to conserve polar bear habitats.

Boat Rescues
High Prior Contribution Same Domain
Thousands are homeless and stranded after Hurricane Katrina. Glenn has spent twelve hours today rescuing people using his boat and ferrying them to nearby shelters. He is very tired from his long day. He is approached by a volunteer who asks him to make a monetary donation that will fund the rescue of about ten more people.

Low Prior Contribution Same Domain
Thousands are homeless and stranded after Hurricane Katrina. Joel has spent twelve hours hanging out with his friends. He is approached by a volunteer who asks him to make a monetary donation that will fund the rescue of about ten more people.

High Prior Contribution Similar Domain
Thousands are homeless and stranded after Hurricane Katrina. Geoff has spent twelve hours today rescuing people using his boat and ferrying them to nearby shelters. He is approached by a volunteer who asks him to make a monetary donation to a local food bank that is feeding the stranded.

Low Prior Contribution Similar Domain
Thousands are homeless and stranded after Hurricane Katrina. Jason has spent twelve hours hanging out with his friends. He is approached by a volunteer who asks him to make a monetary donation to a local food bank that is feeding the stranded.

High Prior Contribution Different Domain
Thousands are homeless and stranded after Hurricane Katrina. Gil has spent twelve hours today rescuing people using his boat and ferrying them to nearby shelters. He is approached by a volunteer who asks him to make a monetary donation to an organization that saves elephants from poachers.

Low Prior Contribution Different Domain
Thousands are homeless and stranded after Hurricane Katrina. Joel has spent twelve hours hanging out with his friends. He is approached by a volunteer who asks him to make a monetary donation to an organization that saves elephants from poachers.

Chapter 5 Study 5 Scenarios

Soup Kitchen

High Contribution Today

Gene is approached by a man raising funds for a local soup kitchen on his way home from putting in a ten hour shift at the local soup kitchen feeding the homeless.

Low Contribution Today

Jeff has had a quiet day relaxing at home. He steps outside his home and is approached by a man raising funds for a local soup kitchen.

High Contribution Week

Gary is approached by a man raising funds for a local soup kitchen. A week ago he put in a ten hour shift at the local soup kitchen feeding the homeless.

High Contribution Year

Gerry is approached by a man raising funds for a local soup kitchen. A year ago he put in a ten hour shift at the local soup kitchen feeding the homeless.

High Contribution Decade

Greg is approached by a man raising funds for a local soup kitchen. A decade ago he put in a ten hour shift at the local soup kitchen feeding the homeless.

Kidney Donation

High Contribution Today

Grace has just donated one of her kidneys to save a distant relative's life. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

Low Contribution Today

Jackie has not done organ donation, and still has both of her kidneys. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

High Contribution Week

Gina donated one of her kidneys to save a distant relative's life a week ago. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

High Contribution Year

Gloria donated one of her kidneys to save a distant relative's life one year ago. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

High Contribution Decade
Gloria donated one of her kidneys to save a distant relative's life a decade ago. She is approached by a person raising money to fund a kidney transplant for a person in the community in desperate need of a transplant who cannot otherwise afford it.

After School Program
High Contribution Today
Glenn is approached by a volunteer who asks him to make a monetary donation to help fund local after-school programs while Glenn is on his way home from spending 4 hours volunteering for an after-school program.

Low Contribution Today
Joel is approached by a volunteer who asks him to make a monetary donation to help fund local after-school programs while Joel is on his way home from hanging out with friends for 4 hours.

High Contribution Week
Geoff is approached by a volunteer who asks him to make a monetary donation to help fund local after-school programs. A week ago, Geoff spent 4 hours volunteering for an after school program.

High Contribution Year
Gil is approached by a volunteer who asks him to make a monetary donation to help fund local after-school programs. A year ago, Gil spent 4 hours volunteering for an after school program.

High Contribution Decade
Gerald is approached by a volunteer who asks him to make a monetary donation to help fund local after-school programs. A decade ago, Geoff spent 4 hours volunteering for an after-school program.

Chapter 5 Study 6 Scenarios

Soup Kitchen
Single Donation Today
Gene is approached by a man raising funds for a local soup kitchen on his way home from having donated $240 to the soup kitchen.

Monthly Donation Today
Gary is approached by a man raising funds for a local soup kitchen on his way home from signing up to make a monthly donation of $20 per month to the soup kitchen for a year, automatically withdrawn from his bank account on the first of the month each month, so $20 has been deducted from his bank account that month. His monthly donations will total $240 over the course of the year.

Single Donation 9 Months Ago
Greg is approached by a man raising funds for a local soup kitchen. Nine months ago he donated $240 to the soup kitchen.

Monthly Donation Started 9 Months Ago
Gerry is approached by a man raising funds for a local soup kitchen. Nine months ago he signed up to make a monthly donation of $20 per month to the soup kitchen for a year, automatically withdrawn from his bank account on the first of the month each month, so $20 has been deducted from his bank account that month. His monthly donations will total $240 over the course of year.

After School Program
Single Donation Today
Jason is approached by a woman seeking volunteers to help out at a local after school program on his way home from donating $600 to the after school program.

Monthly Donation Today
Jonathan is approached by a woman seeking volunteers to help out at a local after school program on his way home from signing up to make a monthly donation of $50 per month to the after school program for a year, automatically withdrawn from his bank account on the first of the month each month, so $50 has been deducted from his bank account that month. His monthly donations will total $600 over the course of year.

Single Donation 9 Months Ago
Jim is approached by a woman seeking volunteers to help out at a local after school program. Ten months ago he donated $600 to the after school program.

Monthly Donation Started 9 Months Ago
Jacob is approached by a woman seeking volunteers to help out at a local after school program. Ten months ago he signed up to make a monthly donation of $50 per month to the after school program for a year, automatically withdrawn from his bank account on

the first of the month each month, so $50 has been deducted from his bank account that month. His monthly donations will total $600 over the course of year.