Veil-of-Ignorance Reasoning and Justification of Moral Judgments

Veil-of-Ignorance Reasoning and Justification of Moral Judgments

A dissertation presented

by

Karen Huang

to

The Committee for the Ph.D. in Business Studies

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Organizational Behavior

Harvard University

Cambridge, Massachusetts

March, 2020

VEIL-OF-IGNORANCE REASONING AND JUSTIFICATION OF MORAL JUDGMENTS

ABSTRACT

The "veil of ignorance" is a moral reasoning device designed to promote impartial decision-making by denying decision-makers access to potentially biasing information about who will benefit most or least from the available options. In this research, I investigate the following questions: Does veil-of-ignorance reasoning influence moral decisions involving tradeoffs between the greater good and competing moral concerns? If so, what is this kind of influence? How do third-party observers perceive veil-of-ignorance reasoning?

In Chapter 1, I introduce moral dilemmas involving tradeoffs between the greater good and competing moral concerns, the use of veil-of-ignorance reasoning in these dilemmas, and the importance of third-party assessments of moral reasoning. In Chapter 2, I test a veil-of-ignorance reasoning intervention and find that this intervention influences moral judgment, specifically in the utilitarian direction. This result generalizes across many decision domains, including decisions with real stakes. Furthermore, this result is explained by impartial reasoning, and not by alternative explanations such as anchoring, probability calculation, or perspective-taking. Chapter 3 examines the interpersonal effects of veil-of-ignorance reasoning. When a decision-maker is dealing with a moral dilemma where the utilitarian response is unpopular, employing a veil-of-ignorance justification of the utilitarian response increases observer trust of the decision-maker, an effect driven by perceived warmth. Overall, VOI reasoning could serve as an intervention to maximize good consequences, while also signaling respect for the individuality of persons and promoting interpersonal trust. These results have implications for both individual and collective decision-making, as well as how democratic deliberation could proceed regarding sensitive moral issues.

TABLE OF CONTENTS

ABSTRACT……iii

TABLE OF CONTENTS……...iv

ACKNOWLEDGEMENTS……v

CHAPTER 1. Introduction……1
Reasoning about Moral Decisions……6
Justification of Moral Decisions……7
Overview……8
References……9

CHAPTER 2. Veil-of-Ignorance Reasoning Favors the Greater Good
Title……12
Abstract……13
Significance Statement……14
Introduction……15
Experimental Designs and Results……21
Study 1……21
Study 2……24
Study 3……25
Study 4……26
Study 5……27
Study 6……28
Study 7……29
Discussion……30
Materials and Methods……34
References……40

CHAPTER 3. Third-Party Judgments of Veil-of-Ignorance Reasoning
Title……43
Abstract……44
Introduction……45
Overview of Studies……50
Study 1……51
Study 2……62
Study 3……77
General Discussion……87
References……93

APPENDICES……97

ACKNOWLEDGEMENTS

I am exceedingly grateful and indebted to my wonderful dissertation committee, Alison Wood Brooks, Joshua D. Greene, Max Bazerman, and Mike Norton, all of whom have been incredibly generous with their time, advice, and commitment to my research and development as a scholar. I am deeply grateful for financial support from the Dissertation Completion Fellowship, HBS Doctoral Programs, Alison Wood Brooks, Joshua D. Greene, and Max Bazerman. Thank you to my HBS doctoral student colleagues, Greene lab, and NERD lab for the camaraderie and scholarly community over the past six years. Thank you to the Institute for Quantitative Social Science, especially Steve Worthington, for statistical advice and consultation. My deepest thanks to my mother, father, and friends for their unconditional love and support.

CHAPTER 1.

INTRODUCTION

At a Paris auto show, a Mercedes Benz executive, who heads the autonomous vehicle division of the company, was approached by a reporter who asked him whether the autonomous vehicle should prioritize the safety of its passengers or value all lives equally. He replied, "Save the one in the car […] that's your first priority." As one could imagine, this response prompted a public outcry, where journalists reported that Mercedes Benz is favoring the lives of its passengers, who are likely very wealthy and privileged, over the lives of ordinary bystanders (Morris, 2016). One could imagine the executive having provided a different response. He could have said that the car should value all lives equally. In that case, there might have been a different public outcry. In a situation where a car is traveling down a road and there is one passenger in the car and nine pedestrians on the road, if the car is programmed to value all lives equally, then the car should be programmed to sacrifice the passenger in order to save the nine pedestrians (Bonnefon, Shariff & Rahwan, 2016). In this case, people would express outrage that there are cars on the road that are killing their passengers. As one can see, there is no easy answer to the moral dilemma of whether the autonomous vehicle should prioritize the safety of its passengers or value all lives equally. How should one reason to resolve such a dilemma, and how should one justify one's decision to the public?

This is not just a dilemma faced by autonomous vehicle executives. Such a moral dilemma – where there is fundamentally a tradeoff between the greater good and competing moral concerns – is faced by leaders, policymakers, and ordinary citizens and decision-makers. In this work, I examine the following research questions: Does veil-of-ignorance reasoning influence moral decisions involving tradeoffs between the greater good and competing moral

concerns? How do third-party observers perceive the use of veil-of-ignorance reasoning? In the following sections, I will first explain these moral dilemmas involving tradeoffs in more detail. Second, I will describe veil-of-ignorance reasoning – both the philosophical foundation and the operationalization of this concept in the current empirical research. Third, I will describe how veil-of-ignorance reasoning influences moral judgment. Fourth, I will describe the importance of third-party assessments of veil-of-ignorance reasoning.

One type of moral dilemma often faced by decision-makers and policymakers is characterized by a tradeoff between utilitarianism and competing moral concerns. Utilitarianism is the view that one ought to choose the action that produces the greatest sum of experienced well-being (or happiness) over suffering (Bentham, 1789/1983; Mill, 1863). For example, if there are an estimated 40,000 road fatalities annually in the United States that could be prevented (National Safety Council, 2019), then policymakers may want to enact policies for automated transportation that increase safety for millions of people. However, these policies need to be enacted in a way where the public will feel comfortable with the policies. Thus, it is important to understand the relevant psychology of people's moral judgments. People may value the policies that minimize the total loss of life, but they may also value competing concerns. For example, people may be averse to sacrificing their individual rights and freedoms. A deontological concern – characterized as rejecting an action that inflicts harm (e.g., sacrificing a person) in order to maximize overall welfare – would be to protect the right of the passenger, since an innocent person should not be used as an instrumental means to an end (e.g. Kant, 1797/2002). Furthermore, people may be averse to harming the passenger – prior research shows that people have an aversion to physically harming others (Rom et al., 2017; Cushman et al., 2012; Gray et al., 2012). In addition, people may feel loyalty toward some individuals over others, as shown by

research on in-group bias (e.g., Brewer, 1979). For example, people may not want their family or friends riding in such a car programmed according to utilitarian principles. Finally, at the policy level, people may value individual autonomy and freedom. According to this libertarian view, people should have the freedom to ride or buy whatever car they want. Since the government shouldn't be restricting people's right to buy or ride in the types of cars they choose, people should not be subject to government regulation on how the car should be programmed. If there are risks, then the insurance market should sort that out, and people should be held responsible for the risks they take on.

In these moral dilemmas involving the tradeoff between utilitarian concerns and competing moral concerns, what may lead ordinary people to decide one way or another? Decades of research in moral psychology has shown that underlying utilitarian judgments are deliberate thought processes about the consequences of an action, whereas the mechanisms driving competing concerns regarding individual rights and in-group loyalties are automatic, affective responses to the action (e.g., Greene et al., 2001, 2009; Cushman et al., 2006; Koenigs et al., 2007; Shenhav & Greene, 2014; Haidt, 2001).

Given an understanding of the psychology of these moral dilemmas, what would be a procedure to resolve these moral dilemmas? Let's say a decision-maker wants to make a fair, or impartial, decision. What is an exercise in impartiality that arbitrates between judgments about the consequences and feelings about the action? A seminal way of thinking about fairness is John Rawls's theory of justice (1971). I will first describe Rawls's philosophy before moving on to how a Rawlsian conception of fairness could help address the moral dilemmas described above.

Rawls's philosophical question was: What is a just society? He posited that a just society is one you would choose if you didn't know who you could be. To capture and formalize this

idea, he put forward a thought experiment involving what he calls a "veil of ignorance." Imagine the members of a society trying to decide the basic ground rules for society. Imagine you took everyone out of their ordinary lives and made them ignorant of their circumstances. What would they choose if they didn't know who they would be in a society where some people win and some people lose to varying degrees? For Rawls, a fair choice is what you would choose if you're not biased. And when you don't know who you're going to be, among all those people affected by the decision, then you don't have any information to bias your decision. According to Rawls, the choice produced by this impartial procedure is the impartial choice.

Rawls applied the veil of ignorance as a procedure for choosing the organizing principles of society. Philosophers have invoked the veil of ignorance in reasoning about sacrificial moral dilemmas (Hare, 2016). Furthermore, empirical researchers in political science have tested the veil of ignorance in the lab, having people reason as if behind the veil of ignorance in choosing how to allocate resources within a group (Frohlich, Oppenheimer & Eavey, 1987; Frohlich & Oppenheimer, 1993). In the current research, I apply the veil of ignorance to reasoning about moral dilemmas that involve the tradeoff between the greater good and competing moral concerns. As an impartial procedure, veil-of-ignorance reasoning could address the conflicts among these competing concerns.

Thus, my first research questions are the following: Does veil-of-ignorance reasoning influence moral decisions involving tradeoffs between the greater good and competing moral concerns? If so, what is this kind of influence? There is reason to believe that veil-of-ignorance reasoning would have no influence at all. Evidence for moral reasoning is rare. There's evidence that people can engage in very simple moral reasoning like thinking about costs/benefits or when given an explicit argument (Paxton & Greene, 2010; Paxton, Ungar, & Greene, 2012; Haidt,

2012; Crockett, 2013; Conway et al., 2018; Patil et al., 2018; Pizarro, Uhlmann, & Bloom, 2003; Frederick, 2005; Paxton, Bruni, & Greene, 2013; Bazerman, Loewenstein, & White, 1992). But evidence for any reasoning resembling a non-obvious philosophical argument is rare. In this work, I investigate a complex form of moral reasoning where participants aren't just given an argument and instructed to follow it. In contrast, participants are encouraged to spontaneously understand a philosophical insight, and to apply that insight to their judgments and decisions.

In the first series of experiments (Chapter 2), in collaboration with Joshua D. Greene and Max Bazerman, I develop a veil-of-ignorance reasoning intervention. We operationalized veil-of-ignorance reasoning by asking participants to consider a given moral dilemma in a different way: Imagine having an equal chance of being any one of the people affected by the decision. To give an example, in the autonomous vehicle dilemma described at the beginning, we told participants to imagine that they had an equal chance of being any one of the people affected by the policy decision. That is, there is a 1 out of 10 chance you will be the passenger and a 9 out of 10 chance you will be a pedestrian. If the law requires the car the swerve, you have 1/10 chance of dying and 9/10 chance of living. If the law requires the car to stay on its current path, you have a 9/10 chance of dying and a 1/10 chance of living. Importantly, after considering the moral dilemma in this veil-of-ignorance way, participants were given the standard moral dilemma – in this case, the standard autonomous vehicle dilemma where they decide whether to endorse a policy -- and indicate whether it is morally acceptable for a state law to require autonomous vehicles (AVs) to swerve in such a situation to save the 9 pedestrians (adapted from Bonnefon, Shariff & Rahwan, 2016).

We hypothesized that veil-of-ignorance reasoning would influence people to favor the greater good. Previewing the results in Chapter 2, we find that veil-of-ignorance reasoning

influences moral judgment, specifically in the utilitarian direction. This result generalizes across many decision domains, including decisions with real stakes. Furthermore, this result is explained by impartial reasoning, and not by alternative explanations such as anchoring, probability calculation, or perspective-taking. Why do these results matter? Broadly, these results matter for both 1) making moral decisions and 2) justifying moral decisions.

Reasoning about Moral Decisions

These results matter if one cares about the procedure by which to make difficult moral decisions. That is, if one cares about using a fair procedure, then one could use veil-of-ignorance reasoning to arrive at a fair decision. Alternatively, if one cares about the outcome – that is, promoting the greater good – then one could use veil-of-ignorance reasoning to maximize general welfare.

These results would be particularly relevant for policymakers. One could say that the job of a policymaker is to maximize the aggregate welfare of the population. If policymakers generally care about utilitarian outcomes, they could use veil-of-ignorance as a procedure to arrive at more utilitarian decisions, which benefit the overall population.

Furthermore, veil-of-ignorance reasoning could be used by citizens before giving their votes. When people make voting decisions, which often involve ethical considerations, it is ideal under democratic principles that people engage in a reasoning process before voting. Taking the example of policies regarding AVs, there are several approaches to determining the ethics for such policies. One is to simply ask people what they would want. However, there are issues with this approach, where people might give contradictory answers, and different groups of people may give contrasting answers. Another approach is to reason philosophically from first principles. However, any solution that results from first principles would be subject to the court

of public opinion in order to be acceptable to a wider population. Taking the case of the ethics of AVs, people need to agree with such principles in order to adopt (i.e., buy and ride in) such cars. This research provides a method of collecting people's opinions while simultaneously asking people to engage in a procedure with more normative constraint and philosophical insight built in to that procedure. Insofar as the deliberation of ordinary citizens matters for democratic processes (Fishkin & Luskin, 2005), the veil-of-ignorance reasoning procedure could be widely applicable in policy domains within deliberative democracies.

Justification of Moral Decisions

If a decision-maker makes a utilitarian choice, how that choice is justified matters. Third-party observers – that is, people not engaged in the decision-making but view the decision process and outcome – make social inferences about the character of the decision-maker, such as a leader or policymaker, based on that decision-maker's moral judgments (Uhlmann, Zhu & Tannenbaum, 2013; Everett et al., 2016, 2018). Third-party observers may be onlookers, or may be directly affected by the decision at hand – such as citizens who would be affected by policy decisions. Moral judgments signal commitment to cooperation, and research shows that people are less likely to trust decision-makers making utilitarian judgments, compared to deontological judgments, because utilitarian judgments show lack of concern for individual persons (Everett et al., 2016, 2018). Particularly in cases where the utilitarian decision is unpopular, utilitarian judgments may violate the norm or be perceived as causing unnecessary harm. Third-party observers often punish those who cause harm (Buckholtz et al., 2008; Martin & Cushman, 2016; Morris, MacGlashan, Littman, & Cushman, 2017), and often punish those who violate norms (Balafoutas & Nikiforakis, 2012; Carpenter & Matthews, 2009; Fehr & Fischbacher, 2004). Furthermore, in a democratic society, the opinion of the public matters in evaluating the

decisions of policymakers. Thus, third-party assessments of moral decisions are of critical importance.

In addition to being important for the individual decision-maker in reasoning about moral decisions, veil-of-ignorance reasoning also matters insofar as it could be used to justify moral decisions to the public. Veil-of-ignorance justifications for the same utilitarian decisions, compared to utilitarian justifications, could mitigate negative perceptions of decision-makers making those utilitarian judgments. In Chapter 3, I investigate the following research question: How do third-party observers view veil-of-ignorance reasoning? As the results of Chapter 3 show, veil-of-ignorance reasoning increases trust from third-party observers, particularly in situations where the utilitarian choice is unpopular. Thus, veil-of-ignorance reasoning boosts trust of the decision-maker who makes utilitarian decisions. For those policymakers whose genuine aim is to promote the greater good, even though that may be the unpopular choice, veil-of-ignorance reasoning could serve as a useful tool for justifying their utilitarian choices.

Overview

In the following chapters, I examine veil-of-ignorance reasoning in the context of moral dilemmas involving the tradeoff between the greater good and competing moral concerns. Chapter 2 examines the effect of engaging in veil-of-ignorance reasoning on one's own moral judgments. Chapter 3 investigates the fundamentally social role of moral judgments, examining the interpersonal effects of veil-of-ignorance reasoning. Taken together, this work introduces and demonstrates a novel intervention in moral judgment and justification. Veil-of-ignorance reasoning increases utilitarian judgment at the intrapsychic level, and observers perceive the decision-maker using veil-of-ignorance reasoning as warmer and more trustworthy at the

interpersonal level. These results have implications for both individual and collective decision-making, as well as how democratic deliberation could proceed regarding sensitive moral issues.

References

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences of the United States of America,* 1–4.

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 220-240.

Bentham, J. (1983). The collected works of Jeremy Bentham: Deontology, together with a table of the springs of action; and the article on utilitarianism. Oxford, England: Oxford University Press (Original work published 1789).

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307-324.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, 60(5), 930–940.

Carpenter, J., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics,* 12(3), 272–288.

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., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion,* 12(1), 2-7.

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

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200-216.

Everett, Jim A. C. and Pizarro, David A. and Crockett, M. J. (2016) Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, 145 (6). pp. 772-787.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4), 185-190.

Fishkin, J. S., & Luskin, R. C. (2005). Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta Politica*, *40*(3), 284-298.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Frohlich, N., & Oppenheimer, J. A. (1993). *Choosing Justice: An Experimental Approach to Ethical Theory* (Vol. 22). Univ of California Press.

Frohlich, N., Oppenheimer, J.A., Eavey, CL (1987). Laboratory results on Rawls's distributive justice. *British Journal of Political Science,* 17(1), 1-21.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*(2), 206-215.

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. 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), 814–834.

Haidt, J. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. (Vintage 2012).

Hare, C. (2016). Should we wish well to all?. *Philosophical Review,* 125(4), 451-472.

Kant, I. (2002). Groundwork for the metaphysics of morals. New Haven, CT: Yale University Press (Original work published 1797).

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908-911.

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition,* 147, 133–143.

Mill, J. S. (1863). Utilitarianism. London, England: Parker, Son, and Bourne.

Morris, A., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences*, 114(39), 10396–10401.

Morris, D.Z. (2016, 15 October). Mercedes-Benz's self-driving cars would choose passenger lives over bystanders. *Fortune*.

National Safety Council (2019.) Fatality Estimates. www.nsc.org

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... & Cushman, F. (2020). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*.

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, *2*(3), 511-527.

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are 'counter-intuitive'deontological judgments really counter-intuitive? An empirical reply to. *Social Cognitive and Affective Neuroscience*, *9*(9), 1368-1371.

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

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653-660.

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

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology,* 69, 44–58.

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.

Uhlmann, E. L., Zhu, L.(. L.)., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, 126(2), 326–334.

CHAPTER 2.

VEIL-OF-IGNORANCE REASONING FAVORS THE GREATER GOOD

Karen Huang

Joshua D. Greene

Max Bazerman

Published in *Proceedings of the National Academy of Sciences of U.S.A.* in October 2019

Abstract

The "veil of ignorance" is a moral reasoning device designed to promote impartial decision-making by denying decision-makers access to potentially biasing information about who will benefit most or least from the available options. Veil-of-ignorance reasoning was originally applied by philosophers and economists to foundational questions concerning the overall organization of society. Here we apply veil-of-ignorance reasoning in a more focused way to specific moral dilemmas, all of which involve a tension between the greater good and competing moral concerns. Across seven experiments (N = 6,261), four pre-registered, we find that veil-of-ignorance reasoning favors the greater good. Participants first engaged in veil-of-ignorance reasoning about a specific dilemma, asking themselves what they would want if they did not know who among those affected they would be. Participants then responded to a more conventional version of the same dilemma with a moral judgment, a policy preference, or an economic choice. Participants who first engaged in veil-of-ignorance reasoning subsequently made more utilitarian choices in response to a classic philosophical dilemma, a medical dilemma, a real donation decision between a more vs. less effective charity, and a policy decision concerning the social dilemma of autonomous vehicles. These effects depend on the impartial thinking induced by veil-of-ignorance reasoning and cannot be explained by anchoring, probabilistic reasoning, or generic perspective-taking. These studies indicate that veil-of-ignorance reasoning may be a useful tool for decision-makers who wish to make more impartial and/or socially beneficial choices.

Keywords: ethics, decision-making, policy-making, procedural justice, fairness

Significance Statement

The philosopher John Rawls aimed to identify fair governing principles by imagining people choosing their principles from behind a "veil of ignorance", without knowing their places in the social order. Across seven experiments with over 6,000 participants, we show that veil-of-ignorance reasoning leads to choices that favor the greater good. Veil-of-ignorance reasoning makes people more likely to donate to a more effective charity and to favor saving more lives in a bioethical dilemma. It also addresses the social dilemma of autonomous vehicles (AVs), aligning abstract approval of utilitarian AVs (which minimize total harm) with support for a utilitarian AV policy. These studies indicate that veil-of-ignorance reasoning may be used to promote decision-making that is more impartial and socially beneficial.

VEIL-OF-IGNORANCE REASONING FAVORS THE GREATER GOOD

The philosopher John Rawls proposed a famous thought experiment, aimed at identifying the governing principles of a just society (1). Rawls imagined decision-makers who've been denied all knowledge of their personal circumstances. They don't know whether they, as individuals, are rich or poor, healthy or ill, or in possession of special talents or abilities. Nor do they know the social groups to which they belong, as defined by race, class, gender, etc. The decision-makers are assumed to be purely self-interested, but their decisions are constrained by the absence of information that they could use to select principles favorable to their personal circumstances. Rawls referred to this epistemically restricted state as being behind a "veil of ignorance".

Rawls conceived of this hypothetical decision as a device for helping people in the real world think more clearly and impartially about the organizing principles of society. A just social order, he argued, is one that selfish people would choose if they were constrained to choose impartially, in the absence of potentially biasing information. Some empirical researchers have adapted Rawls' thought experiment to the lab, asking ordinary people to evaluate candidate organizing principles by engaging in veil-of-ignorance reasoning (2). Here, we depart from the conventional use of the veil of ignorance as a device for thinking about the general organization of society. Instead, we apply veil-of-ignorance reasoning to a set of more specific moral and social dilemmas. These dilemmas, though more restricted in scope than Rawls' foundational dilemma, are nevertheless of broad social significance, with life-and-death consequences in the domains of healthcare, international aid, and automated transportation. What effect, if any, does veil-of-ignorance reasoning have on people's responses to such dilemmas?

We predict that veil-of-ignorance reasoning will cause people to make more utilitarian judgments, by which we mean judgments that maximize collective welfare.[1] This result is by no means guaranteed, as there are reasons to think that veil-of-ignorance reasoning could have the opposite effect, or no effect at all. Rawls was one of utilitarianism's leading critics (1), suggesting that veil-of-ignorance reasoning might reduce utilitarian judgment. And even if veil-of-ignorance reasoning were to support utilitarian choices, it's possible that people's ordinary responses to moral dilemmas implicitly reflect the lessons of veil-of-ignorance reasoning, such that engaging in explicit veil-of-ignorance reasoning would have no additional effect.

Despite Rawls' renown as a critic of utilitarianism, our predicted results are not necessarily incompatible with Rawls' philosophy, as the dilemmas employed here are not designed to distinguish between a utilitarian decision principle and Rawls' "maximin" principle (1, 3). Rawls' "maximin" principle favors whatever outcome maximizes the welfare of the least well-off person. For example, in the well-known footbridge case (see below), the least well-off people under each option experience equally bad outcomes, namely death by trolley. Thus, one might expect a Rawlsian to be indifferent between the two options or to favor the utilitarian

[1] Following convention in the psychology and cognitive neuroscience literatures, we refer to these judgments as "utilitarian", but one could also call them "consequentialist", a label that does not assume that saving more lives necessarily implies greater overall happiness. In addition, in calling these judgments "utilitarian" we are not claiming that the people who make them are in any general way committed to utilitarianism (14).

option, invoking the utilitarian principle as a secondary consideration.[2] Our predictions are, however, most closely aligned with the ideas of Rawls' contemporary and critic, John Harsanyi, an influential economist who independently conceived of veil-of-ignorance reasoning and argued that it provides a decision-theoretic foundation for a utilitarian social philosophy (4-5).

To illustrate our application of veil-of-ignorance reasoning, consider the aforementioned *footbridge* dilemma, in which one can save five people in the path of a runaway trolley by pushing a person off of a footbridge and into the trolley's path (6). The utilitarian option is to push, as this maximizes the number of lives saved. However, relatively few people favor the utilitarian option in this case, a result of negative affective responses to this actively, directly, and intentionally harmful action (7-8).

What effect might veil-of-ignorance reasoning have on a case such as this? Following Hare (9), you might imagine that you are going to be one of the six people affected by the footbridge decision (one of the five on the tracks or the one who could be pushed). You might assume that you have even odds[3] of being any one of them. (We note that Rawls' version of the

---

[2] Alternatively, one might expect a Rawlsian to reject the utilitarian option on the grounds that, "each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override" (1, p. 3). See, for example, Sandel (32, Chapter 2).

[3] Our understanding of "even odds" is motivated by a principle of impartiality, which provides the motivation for veil-of-ignorance reasoning. Veil-of-ignorance reasoning, as applied here, assigns even odds of being each person affected by the specific decision in question. One could, however, incorporate other types of probabilistic information, such as the odds that a decision-maker would, in real life, occupy one position rather than another (e.g. being on a footbridge vs.

veil of ignorance assumes unknown odds rather than even odds.[4]) Would you, from a purely self-interested perspective, want the decision-maker to push, giving you a 5 out of 6 chance of living? Or would you want the decision-maker to not push, giving you a 1 out 6 chance of living? Here,

---

on trolley tracks). Incorporating such probabilities could be highly relevant for other purposes, but doing so would weaken the connection between VOI reasoning and impartiality, which is essential for present purposes. See Discussion.

[4] In Rawls' version of the veil of ignorance, the decision makers assume that their odds of occupying any particular position in society are unknown. Following Harsanyi, we make an "equi-probability" assumption, instructing participants that they have even odds of being each of the people affected by the decision. In other words, we make the decision a matter of "risk", rather than "ambiguity" (33). We do this for two related reasons. First, it is not our purpose to examine the effect of Rawls' specific version of veil-of-ignorance reasoning, but rather to determine whether some kind of veil-of-ignorance reasoning, embodying a principle of impartiality, can influence moral judgment and, more specifically, induce people to make choices that more reliably favor the greater good. Second, and more positively, we believe that Rawls' assumption of unknown odds, rather than even odds, makes little sense given the rationale behind the veil-of-ignorance thought experiment: If the function of the veil is to constrain the decision-makers into perfect impartiality, then why not give the interests of each person exactly equal weight by giving the decision-makers exactly even odds of being each person? We know of no compelling justification for assuming that the odds are anything other than exactly equal. (See 12, p. 383-385).

we suspect (and confirm in Study 1) that most people prefer that the decision-maker push, increasing one's odds of living.

But, what, if anything, does this imply about the ethics of the original moral dilemma? As noted above, most people say that it would be wrong to push the man off the footbridge. And yet a Rawlsian (or Harsanyian) argument seems to imply that pushing is the fairer and more just option: It's what those affected by the decision would want if they didn't know which positions they would occupy. In the studies presented here, we follow a two-stage procedure, as suggested by the foregoing argument. First, participants consider a veil-of-ignorance version of a moral dilemma, reporting on what they would want a decision-maker to do if they did not know who they would be among those affected by the decision. Second, participants respond to a standard version of the same dilemma, reporting on the moral acceptability of the proposed action. (In Study 3 participants in the second stage make a real-stakes choice instead of a hypothetical judgment.) The question, then, is whether engaging in veil-of-ignorance reasoning in the first stage influences ordinary moral judgment in the second stage. To be clear, we are not comparing the responses in the veil-of-ignorance exercise with responses to the standard moral dilemma. Instead, we investigate the influence of veil-of-ignorance reasoning, induced through the veil-of-ignorance exercise, on responses to the standard moral dilemma.

In applying veil-of-ignorance reasoning, we are not only attempting to influence moral judgment, but to do so through a kind of *reasoning*. By this, we mean that the influence occurs through a conscious valuation process that is constrained by a need for *consistency*—either with a general principle, with related judgments, or both (10). We expect that, in the second stage, participants will explicitly consider the normative relationship between the moral judgment they are currently making and the judgment that they made in the first stage, a self-interested decision

from behind a veil of ignorance. Moreover, we expect that they will be inclined to make their current moral judgment consistent with their prior veil-of-ignorance judgment. In other words, we predict that participants will think something like this: "If I didn't know which of the six people I was going to be, I would want the decision-maker to push. But when I think about pushing, it feels wrong. Nevertheless, if pushing is what I'd want from an impartial perspective, not knowing who I was going to be, then perhaps it really is the right thing to do, even if it feels wrong."

Thus, through this procedure, we encourage participants to absorb the philosophical insight of the veil-of-ignorance thought experiment and apply this idea in their subsequent judgments. We note that the predicted effect of veil-of-ignorance reasoning would provide evidence for an especially complex form of moral reasoning. This would be notable, in part, because there is little evidence for moral reasoning beyond the application of simple rules such as simple cost-benefit reasoning (11-16).

Beyond moral psychology, the effects of veil-of-ignorance reasoning may be of practical significance, as people's responses to moral dilemmas are often conflicted and carry significant social costs (12). Consider, for example, the social dilemma of autonomous vehicles (AVs) (17), featuring a tradeoff between the safety of AV passengers and that of others, such as pedestrians. As a Mercedez-Benz executive discovered (18), an AV that prioritizes the safety of its passengers will be criticized for devaluing the lives of others. But a "utilitarian" AV that values all lives equally will be criticized for its willingness to sacrifice its passengers. This paradox is reflected in the judgments of ordinary people, who tend to approve of utilitarian AVs in principle, but disapprove of enforcing utilitarian regulations for AVs (17). Likewise, people may feel conflicted about bioethical policies or charities that maximize good outcomes, with costs to

specific, identifiable parties (12). We ask whether the impartial perspective encouraged by veil-of-ignorance reasoning can influence people's responses to these and other dilemmas. This research does not assume that such an influence would be desirable. However, to the extent that people value impartial decision procedures or collective well-being, such an influence would be significant.

Across seven studies, we investigate the influence of veil-of-ignorance reasoning on moral judgment. We begin with the *footbridge* dilemma (Study 1) because it is familiar and well characterized. In subsequent studies we employ cases with more direct application, including a decision with real financial stakes (Study 3). Across all cases, we predict that participants' responses to the veil-of-ignorance versions will tend to be utilitarian, simply because this maximizes their odds of a good outcome. Critically, we expect participants to align their later moral judgments with their prior veil-of-ignorance preferences, causing them to make more utilitarian moral judgments, favoring the greater good, as compared to control participants who have not engaged in veil-of-ignorance reasoning.

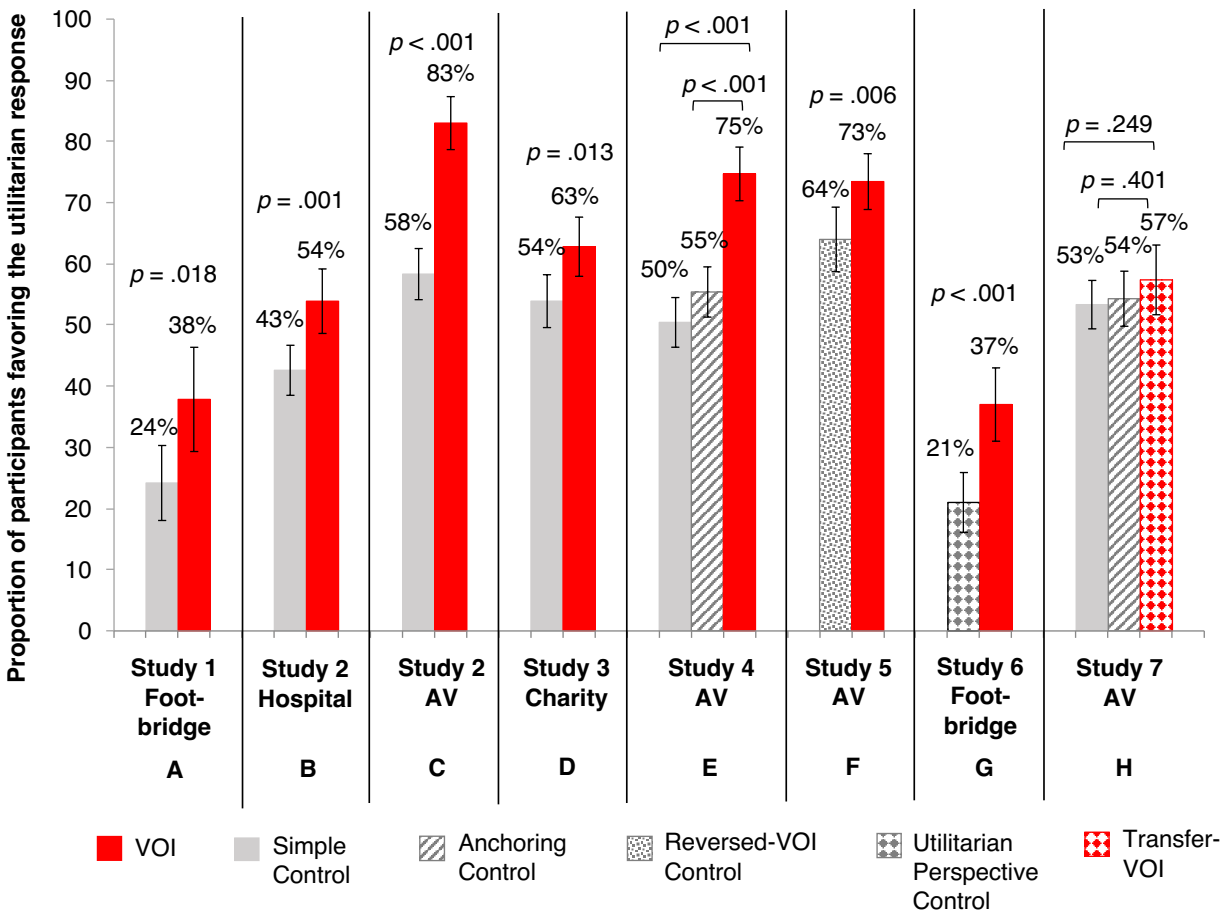
Experimental Designs and Results

Study 1 (N=264) employs the *footbridge* dilemma: The decision-maker can save five people in the path of a runaway trolley by pushing a person off of a footbridge and into the trolley's path (6). The utilitarian option is to push, as this saves more lives, but relatively few favor this option, largely due to negative affective responses (7-8). Study 1's veil-of-ignorance (VOI) condition employs the two-stage procedure described above. In the VOI version (stage 1), participants imagined having equal odds of being each of the six people affected by the decision: the five people on the tracks and the 6th person who could be pushed. Participants were asked whether they would want the decision-maker to push, giving the participant a 5 out of 6 chance

of living, or not push, giving the participant a 1 out of 6 chance of living. Here (and in all subsequent studies) most participants gave utilitarian responses to the VOI version of the dilemma. In stage 2 of the VOI condition, participants responded to the standard footbridge dilemma as the decision-maker, evaluating the moral acceptability of pushing, using a dichotomous response and a scale rating. In the control condition there is only one stage, wherein participants respond to the standard dilemma. Critically, the key dependent measures for both conditions were responses to the standard dilemma. In other words, we ask whether first completing the VOI version affects subsequent responses to the standard dilemma.

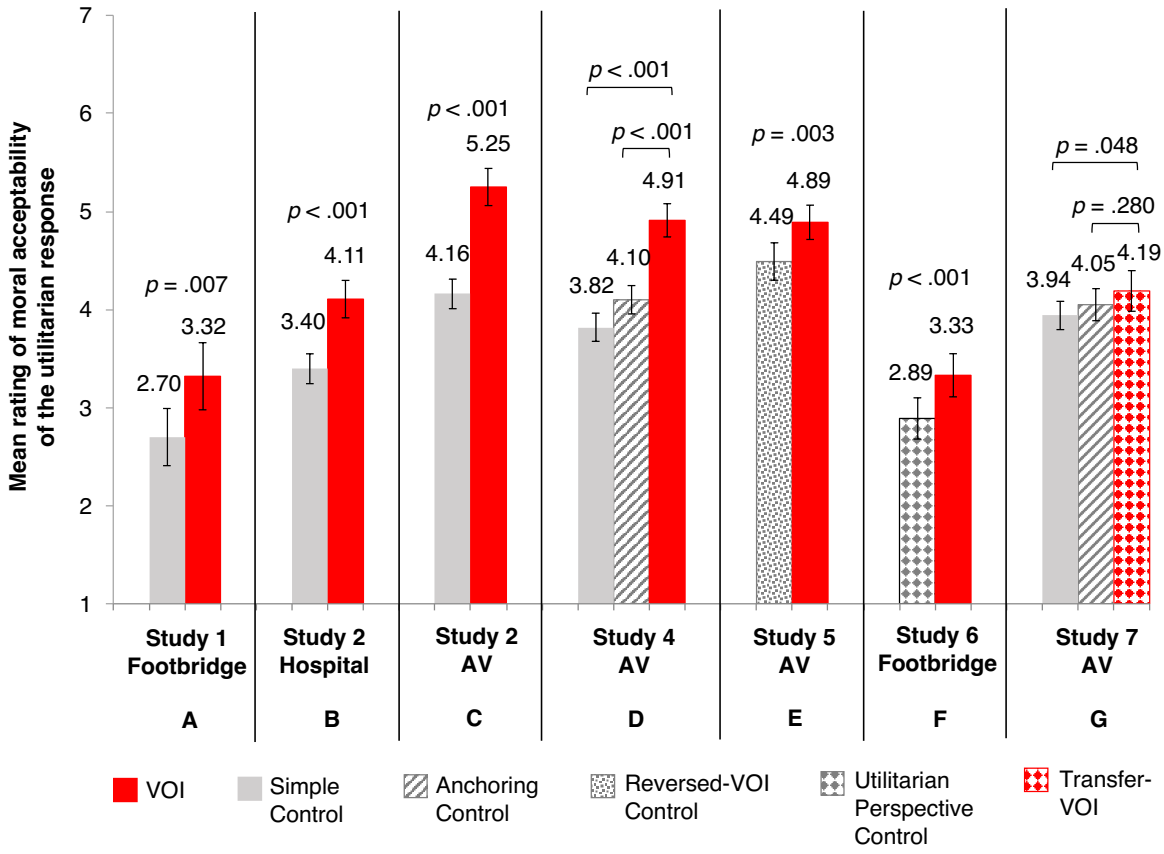
As predicted, participants in the VOI condition gave more utilitarian responses to the standard footbridge dilemma (38%, [*95% CI:* 30%, 47%]), as compared to control participants (24%, [*95% CI:* 18%, 32%]; logistic regression, $p = .018$). Likewise, participants rated the utilitarian response as more morally acceptable in the VOI condition ($M = 3.32$, $SD = 2.05$) as compared to the control condition ($M = 2.70$, $SD = 1.65$) (linear regression, $t(262) = 2.74$, $p = .007$). (Fig. 1.1A-1.2A. See SI Appendix for detailed procedures and results for all studies, including results without excluding participants who failed attention and/or comprehension checks. See Tables S1-S4. See Studies 4-6 for additional control conditions.)

FIGURE 1.1. Dichotomous responses for all studies (N = 6,261).



p-values from logistic regression. Error bars indicate 95% CI. (A) Study 1 footbridge case; (B) Study 2 hospital case; (C) Study 2 AV case; (D) Study 3 charity case; (E) Study 4 AV case; (F) Study 5 AV case; (G) Study 6 footbridge case; (H) Study 7 AV case.

FIGURE 1.2. Scale responses for Studies 1-2, 4-7 (N = 5,428).



p-values from linear regression. Error bars indicate 95% CI. (A) Study 1 footbridge case; (B) Study 2 hospital case; (C) Study 2 AV case; (D) Study 4 AV case; (E) Study 5 AV case; (F) Study 6 footbridge case, (G) Study 7 AV case.

Study 2 (*N*=894) employs dilemmas concerning bioethics and the ethics of AVs. In the bioethics case, participants considered taking oxygen away from a single hospital patient to enable the surgeries of 9 incoming earthquake victims. In the VOI version of the bioethical case, participants were asked how they would want the oxygen to be allocated if they knew they had a 1 in 10 chance of being the single patient and a 9 in 10 chance of being one of the 9 earthquake

victims (19). In the AV policy case, participants considered whether AVs should be required to minimize the total loss of life (i.e., be utilitarian), for example, saving 9 pedestrians by swerving into a wall, but killing the AV's passenger (17). In the VOI AV case, participants were asked if they would want the AV to swerve into the wall given a 1 in 10 chance of being in the AV and 9 in 10 chance of being one of the 9 pedestrians. As predicted, participants in the VOI condition gave more utilitarian responses to the standard bioethical dilemma (54%, [*95% CI:* 49%, 59%]), as compared to control (43%, [*95% CI:* 39%, 47%], $p = .001$). Likewise, participants in the VOI condition gave more utilitarian responses to the standard AV dilemma (83%, [*95% CI:* 79%, 87%]), as compared to control (58%, [*95% CI:* 54%, 62%], $p < .001$) (Figs. 1.1B-C). The rating scale results showed a similar pattern: Participants in the VOI condition reported taking the patient off oxygen as more morally acceptable ($M = 4.11$, $SD = 1.90$) compared to participants in the control condition ($M = 3.40$, $SD = 1.75$; $t(892) = 5.71$, $p < .001$). Similarly, participants in the VOI condition reported swerving as more morally acceptable ($M = 5.25$, $SD = 1.74$) compared to participants in the control condition ($M = 4.16$, $SD = 1.83$; $t(892) = 8.86$, $p < .001$) (Figs. 1.2B-C).

Study 3 (N=833) examines a real-stakes setting: charitable donations. U.S. participants chose to donate $200 to one of two real charities (with one randomly selected participant's decision determining the actual outcome). Donating to the more effective/utilitarian charity can be expected to cure two people of blindness in India. Donating to the other charity can be expected to cure one person of blindness in the U.S. In the VOI condition, participants were first asked where they would want the $200 to go if they knew they had a 1 in 3 chance of being an American who would be cured by a donation to the U.S. charity and a 2 in 3 chance of being an Indian who would be cured by a donation to the Indian charity. They then made their real

donation decisions. As predicted, participants in the VOI condition more often chose to donate to the more effective/utilitarian charity (63%, [*95% CI:* 57%, 68%]), as compared to control participants, who only made the real donation decision (54%, [*95% CI:* 50%, 58%], $p = .013$) (Fig. 1.1D).

We've hypothesized that the effects observed in Studies 1-3 are due to the influence of VOI reasoning itself, inducing a more impartial mindset that promotes concern for the greater good. An alternative explanation is that participants in the VOI condition are simply "anchoring" on their prior utilitarian responses to the VOI versions, giving subsequent utilitarian responses due to a generic tendency toward consistency. (By "anchoring," we mean giving a subsequent response that is the same as a prior response, rather than referring to anchoring in the more specific sense defined by Tversky and Kahneman (20)). Our hypothesis also appeals to a desire for consistency, but we hypothesize that participants are engaging in a specific kind of moral reasoning, mirroring the reasoning of Rawls and Harsanyi, whereby participants perceive a connection between what is morally defensible and what they would want if they did not know whom they would be among those affected by the decision.

Study 4 (N=1,574; pre-registered) tests this alternative hypothesis while replicating the VOI effect using the AV dilemma. Study 4 employs an additional anchoring control condition in which participants first respond to a standard (non-VOI) dilemma that reliably elicits utilitarian responses. This non-VOI dilemma asks participants whether they favor destroying a sculpture to save the lives of two people. As predicted, participants in the VOI condition gave more utilitarian responses (75%, [*95% CI:* 70%, 79%]), as compared to simple control (50%, [*95% CI:* 46%, 54%]; $p < .001$), and anchoring control (55%, [*95% CI:* 51%, 60%]; $p < .001$). Likewise, participants rated the utilitarian policy as more morally acceptable in the VOI condition (M =

4.91, *SD* = 1.71), as compared to those in the simple control condition (*M* = 3.82, *SD* = 1.86; *t*(1571) = 9.65, *p* < .001), and as compared to those in the anchoring control condition (*M* = 4.10, *SD* = 1.72; *t*(1571) = 7.13, *p* < .001). (Figs. 1.1E, 1.2D.)

Further alternative explanations appeal to features of the VOI dilemma not captured by Study 4's anchoring control condition. More specifically, participants in the VOI condition are asked to engage in numerical reasoning and probabilistic reasoning. This could, perhaps, induce a mindset favoring expected utility calculations, as prescribed by rational choice models of decision-making (21). The VOI condition also asks participants to engage in a limited kind of perspective-taking (22), as participants are asked to consider the effects of the decision on all affected. These and other task features could potentially induce more utilitarian responses to the standard dilemma. These features are essential to VOI reasoning, but a further essential feature of VOI reasoning (as implemented here) is its relation to impartiality, whereby one has an equal probability of being each person affected.

Thus, Study 5 (*N*=735; pre-registered), which again uses the AV dilemma, employs a more stringent control condition in which participants first respond to a modified VOI dilemma in which the probabilities are *reversed*. That is, one has a 9 in 10 chance of being the single person in the AV and a 1 in 10 chance of being one of the 9 pedestrians in the AV's path. Reversing the probabilities disconnects VOI reasoning from impartiality, as one no longer has an equal probability of being each person affected. Because it is the impartiality of VOI reasoning that gives it its moral force, we do not expect this reversed-VOI reasoning to have the same effect. As predicted, participants in the VOI condition gave more utilitarian responses (73%, [*95% CI:* 69%, 78%]), as compared to the reversed-VOI control condition (64%, [*95% CI:* 59%, 69%]; *p* = .006). Likewise, participants rated the utilitarian policy as more morally acceptable in

the VOI condition (M = 4.89, SD = 1.77), as compared to those in the reversed-VOI control condition (M = 4.49, SD = 1.80; $t(733) = 3.03$, $p = .003$). (Figs. 1.1F, 1.2E.)

Study 6 (N=571; pre-registered) aims to rule out a further alternative explanation for the VOI effect: The effect of VOI may simply be due to "narrow anchoring", whereby giving a specific response to a specific dilemma in the first phase (involving the VOI exercise) then causes the participant to give the same response to the same dilemma in the second phase. We therefore employ an additional control condition in which we expect participants to give utilitarian responses to the dilemma in the first phase, but not in the second phase when they encounter the standard version of the dilemma. This additional control condition asks participants to adopt the perspective of a person (named Joe) who is strongly committed to utilitarianism and who is therefore willing to sacrifice the interests of some individuals for the greater good of others. We predicted that participants would tend to give utilitarian responses to the footbridge dilemma when asked to adopt Joe's utilitarian perspective during the first phase of the control condition. But we predicted that participants would tend not to give utilitarian responses in the second phase of the control condition, when they are no longer instructed to adopt Joe's perspective and are instead simply responding to the footbridge dilemma in its standard form. We hypothesized that participants in the VOI condition, compared to those in the utilitarian-perspective control condition (in which they adopt Joe's perspective in the first phase), would be more likely to make the utilitarian judgment in response to the standard footbridge case in the second phase.

As predicted, participants in the VOI condition gave more utilitarian responses to the standard footbridge dilemma (37%, [*95% CI:* 31%, 43%]), as compared to the utilitarian-perspective control condition (21%, [*95% CI:* 16%, 25%]; $p < .001$). Likewise, participants rated

the utilitarian judgment in the standard footbridge dilemma as more morally acceptable in the VOI condition (M = 3.33, SD = 1.86), as compared to in the utilitarian-perspective control condition (M = 2.89, SD = 1.83; $t(569) = 2.83$, p = .005). (Figs. 1.1G, 1.2F.) These results indicate that the effect of VOI reasoning cannot be explained by a tendency to anchor on a specific response to a specific dilemma.

Finally, Study 7 (N=1,390; pre-registered) asks whether VOI reasoning transfers across cases. Participants in the transfer-VOI condition first responded to two VOI cases that are not tightly matched to the AV case, before responding to the standard AV case. Study 7 employed a simple control condition as in Study 1 along with a two-dilemma anchoring control condition similar to that of Study 4. We predicted that participants in the transfer-VOI condition would be more likely to make the utilitarian judgment in the standard AV case, relative to the two control conditions. Contrary to our predictions, we found no significant differences in participants' responses to the standard AV case in the transfer-VOI condition (57%; [*95% CI:* 52%, 63%]), as compared to simple control (53%; [*95% CI:* 49%, 57%]; p =.249), and anchoring control (54%; [*95% CI:* 50%, 59%]) p = .401). For the scale measure, we found that participants rated the utilitarian response as more morally acceptable in the transfer-VOI condition (M = 4.19, SD = 1.80), as compared to those in the simple control condition (M = 3.94, SD = 1.86; $t(1387) = 1.98$, p = .048). However, there were no significant differences in participants' scale responses between the transfer-VOI condition and the anchoring control condition (M = 4.05, SD = 1.78; $t(1387)= 1.08$, p = .280). (Figs. 1.1H, 1.2G.) These results establish a boundary condition on the effect of VOI reasoning. We note, however, that further training in VOI reasoning may enable people to transcend this boundary.

Discussion

Across multiple studies we show that veil-of-ignorance reasoning influences responses to moral dilemmas, encouraging responses that favor the greater good. These effects were observed in response to a classic philosophical dilemma, a bioethical dilemma, real-stakes decisions concerning charitable donations, and in judgments concerning policies for AVs. While previous research indicates net disapproval of utilitarian regulation of AVs (17), here we find that veil-of-ignorance reasoning shifts approval to as high as 83% (Fig. 1.1C, 1.1E-F). (We note that these findings address attitudes toward regulation, but not individual consumption.) The effect of veil-of-ignorance reasoning was replicated in three pre-registered studies. These studies showed that this effect cannot be explained by a generic tendency toward consistency (general anchoring) or by a tendency to give the same response to a subsequent version of the same dilemma (narrow anchoring). Most notably, we show that the effect of veil-of-ignorance reasoning depends critically on assigning probabilities aligned with a principle of impartiality.

Arguably the most central debate in the field of moral psychology concerns whether and to what extent people's judgments are shaped by intuition as opposed to reason or deliberation (10-12, 16). There is ample evidence for the influence of intuition, while evidence for effectual moral reasoning is more limited (16). Beyond simple cost-benefit utilitarian reasoning (11-15), people's judgments are influenced by explicit encouragement to think rationally (23) and by simple debunking arguments (11). Performance on the Cognitive Reflection Test (24) is correlated with utilitarian judgment (11, 15, 25), and exposure to this test can boost utilitarian judgment (11, 15), but this appears to simply shift the balance between intuitive responding and utilitarian reasoning, rather than eliciting a more complex form of reasoning. Closer to the

present research is the use of joint (versus separate) evaluation, which induces participants to make a pair of judgments based on a common standard of evaluation (26).

Here we provide evidence for effectual moral reasoning in ordinary people that is arguably more complex than any previously documented. The VOI condition requires a kind of spontaneous "micro-philosophizing" to produce its effect, recapitulating the insights of Rawls and Harsanyi, who perceived an equivalence between self-interested decisions made from behind a veil of ignorance and justifiable moral decisions. Here, participants are not presented with an explicit argument. Instead, they're given the raw materials with which to construct and apply an argument of their own making. Participants in the VOI condition are not told that there is a normative relationship between the VOI exercise and the subsequent judgment, but many participants nevertheless perceive such a relationship. Without explicit instruction, they perceive that a self-interested choice made from behind a veil of ignorance is an impartial choice, and therefore likely to be a morally good choice when the veil is absent. And, once again, this effect disappears when the probabilities are reversed, indicating that participants are sensitive to whether the veil of ignorance is fostering a kind of impartial thinking. We are not claiming, of course, that people engage in this kind of moral reasoning under ordinary circumstances. But these findings indicate that ordinary people can actively engage in a rather sophisticated kind of moral reasoning with no special training and minimal prompting.

We wish to note several limitations of the present findings. First, we do not claim that veil-of-ignorance reasoning must always promote utilitarian judgment. In particular, we are not attempting to resolve the debate between Rawls and Harsanyi over whether veil-of-ignorance reasoning favors a utilitarian principle over Rawls' "maximin" principle, as the dilemmas employed here do not distinguish between them. Second, we note that our studies aimed at ruling

out competing explanations (Studies 4-6), as well as our study establishing limited generalization (Study 7), all used either the footbridge case or the AV policy case. Nevertheless, the most parsimonious interpretation of the evidence is that the VOI effect observed for these cases is psychologically similar to those observed for other cases. Finally, we note that in Study 5 the proportion of utilitarian judgments in the standard AV case, following the reversed-VOI exercise, was relatively high (64%), as compared to the stand-alone AV cases tested in Studies 2, 4, and 7 (58%, 50%, 53%, respectively). Thus, it is possible that component features of the VOI exercise, such as the engagement of probabilistic reasoning, may play some role in promoting subsequent utilitarian judgment. Alternatively, it could be that engaging in reversed-VOI reasoning is enough to prompt some participants to engage in standard VOI reasoning.

We wish to emphasize that these findings, by themselves, neither assume nor demonstrate that the effects of veil-of-ignorance reasoning are desirable. Nevertheless, these findings may have significant implications when combined with certain widely shared moral values (27). For those who regard promoting the greater good as an important moral goal, the present findings suggest a useful tool for encouraging people to make decisions that promote the greater good. Likewise, this approach may be of interest to those who value impartial procedures, independent of any commitment to maximizing aggregate well-being. Lawmakers and policymakers who value impartial procedures and/or promoting the greater good may find veil-of-ignorance reasoning to be a useful tool for making complex social decisions and justifying the decisions they have made.

Here, it is worth noting connections between veil-of-ignorance reasoning and other policy tools. For example, others have used structured decision procedures to encourage a more impartial or detached perspective on matters of distributive justice, including one of Rawls' and

Haransyi's central concerns, the (re)distribution of wealth (28, 29). We also note that decision procedures similar to veil-of-ignorance reasoning could be used for evaluating policies from a less impartial perspective. This might involve rejecting the equi-probability assumption that we (following Haransyi) have used in our VOI reasoning procedures. For example, if one expects to have a high probability of being a passenger in an AV, but one expects to have a low probability of being a pedestrian who could be threatened by AVs, then one might wish to incorporate these individualized probabilities into a decision procedure that in some ways resembles VOI reasoning. However, if the aim is to incorporate a principle of impartiality into one's decision procedure then, in our view, it makes the most sense to adopt Harsanyi's equi-probability assumption.

Veil-of-ignorance reasoning may be most useful when people are forced to make, and publicly justify, decisions involving difficult tradeoffs. Decisions that promote the greater good may involve emotionally aversive sacrifices and/or an unwillingness to allocate resources based on personal or group-based loyalties (12, 30). Observers tend to be highly suspicious of people who make utilitarian decisions of this kind (31). Indeed, we found in Study 6 that participants who adopted a utilitarian perspective in the first phase (because we asked them to) were not especially likely to maintain that perspective in the second phase. How, then, can decision-makers whose genuine aim is to promote the greater good advance policies that are so readily perceived as anti-social or disloyal? We suggest that veil-of-ignorance reasoning can help people—both decision-makers and observers—distinguish between policies that are truly anti-social or culpably disloyal from socially beneficial policies that are simply aversive. Emotionally uncomfortable tradeoffs may seem more acceptable if one can credibly say, "This is what I would want for myself if I did not know who I was going to be."

Where there is conflict—either within or between people—about important moral decisions, mechanisms that might promote agreement are worth considering. Veil-of-ignorance reasoning may be especially useful because it influences people's judgments without telling them how to think or what to value. Nor does it manipulate people through non-conscious influences or the restriction of information. Instead, it is Socratic. It openly and transparently asks people to consider their decisions from a different perspective, leaving it up to decision-makers to determine whether that perspective is valuable. Across a range of decisions, from bioethics to philanthropy to machine ethics, people seem to find this perspective illuminating.

Materials and Methods

The procedures and materials for all studies were reviewed and approved by the Harvard University Institutional Review Board. All participants provided informed consent. We have uploaded all study materials, preregistrations, raw data, and analyses code on Open Science Framework, accessible at the following link: https://osf.io/6xyct/?view_only=13cafb7e7c654c1e84afcf6401716f7b.

All statistical analyses were conducted using R statistical software.

Study 1. In both conditions, participants entered their Amazon Mechanical Turk (MTurk) IDs and completed an attention check. Participants who failed the attention check were excluded from analysis. In the control condition, participants responded to the standard version of the footbridge dilemma with a dichotomous choice ("Is it morally acceptable for you to push the second person on to the tracks in order to save the five workmen?") and a scale item ("To what extent is this action morally acceptable?").

In the VOI condition, participants first responded to a VOI version of the footbridge dilemma in which the participant is asked to imagine having an equal probability of being each

of the six people affected by the decision. They then indicated what they would like the decision-maker to do using a dichotomous choice ("Do you want the decision-maker to push or not push?") and a scale measure (To what extent do you want the decision-maker to push?). The VOI version of the footbridge dilemma was then followed by the standard version, as used in the control condition. Once again, in both conditions our primary dependent measures are responses to the standard footbridge dilemma.

In both conditions, after participants responded to the dilemma(s), they completed comprehension checks (one for each dilemma). We hypothesized *a priori* that only participants who engaged in careful and attentive thinking would be affected by the VOI manipulation, and therefore we excluded from analysis participants who failed at least one attention check or comprehension check. For all studies, we report exclusion rates by condition, and we present results including all participants. This provides assurance that our conclusions are not artifacts of differential rates of exclusion across conditions (See Tables S1-S4 in SI Appendix.)

At the end of their sessions, participants in the VOI condition were asked about whether their responses to the standard case were influenced by the VOI exercise. All participants assessed their prior familiarity with the testing materials, supplied their age and gender, and were asked for general comments.

Study 2. Procedures followed those of Study 1, but using the hospital and AV dilemmas. Participants were assigned to the same condition for both dilemmas, and all participants responded to both the hospital and AV dilemmas. In the VOI condition, participants were always presented with the VOI version of a dilemma immediately prior to the standard version of that dilemma. In both conditions, and in both stages of the VOI condition, the order of the dilemmas (AV vs. hospital) was counterbalanced.

In the standard AV case, participants responded to the dichotomous measure ("Is it morally acceptable for a state law to require autonomous vehicles to swerve in such a situation to save the 9 pedestrians?") followed by the corresponding scale measure. In the standard hospital case, participants responded to the dichotomous measure ("Is it morally acceptable for you to take the patient at the hospital off oxygen?") followed by the scale measure.

In the VOI version of the AV dilemma, participants responded to a dichotomous measure ("Please respond from a purely self-interested perspective: Would you want to be in a state where the law requires autonomous vehicles to swerve in such a situation?") and a corresponding scale measure. Likewise, in the VOI hospital dilemma, participants responded to a dichotomous measure ("Please respond from a purely self-interested perspective: Do you want the ethics committee to take the patient off oxygen?") and a corresponding scale measure. In Study 1, we expected participants responding to the VOI dilemma to respond from a self-interested perspective when asked what they would want the decision-maker to do, as the decision-maker's choice would determine whether they would probably live or probably die. In Study 2 and all subsequent studies we explicitly instructed participants responding to VOI dilemmas to respond from a self-interested perspective. This change only affects the VOI exercise and not the standard moral dilemma used in the second phase of the VOI condition.

Study 3. Here participants chose between a more effective and less effective charitable donation. We presented all participants with descriptions of two real charities, although the charity names were not given. Donating $200 to the Indian charity would fund cataract surgeries for 2 people in India. Each of the 2 people need surgery in one eye, and without the surgery, each will go permanently blind in one eye. Donating $200 to the U.S. charity would contribute to funding surgery for a person living in the U.S. Here the recipient is going blind from an eye

disease called pars planitis, and without the surgery, this person will go permanently blind in one eye. These charities were designated “Charity A” and “Charity B”, with label/order counterbalanced.

The Indian charity is more effective because the same level of donation cures two people instead of one. More precisely, a donation to the U.S. charity is expected to *contribute to* the curing of a single person. However, for the purposes of assigning probabilities in the VOI version, we assumed that the single person would be cured as a result of the donation. This is a conservative assumption, since it increases the appeal of the U.S. charity, and our prediction is that considering the VOI version of the charity dilemma will make people less likely to choose the U.S. charity in the subsequent decision. The two charities differ in other ways, most notably in the nationality of beneficiaries, but that is by design, as the most effective charities tend to benefit people in poorer nations, where funds go further.

For the real donation decision employed in both conditions, we told participants, “We will actually make a $200 donation and one randomly chosen participant’s decision will determine where the $200 goes.” They were then asked, “To which charity do you want to donate the $200?” with the options, “I want to donate the $200 to Charity A” or “I want to donate the $200 to Charity B.”

In the VOI condition, participants were first presented with a VOI version of the charity dilemma (hypothetical) which used this prompt: “Please respond from a purely self-interested perspective: To which charity do you want the decision-maker to donate the $200?” There was no continuous measure in Study 3. Participants in the VOI condition then made their real charity decision, as in the control condition. We compared the real charity decisions between the VOI and control conditions.

Study 4. Study 4 methods (pre-registered; #6425 on AsPredicted.org) follow those used for the AV policy case in Study 2, but with no accompanying bioethical dilemma, and, critically, with the inclusion of a new anchoring control condition. In the anchoring control condition, participants first responded to the *sculpture* case (which reliably elicits utilitarian responses) before responding to the AV policy case.

Study 5. Study 5 methods (pre-registered; #11473 on AsPredicted.org) follow that of Studies 2 and 4, using variants of the AV policy dilemma for both the VOI condition and the reversed-VOI control condition. As before, in the VOI condition, participants completed a VOI version of the AV dilemma, in which they imagined having a 9 in 10 chance of being one of the 9 people in the path of the AV and a 1 in 10 chance of being the single passenger in the AV. In the reversed-VOI version of the AV dilemma, they imagined having a 9 in 10 chance of being the single passenger in the AV and a 1 in 10 chance of being one of the nine pedestrians in the AV's path.

Study 6. In Study 6 (pre-registered; #27268 on AsPredicted.org), participants responded to the standard footbridge case, as in Study 1, as the main dependent variable. In the VOI condition, participants first responded to the VOI version of the footbridge case, prior to the standard footbridge case, as in Study 1. In the utilitarian-perspective control condition, participants responded to a version of the footbridge case in which they were instructed to adopt the perspective of a person committed to utilitarianism. This was then followed by the standard footbridge case, as in the VOI condition.

Study 7. In Study 7 (pre-registered, #6157 on AsPredicted.org), as in Studies 4-5, all participants responded to the standard AV case. This study tested for a transfer effect. In the transfer-VOI condition, participants completed the VOI hospital case and a hypothetical version

of the VOI charity case (respectively from Studies 2 and 3) before responding to the standard AV case. We intentionally included two VOI cases for the VOI manipulation, before the standard AV case, to boost the possibility of transfer. In the simple control condition, participants responded only to the standard AV case. In the anchoring control condition, prior to the standard AV case, participants responded to the sculpture case from Study 4 and an additional case, the *speedboat* case, which also reliably elicits utilitarian judgments. Thus, as in Study 4, this control condition is intended to control for participants' making two affirmative utilitarian responses prior to responding to the standard AV case. We note that this control condition lacks the features introduced in Study 5 (which was run after Study 7). However, because our prediction concerning this control condition was not confirmed, the absence of these features does not affect our conclusions. In both the transfer-VOI condition and the anchoring control conditions, we counterbalanced the order of the two cases preceding the standard AV case.

References

1. J. Rawls, *A Theory of Justice.* (Harvard, 1971).
2. N. Frohlich, J. A. Oppenheimer, C. L. Eavey, Laboratory results on Rawls's distributive justice. *Brit. J. Poli. Sci.* **17**(1), 1-21 (1987).
3. T. Kameda, K. Inukai, S. Higuchi, A. Ogawa, H. Kim, T. Matsuda, M. Sakagami, Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decisions. *Proc. Natl. Acad. Sci. U.S.A.* **113**(42), 11817-11822 (2016).
4. J. C. Harsanyi, Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *J Poli Econ* **63**(4), 309-321 (1955).
5. J. C. Harsanyi, Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. *Amer. Poli. Sci. Rev.* **69**(2), 594-606 (1975).
6. J. J. Thomson, The trolley problem. *Yale Law J.* **94**,1395-1415 (1985).
7. J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, J. D. Cohen, An fMRI investigation of emotional engagement in moral judgment. *Sci.* **293**(5537), 2105-2108 (2001).
8. M. Koenigs, L. Young, R. Adolphs, D. Tranel, F. A. Cushman, M. Hauser, A. Damasio, Damage to the prefrontal cortex increases utilitarian moral judgements. *Nat.* **446**(7138), 908-911 (2007).
9. C. Hare, Should we wish well to all? *Phil. Rev.* **125**(4), 251-272 (2016).
10. J. M. Paxton, J. D. Greene, Moral reasoning: Hints and allegations. *Top. Cog. Sci.*, **2**(3), 511-527 (2010).
11. J. M. Paxton, L. Ungar, J. D. Greene, Reflection and reasoning in moral judgment. *Cog. Sci.* **36**(1), 163-177 (2012).

12. J. D. Greene, *Moral Tribes: Emotion, Reason, and the Gap between Us and Them.* (Penguin, 2013).

13. M. J. Crockett, Models of morality. *Tre. Cog. Sci.* **17**(8), 363-366 (2013).

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

15. I. Patil, M. M. Zucchelli, W. Kool, S. Campbell, F. Fornasier, M. Calò, G. Silani, M. Cikara, F. Cushman, Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures PsyArXiv doi:10.31234/osf.io/q86vx (22 December, 2018).

16. J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. (Vintage 2012).

17. J. F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles *Sci.* **352**(6293), 1573-1576 (2016).

18. D. Z. Morris, Mercedes-Benz's self-driving cars would choose passenger lives over bystanders. *Fortune* (15 October, 2016).

19. C. Robichaud, Liberty hospital simulation. Classroom exercise. (2015).

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

21. J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior (Commemorative Edition)*. (Princeton 2007).

22. A. D. Galinsky, G. B. Moskowitz, Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *J. Pers. Soc. Psy.* **78**(4), 708-724 (2000).

23. D. A. Pizarro, E. Uhlmann, P. Bloom, Causal deviance and the attribution of moral responsibility. *J. Exp. Soc. Psy*. **39**(6), 653–660 (2003).

24. S. Frederick, Cognitive reflection and decision making. *J. Econ. Pers.* **19**(4), 25-42 (2005).

25. J. M. Paxton, T. Bruni, J. D. Greene, Are 'counter-intuitive' deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Soc. Cog. Aff. Neur.* **9**(9), 1368-1371 (2013).

26. M. H. Bazerman, G. F. Loewenstein, S. B. White, Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Adm. Sci. Quar.* **37**(2), 220–240 (1992).

27. J. D. Greene, Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics* **124**(4), 695-726 (2014).

28. M. I. Norton, D. Ariely, Building a better America—One wealth quintile at a time. *Pers. Psy. Sci.* **6**(1), 9-12 (2011).

29. O. P. Hauser, M. Norton, (Mis) perceptions of inequality. *Cur. Opin. Psy.* **18**, 21-25. (2017).

30. P. Bloom, *Against Empathy: The Case for Rational Compassion*. (Ecco, 2016).

31. J. A. C. Everett, N. S. Faber, J. Savulescu, M. J. Crockett, The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *J. Exp. Soc. Psy.* **79**, 200-216 (2018).

32. M. J. Sandel, *Justice: What's the Right Thing To Do?*. (Macmillan, 2010).

33. F. H. Knight, *Risk, Uncertainty and Profit.* (Hart, Schaffner & Marx, 1921).

CHAPTER 3.

THIRD-PARTY JUDGMENTS OF VEIL-OF-IGNORANCE REASONING

Karen Huang

Abstract

Veil-of-ignorance reasoning may signal respect for persons and trustworthiness in a cooperation market. Across three experiments (two preregistered), I investigate how third-party observers perceive decision-makers based on their justifications of their judgments in sacrificial moral dilemmas. In Study 1, I find that people report more trust toward a decision-maker who expresses a VOI compared to utilitarian justification for the same utilitarian decision. Study 2 shows that people transfer more money in the trust game to the partner who expresses a VOI compared to utilitarian justification. Study 3 shows that in the context of a moral dilemma where people are generally split between the utilitarian and deontological responses, the use of a VOI justification primarily increases trust from deontologists. Taken together, these results show that when a decision-maker is dealing with a situation where the utilitarian response is unpopular, the decision-maker is better off defending the utilitarian response using VOI justification. Results across studies suggest that the primary mechanism driving the increase in observer trust is perceived warmth of the VOI decision-maker. Overall, VOI reasoning could serve as an intervention to maximize good consequences, while also signaling respect for the individuality of persons and thus promoting social cooperation.

Keywords: moral judgment; ethics; person perception; trust game

THIRD-PARTY JUDGMENTS OF VEIL-OF-IGNORANCE REASONING

Moral judgments are not made in social isolation. When decision-makers make moral judgments, those judgments also serve as social signals, whereby observers could infer personal characteristics and emotional responses of the decision-maker. In the realm of sacrificial moral dilemmas – that is, dilemmas involving the tradeoff between an action that favors the greater good, and competing moral concerns (e.g., the harm inflicted upon an individual) – people's responses often fall into two categories of moral judgments. A deontological judgment is characterized as rejecting an action that inflicts harm (e.g., sacrificing a person) in order to maximize overall welfare. Such a judgment aligns with deontological theories that notions of duties, rights, and obligations determine the rightness or wrongness of an action (e.g. Fried, 1978; Kant, 1797/2002; Rawls, 1971; Scanlon, 1998; W.D. Ross, 1930). On the other hand, a utilitarian judgment is characterized as endorsing an action that produces the best overall consequences, reflecting consequentialist moral theories positing that consequences (e.g., overall well-being or happiness) determine the rightness or wrongness of the moral decision (Bentham, 1789/1983; Mill, 1863). Decades of research in moral psychology have shown that automatic moral intuitions often align with deontological judgments, whereas utilitarian moral judgments are often driven by more deliberate processes (e.g., Greene, 2014; Haidt, 2001; Greene, 2007; Greene et al, 2008; Koenigs et al, 2007). To explain why ordinary moral intuitions more often align with deontology rather than utilitarianism, a partner choice account argues that deontologists are evolutionarily more favored in cooperative exchanges (Everett, Pizarro, & Crockett, 2016; Everett, Faber, Savulescu, & Crockett, 2018).

Moral judgments serve a fundamentally social role in cooperative exchanges. Partner choice accounts of the evolution of morality posit that people are more likely to choose as cooperation partners those who can be relied upon to act in a mutually beneficial way (Everett et al., 2016; Alexander, 1987; Baumard, André, & Sperber, 2013; Krebs, 2008; Noë & Hammerstein, 1994; Trivers, 1971). When decision-makers make moral judgments, third-party observers – that is, those who are neither the victim nor agent of the moral judgments, but uninvolved onlookers of the judgment – may make social inferences about the decision-maker. People may select more often as cooperation partners those who make certain types of moral judgments that signal a commitment to cooperation. Specifically, this partner choice account posits that deontologists are likely to be selected as social partners, hence why deontological judgments become prevalent as defaults (Everett et al., 2016).

What are the negative reputational consequences of making utilitarian judgments? Research shows that third-party observers perceive utilitarian decision-makers as less trustworthy and warm, cooperate with them less in economic games, and select them less frequently as social partners, compared to deontological decision-makers (e.g., Bostyn & Roets, 2017; Sacco, Brown, Lustgraaf, & Hugenberg, 2017; Everett et al., 2016; Everett et al., 2018; Rom, Weiss, & Conway, 2017; Uhlmann, Zhu, & Tannenbaum, 2013). Third-party observers perceive utilitarians as suppressing emotional responses such as empathy in order to deliberately calculate the consequences of each action (Uhlmann, Zhu, & Tannenbaum, 2013). Furthermore, people who make utilitarian judgments in sacrificial dilemmas -- endorsing harming one person to maximize outcomes for many -- may also appear to lack an aversion to harming other people in general (Kahane, Everett, Earp, Farias, & Savulescu, 2015). As such, utilitarian judgments fail to signal a respect for individual persons (Everett et al., 2018).

Extant research shows that people are more likely to trust decision-makers who express deontological judgments, compared to utilitarian judgments, in economic games (Everett et al. 2016; 2018; Bostyn & Roets, 2017). Why do people trust deontologists more than utilitarians? One important mechanism is that expressing a deontological judgment, compared to a utilitarian judgment, may signal socially valued emotional responses (Everett et al., 2016). Specifically, these socially valued emotional responses indicate *respect for individual persons*. Consistent with this account, people who make deontological judgments also show an aversion to harming others (Bartels & Pizarro, 2011; Cushman, Gray, Gaffey, & Mendes, 2012). Furthermore, third-party observers rate deontological decision-makers as more empathic, compared to utilitarian decision-makers (Uhlmann, Zhu, & Tannenbaum, 2013). This is consistent with research showing that people perceive decision-makers making deontological judgments – those who reject causing harm for the greater good -- as warmer compared to decision-makers making utilitarian judgments (Rom, Weiss, & Conway, 2017). According to the partner choice account, deontological judgments signal being a cooperative social partner.

Although deontologists are preferred to utilitarians as social partners, and deontological intuitions predominate when people make moral decisions, sometimes utilitarian judgments are necessary for promoting positive social outcomes (Greene, 2014). Research in moral psychology shows that it is very difficult for people to engage in complex reasoning that would influence their moral judgments in the utilitarian direction (e.g., Paxton & Greene, 2010; Paxton, Ungar, & Greene, 2012; Haidt, 2012; Pizarro, Uhlmann, & Bloom, 2003; Frederick, 2005; Paxton, Bruni, & Greene, 2013). Recent research shows that having people engage in veil-of-ignorance reasoning influences people to favor the greater good (Huang, Greene, & Bazerman, 2019). Veil-of-ignorance reasoning involves considering the experience and well-being of each of the people

potentially affected by the decision. This reasoning reflects John Rawls' moral theory, which posits that a fair society is one decision-makers would choose without information about their own personal circumstances that could bias their decisions (Rawls, 1971).

Since veil-of-ignorance reasoning increases utilitarian judgments, veil-of-ignorance reasoning could be used to arrive at utilitarian decisions that would otherwise seem aversive, and could be used to justify utilitarian decisions. However, does veil-of-ignorance reasoning incur a negative social cost, just as utilitarian judgments incur a negative social cost? Or are veil-of-ignorance decision-makers seen more favorably than utilitarians? In the current research, I investigate the following question: Does a veil-of-ignorance justification for a utilitarian decision, compared to a utilitarian justification, increase trust from third-party observers?

To illustrate a veil-of-ignorance justification compared to a utilitarian justification, consider the following example. In the classic philosophical dilemma, the footbridge case (Thomson, 1985), a decision-maker must choose whether to push the person wearing a backpack off the footbridge and onto the tracks. If the decision-maker pushes, then the person wearing the backpack will die, but will stop the trolley so that the five on the tracks will be saved. If the decision-maker does not push, then the five on the tracks will die. A utilitarian justification for pushing could proceed as follows:

> "I believe that it is most important to think about the total costs and benefits associated with each option. If the decision-maker pushes the person onto the tracks, 1 person dies and 5 people live. If the decision-maker does not push the person onto the tracks, then 5 people die and 1 person lives. Pushing the person off the tracks produces the best overall balance of costs and benefits. I think that it is better to save many lives rather than just one."

This justification considers the outcomes of each potential action, and arrives at the action that maximizes overall well-being. By contrast, for the same utilitarian decision to push, a veil-of-ignorance justification could proceed as follows:

> "I believe it is most important to think about this decision from an impartial perspective. What would I want the decision-maker to do if I had an equal chance of being each of these six people affected by the decision? Imagine I had a 1 out of 6 chance of being the person wearing the backpack, and a 5 out of 6 chance of being one of the people on the tracks. If I didn't know who I was going to be, I would want the decision-maker to push the person wearing the backpack, since this would mean a greater chance of living. Imagining I could be in the shoes of each of these people equally, I would want the decision-maker to make the most impartial choice."

Considering the dilemma in this way, the most impartial choice would be to push, since it means a greater chance of living, not knowing who one was going to be. I hypothesize that third-party observers are more likely to trust decision-makers using the veil-of-ignorance justification, compared to the utilitarian justification, for the same utilitarian decision.

Why would veil-of-ignorance justifications increase trust, compared to utilitarian justifications? I posit that the same mechanism that drives the increase in trust for decision-makers making deontological compared to utilitarian judgments may also drive an increase in trust for decision-makers using veil-of-ignorance compared to utilitarian justifications. Given the partner choice account for why deontologists are evolutionarily favored in cooperative exchanges, what social signals would allow people to favor veil-of-ignorance justifications? It would be helpful to look at the ways in which veil-of-ignorance reasoning may be similar to deontological reasoning. Veil-of-ignorance reasoning signals a respect for persons – that is, considering the experience and well-being of each individual potentially affected by the decision.

Third-party observers may view utilitarian decision-makers as looking for trouble – that they are okay with doing harm that does not benefit the greater good, that they could endorse harm for ignoble reasons, or that they suppress their empathy for the identifiable victims (Kogut & Ritov, 2005; Small & Loewenstein, 2003; Uhlmann et al., 2013; Kahane et al., 2015). But consequentialism isn't necessarily incompatible with empathy or respect for individual persons. Indeed, veil-of-ignorance reasoning involves respecting each person equally, and by doing so,

one is compelled to make the choice that favors the most of these individuals. Furthermore, veil-of-ignorance reasoning involves active perspective-taking and empathy. Thus, veil-of-ignorance reasoning counteracts a potential signal that the decision-maker endorses harm frivolously, or suppresses empathy. Third-party observers may perceive decision-makers using veil-of-ignorance reasoning as more warm, empathic, better at perspective-taking and respecting individual persons, compared to decision-makers using utilitarian reasoning. By signaling these socially desirable personality characteristics and emotional responses, veil-of-ignorance justifications may increase observer trust.

Even though a decision-maker using a veil-of-ignorance justification may not be as likely to be chosen as a cooperation partner compared to a decision-maker using a deontological justification, the decision-maker using a veil-of-ignorance justification may be more likely to be chosen as a cooperation partner compared to the decision-maker using a utilitarian justification. Thus, veil-of-ignorance justifications, compared to utilitarian justifications, may confer an adaptive function in the selection of utilitarian decision-makers. Overall, veil-of-ignorance reasoning could maximize good consequences, while also signaling respect for the individuality of persons. Veil-of-ignorance justifications may serve the function of increasing social cooperation, as deontological judgments do, while promoting positive social outcomes, as utilitarian judgments do.

Overview of Studies

Across a series of experiments, I investigate the following question: How do third-party observers perceive veil-of-ignorance reasoning? I hypothesize that people are more likely to trust decision-makers who use veil-of-ignorance reasoning, compared to utilitarian reasoning. Study 1 provides preliminary evidence that participants are more likely to trust decision-makers who use

veil-of-ignorance reasoning, compared to utilitarian reasoning. Furthermore, Study 1 shows that participants view decision-makers who use veil-of-ignorance reasoning in a sacrificial moral dilemma as warmer and more competent compared to decision-makers who use utilitarian reasoning, and that these increased perceptions of warmth and competence explain the effect of veil-of-ignorance reasoning, compared to utilitarian reasoning, on increased trust. Study 2 shows that participants transfer more money in a trust game to the decision-maker who uses veil-of-ignorance reasoning, compared to utilitarian reasoning. Study 2 also tests several candidate mechanisms underlying this effect: increased perceptions of socially valuable responses (i.e., empathy, humanization, perspective-taking, warmth, competence). Furthermore, Study 2 addresses a confound, expression of emotional conflict, in the veil-of-ignorance manipulation in Study 1. Study 3 investigates generalizability of the findings from Study 2 by testing whether participants also transfer more in the trust game to a partner who shows veil-of-ignorance reasoning, compared to utilitarian reasoning, in a different moral dilemma – a bioethical dilemma regarding the distribution of resources. In addition, Study 3 tests the mechanisms of warmth and competence derived from Studies 1-2, and shows a boundary condition of the veil-of-ignorance-justification intervention on trust from observers.

Study 1

Study 1, as an exploratory study, investigates how third-party observers perceive decision-makers who use veil-of-ignorance reasoning (VOI), compared to utilitarian and deontological reasoning. Since prior research investigating perceptions of moral decision-making has found that people prefer deontological to utilitarian decision-makers (Everett et al., 2016; 2018), I investigate how VOI decision-makers compare to these two groups. Therefore, this study employs a between-subjects design with three conditions, where participants give their

impression of a partner who gives a deontological, utilitarian, or VOI justification to their judgment in a moral dilemma.

Method

Sample. I recruited 400 participants from Amazon's Mechanical Turk (MTurk) in exchange for $2.00 per participant. All participants were U.S. residents. All exclusion criteria were decided *a priori.* I excluded 5 participants with duplicate MTurk IDs, and 87 participants who did not pass at least one comprehension check or attention check. This left a sample of 313 participants (162 male, 148 female, M_{age} = 36.16, SD_{age} = 11.67) for analysis.

Procedure. Participants first responded to the classic footbridge dilemma (Thomson, 1985) as a measure of participant moral judgment (deontological or utilitarian) using both a choice measure and a Likert-scale measure. I used the footbridge dilemma because it has been reliably shown in prior research to elicit non-utilitarian responses (e.g., Greene et al., 2001).

Participants were then shown the decision and justification (deontological, utilitarian, or VOI) given by another person, named Sam, to the same footbridge case. In the deontological condition, participants were told that Sam said the decision-maker should not push, with the following justification:

> "I believe that it is most important to think about the fundamental rights of persons. I think that it is wrong to sacrifice one person to save the five people on the tracks. I think that killing is just wrong regardless of the consequences."

In the utilitarian condition, participants were told that Sam said that the decision-maker should push, with the following justification:

> "I believe that it is most important to think about the total costs and benefits associated with each option. If the decision-maker pushes the person onto the tracks, 1 person dies and 5 people live. If the decision-maker does not push the person onto the tracks, then 5 people die and 1 person lives. Pushing the person off the tracks produces the best overall balance of costs and benefits. I think that it is better to save many lives rather than just

one."

In the VOI condition, participants were told that Sam said that the decision-maker should push, with the following justification:

> "I believe it is most important to think about this decision from an impartial perspective. What would I want the decision-maker to do if I had an equal chance of being each of these six people affected by the decision? Imagine I had a 1 out of 6 chance of being the person wearing the backpack, and a 5 out of 6 chance of being one of the people on the tracks. If I didn't know who I was going to be, I would want the decision-maker to push the person wearing the backpack, since this would mean a greater chance of living. But when I think about pushing, it feels wrong. Nevertheless, if pushing is what I'd want from an impartial perspective, not knowing who I was going to be, then perhaps it really is the right thing to do, even if it feels wrong."

Participants then responded to a series of scale measures. Since this was an exploratory study, I included an array of measures used by prior research investigating perceptions of utilitarian decision-making (Everett et al., 2016; 2018), to see if VOI reasoning would mitigate these negative perceptions. The main dependent variable of interest was trust, measured on a 1-7 scale. In addition, I asked participants to what extent they thought the partner was moral, warm/cold, competent, capable, sociable, and to what extent they would prefer the partner as a friend, spouse, boss and leader, measured on 1-7 scales. (See Appendix A for all measures and Appendix B for study stimuli.)

Finally, participants completed a series of comprehension checks regarding the dilemma, Sam's decision, and Sam's procedure for making the decision. (See Appendix C for comprehension checks.)

Results

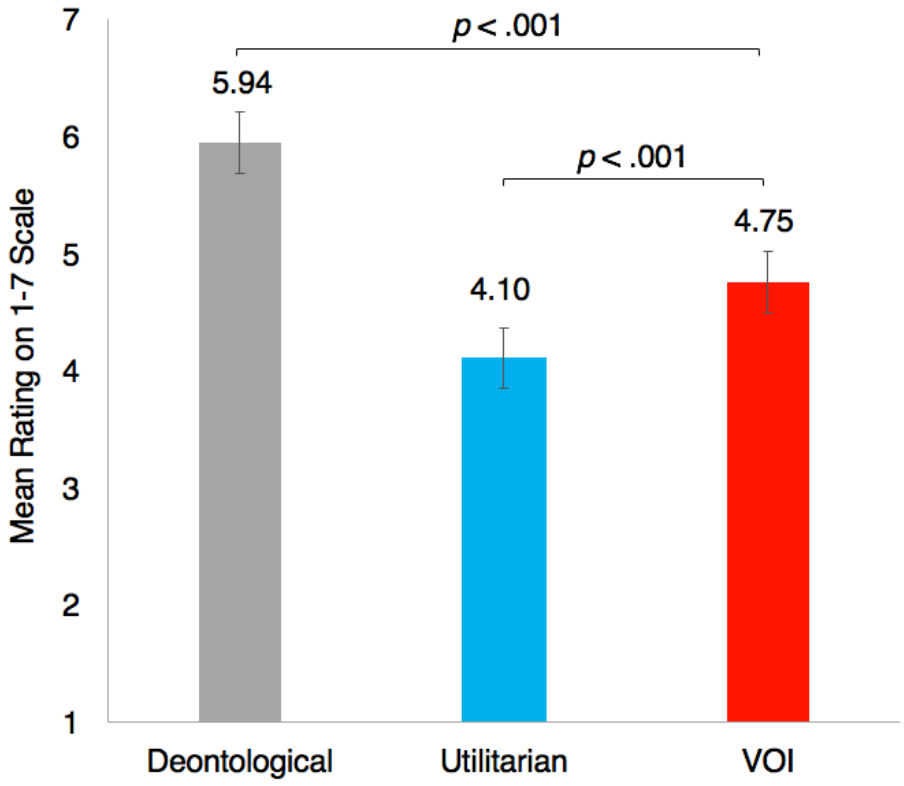
Participant Moral Judgment. In response to the footbridge case, 71.57% of participants said that it was not morally acceptable to push the person off the footbridge (deontological

judgment), and 28.43% of participants said it was morally acceptable (utilitarian judgment). Participants responded with a mean of 3.09 (*SD*=1.79) on the scale measure of moral judgment.

Trust. Participants reported higher trust toward the decision-maker using the VOI justification ($M = 4.75$, *95% CI:* [4.49, 5.01]), compared to the decision-maker using the utilitarian justification ($M = 4.10$, *95% CI:* [3.84, 4.36]), $\beta = .65$ (*95% CI:* [.29, 1.02]), $t(310) = 3.50$, $p < .001$) for the same utilitarian decision.

Participants reported lower trust toward the decision-maker using the VOI justification compared to the decision-maker using the deontological judgment ($M = 5.94$, *95% CI:* [5.68, 6.20]; $\beta = 1.19$, *95% CI:* [.82, 1.56], $t(310) = 6.39$, $p < .001$). (See Figure 2.1.)

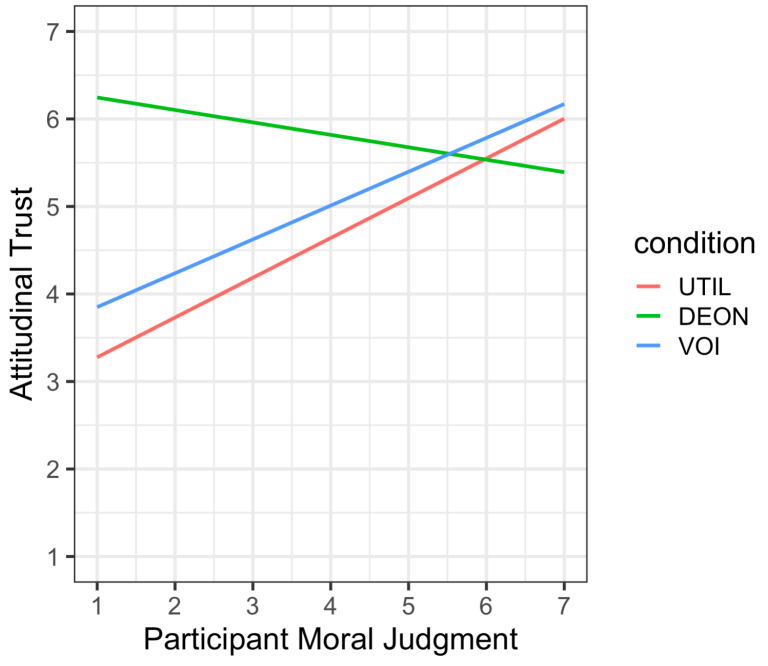
FIGURE 2.1. Results for attitudinal trust from Study 1.



Analyses employ linear regression. Bracketed values show 95% CI

Prior research on third-party perceptions of moral judgments has shown the importance of controlling for participants' own moral judgments (Everett et al., 2018). For completeness, I analyzed the effect of VOI vs. utilitarian justification on transfers in the trust game, controlling for participants' own moral judgments in the footbridge case, and controlling for the interaction between the partner's justification and participants' moral judgments. In this interaction model, there remained a marginally significant main effect of VOI justification compared to utilitarian justification on observer trust (β = .64, *95% CI:* [-.02, 1.30], $t(307) = 1.91$, $p = .058$). There was not a significant interaction effect of VOI vs. utilitarian response and participant moral judgment, meaning there was no detected difference in how deontologists and utilitarians[5] responded to the VOI vs. utilitarian justification (β = -.07, *95% CI:* [-.26, .12], $t(307) = -.71$, $p = .479$). (See Figure 2.2.)

FIGURE 2.2. Interaction model from Study 1.



[5] I use "deontologists" and "utilitarians" as shorthands to refer to participants who responded in the more deontological or utilitarian direction.

There was a significant main effect of deontological vs. utilitarian justification (β = 3.57, *95% CI:* [2.91, 4.22], $t(307)=10.79$, $p < .001$), aligning with prior research that deontologists are more trusted generally. There was also a significant interaction effect of deontological vs. utilitarian response and participant judgment, indicating that participants who responded more in the utilitarian direction reported less trust for the decision-maker who made the deontological judgment, compared to the utilitarian judgment (β = -.60, *95% CI:* [-.79, -.41], $t(307)= -6.19$, $p < .001$). Furthermore, the more participants responded in the utilitarian direction, the more trust they reported toward the decision-maker using the utilitarian justification (β = .45, *95% CI:* [.31, .59], $t(307)=6.41$, $p < .001$). These results align with a large body of evidence showing a similarity bias in people's perceptions of others (Lydon, Jamieson, & Zanna, 1988).

Warmth and Competence. An exploratory factor analysis of the single-item measures revealed two distinct factors of warmth and competence that explained 67% of the total variance. The warmth factor (*alpha* = .91) comprised of 5 items ("moral", "sociable", "warm", "friend", and "spouse"). The competence factor (*alpha* = .87) comprised of 4 items ("competent", "capable", "boss", and "leader").

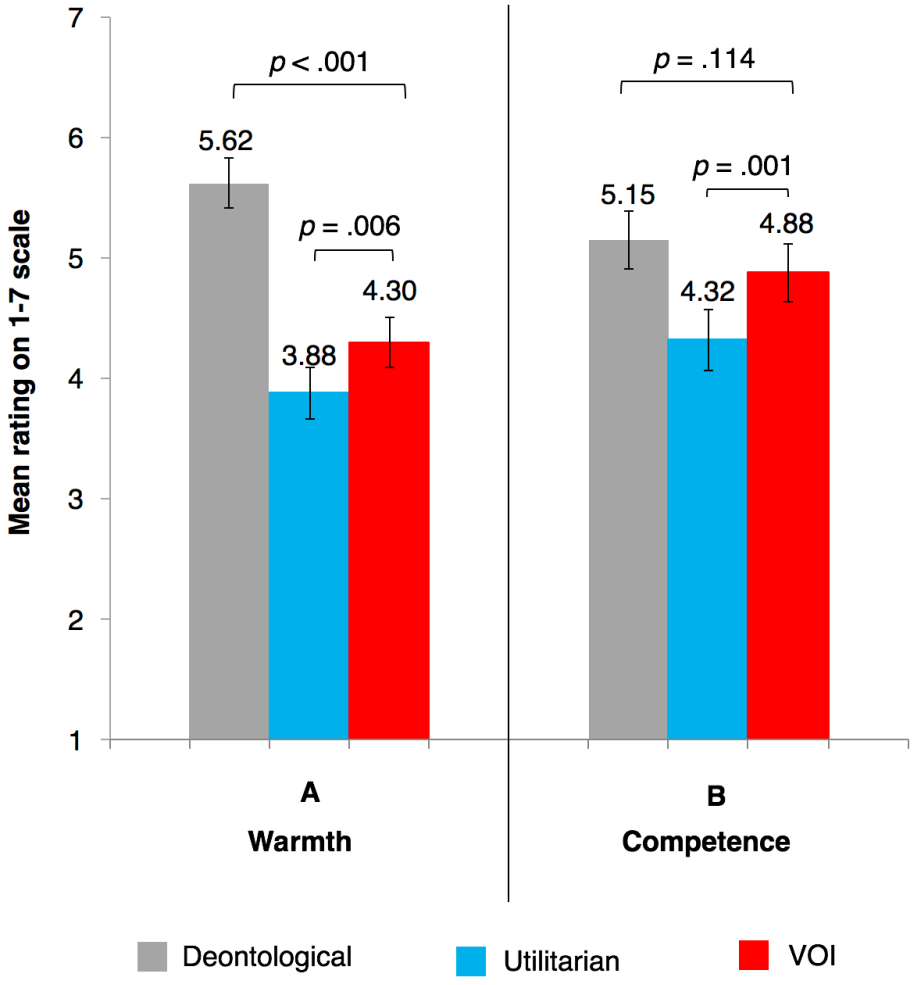
Warmth. Participants perceived the decision-maker using VOI reasoning to be warmer (M=4.30, *95% CI:* [4.09, 4.51]), compared to the decision-maker using utilitarian reasoning (M=3.88, *95% CI:* [3.67, 4.09]; β = .42 (*95% CI:* [.12, .72]), $t(310) = 2.80$, $p = .006$).

To test warmth as the mechanism underlying the effect of the VOI justification compared to utilitarian justification on trust, I conducted a mediation analysis (Preacher & Hayes, 2008; Preacher & Kelley, 2011). I estimated the causal pathway linking the partner's VOI vs. utilitarian justification with trust of the partner, as mediated by perceived warmth of the partner. I tested for and bootstrapped the indirect effect over 10,000 simulations and found an effect that was

significantly different from zero (indirect effect = .38, *95% CI:* [.09, .67]). This suggests that perceived warmth plays an important role in explaining why providing a VOI justification for a utilitarian decision increases trust from observers.

Participants perceived the decision-maker using VOI reasoning to be less warm (*M*=3.88, *95% CI:* [3.67, 4.09]) compared to the decision-maker using deontological reasoning (*M*=5.62, *95% CI:* [5.41, 5.83]; *ß* = 1.32 (*95% CI*: [1.02, 1.62]), *t*(310)=8.80, *p* < .001). (See Figure 2.3A.)

FIGURE 2.3. Results for warmth and competence factors from Study 1.



Analyses employ linear regression. Bracketed values show 95% CI.

I conducted a mediation analysis to test decreased perceived warmth as the mechanism underlying the effect of providing a VOI justification, compared to a deontological justification, on decreased trust. I tested for and bootstrapped the indirect effect over 10,000 simulations and found an effect that was significantly different from zero (indirect effect = -1.15, *95% CI:* [-1.47, -.87]). This suggests that a decrease in perceived warmth plays an important role in explaining why observers trust the VOI decision-maker less compared to the deontological decision-maker.

Competence. Participants perceived the decision-maker using VOI reasoning to be more competent (M=4.88, *95% CI:* [4.64, 5.12]), compared to the decision-maker using utilitarian reasoning (M = 4.32, *95% CI:* [4.08, 4.57]; β = .56 (*95% CI:* [.22, .90]), $t(310)$ =3.22, p=.001).

I fit a mediation model with trust as the dependent variable, VOI vs. deontological justification as the treatment variable, and perceived competence as the mediator variable. I tested for and bootstrapped the indirect effect over 10,000 simulations, and found an effect that was significantly different from zero (indirect effect = .46, *95% CI:* [.17, .75]). This suggests that perceived competence also plays an important role in explaining why providing a VOI justification for a utilitarian decision increases trust from observers.

There were no differences in perceived competence of the decision-maker using VOI reasoning compared to deontological reasoning (M = 5.15, *95% CI:* [4.91, 5.39]; β = -.27 (*95% CI:* [-.61, .07]), $t(310)$=-1.59, p = .114). (See Figure 3B.)

Multiple Mediation Analysis.

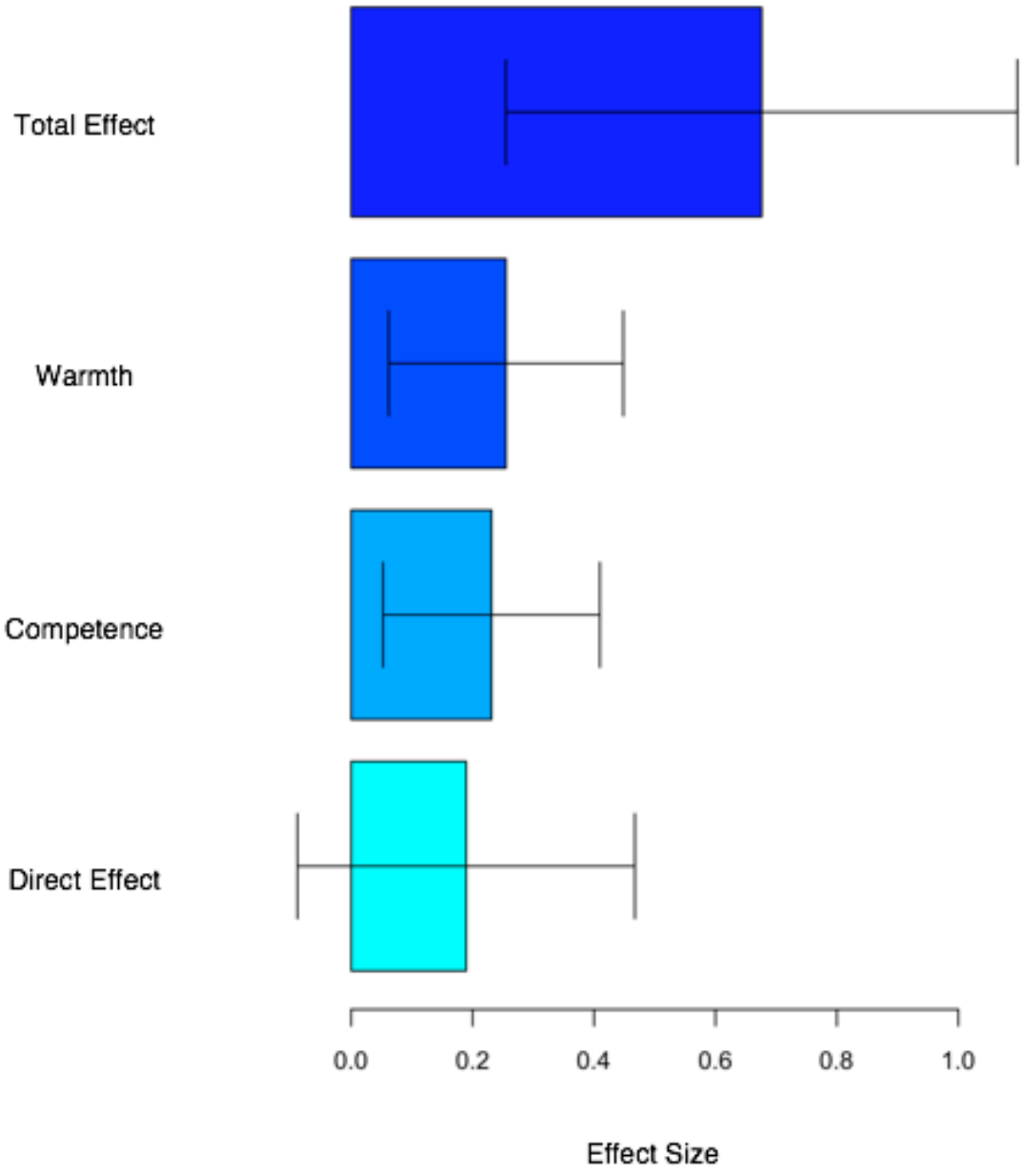
To estimate the individual mediation effects of warmth and competence on increased trust toward a partner who gives a VOI compared to utilitarian justification, I conducted a mediation analysis with multiple mediators using the "mma" R package (Yu & Li, 2017; Yu, Fan, & Wu, 2014). The mediation analysis decomposes the total effect into the direct effect from

condition on trust, and the indirect effects of condition on trust through multiple individual mediators (warmth and competence).

I used a generalized linear model to estimate the mediation effects – including the total effect, direct effect, and indirect effects -- and used the bootstrap method (500 times of bootstrap resampling) to measure the uncertainty in estimating the mediation effects (by calculating the estimated 95% confidence intervals).

The model estimated an indirect effect of warmth that was significantly different from zero (indirect effect = .28, *95% CI:* [.06, .45]) and an indirect effect of competence significantly different from zero (indirect effect = .23, *95% CI:* [.05, .41]). The model estimated a non-significant direct effect of condition (direct effect = .19, *95% CI:* [-.09, .47]), suggesting that warmth and competence fully mediate the effect of VOI vs. utilitarian justification on observer trust. The estimated total effect of all predictors (experimental condition and mediators) was .68, *95% CI:* [.26, 1.10]. (See Figure 2.4.)

FIGURE 2.4. Estimated mediation effects on attitudinal trust from Study 1.



Plots show estimated effect sizes (distance from 0), and bracketed values show 95% CI of the effect sizes. Mediation analysis with multiple mediators, using generalized linear model.

Discussion

These results provide preliminary evidence that people are more likely to trust a decision-maker using VOI reasoning, compared to utilitarian reasoning, for the same utilitarian decision. The present results also suggest that VOI reasoning could increase perceptions of both warmth and competence, compared to utilitarian reasoning, and that these increased perceptions explain the effect of VOI reasoning, compared to utilitarian reasoning, on increased observer trust.

In addition, the results show that people report decreased trust toward a decision-maker using VOI reasoning, compared to deontological reasoning, because they view the decision-maker using VOI reasoning as less warm. Therefore, results from Study 1 suggest that the mechanism of perceived warmth driving increases in trust in VOI compared to utilitarian decision-makers, also decreases trust in VOI compared to deontological decision-makers. The same mechanism seems to operate in how people view VOI decision-makers in comparison to utilitarian and deontological decision-makers.

However, the VOI manipulation in this study is confounded with the expression of emotional conflict. Along with VOI reasoning, the VOI manipulation included the following expression: "But when I think about pushing, it feels wrong. Nevertheless, if pushing is what I'd want from an impartial perspective, not knowing who I was going to be, then perhaps it really is the right thing to do, even if it feels wrong." Although the psychological mechanism regarding VOI reasoning may involve this process (Huang, Greene & Bazerman, 2019), the philosophical justification of VOI reasoning does not necessitate such a conflict expression. That is, such an expression is not unique to VOI reasoning and could be used alongside any moral reasoning. Thus, Study 2 aims to address this confound by employing an un-confounded manipulation of VOI reasoning.

Study 2

A primary motivation of Study 2 (preregistered on AsPredicted, #36296) is to investigate third-party judgments of VOI justification using a behavioral measure of trust. Since prior research has consistently shown that attitudinal and behavioral measures of trust regarding moral decision-makers may not always align (e.g., Everett et al. 2016; 2018), in Study 2 I use a behavioral measure of trust – transfers in the trust game – as a more reliable measure of trust with real consequences. I hypothesized that people would transfer more money in the trust game to decision-makers giving a VOI justification, compared to those giving a utilitarian justification.

Another motivation of Study 2 is to further investigate the mechanism by including additional measures of socially valuable responses. Since the VOI justification expresses the consideration of the mental state of each individual affected by the decision, additional socially valued emotional responses such as perceived empathy, perceived perspective-taking, and perceived humanization may drive the effect of VOI justification on trust.

Since veil-of-ignorance reasoning involves considering the experience of each individual affected by the decision, observers may perceive that the veil-of-ignorance decision-maker has greater empathy compared to the utilitarian decision-maker. Furthermore, since veil-of-ignorance reasoning involves imagining the experiences and outcomes for each individual affected by the decision, observers may perceive that the veil-of-ignorance decision-maker cares more about each person's harm. In addition, veil-of-ignorance reasoning may signal to observers that the decision-maker humanizes each person affected by the situation, treating them as individuals with experiences, rather than as numbers or inanimate objects (Haslam & Loughnan, 2014; Gray, Young & Waytz, 2012). A person using veil-of-ignorance reasoning may mostly attend to mental states, whereas a person using utilitarian reasoning may mostly attend to outcomes. Taken

together, veil-of-ignorance reasoning, through the expression of socially valued emotional responses, may increase trust from third-party observers.

Method

Sample. The sample size was determined *a priori* by a power analysis for a linear regression (ANOVA with 3 groups) capable of detecting a small effect size $f = .15$ at power = .80. According to the power analysis, the targeted final sample size was 432. Taking into account an exclusion rate of 66% of the total recruitment sample size (determined by a pilot study[6]), the total recruitment sample size would be 1270 participants.

I recruited 1333 participants from Amazon's Mechanical Turk. All participants were paid $2.00 each with a $0.30 bonus. All participants were U.S. residents. All exclusion criteria were decided *a priori.* I excluded 23 participants with duplicate MTurk IDs and 367 participants who did not pass at least one comprehension check. This left a final sample of 943 participants (402 male, 535 female, $M_{age} = 35.59$, $SD_{age} = 11.36$) for analysis.

Procedure. The design mirrored the design of Study 1, with several notable additions. First, the VOI manipulation in Study 2 addressed the confound in the VOI manipulation in Study 1 by removing the statements expressing emotional conflict. Thus, this VOI manipulation was a more accurate and un-confounded operationalization of VOI reasoning. (See Appendix B for study stimuli.)

Furthermore, Study 2 measured behavioral trust using the trust game (adapted from Everett et al., 2016; 2018). All participants first reported their judgment in the footbridge case, as

[6] In this pilot study, I found that participants who responded 3-5 on the scale measure for moral judgment in the footbridge dilemma were more likely to trust the decision-maker using VOI rather than utilitarian reasoning. Therefore, I had subsetted the data to include only these participants. In Study 2, I report the results for the full sample, since the results replicate with both this subsetted group and the full sample.

in Study 1. Then, they were introduced to the trust game and given information about a "partner" who made a decision in the footbridge case. Participants received information about the partner's VOI, deontological, or utilitarian justification, respectively. This study involved deception, because this partner was actually hypothetical. Participants then played the trust game with the partner, where they had a 30-cent bonus and could allocate 0-30 cents to the partner. This was the primary dependent measure. Participants also reported their prediction of how much money their partner would return to them.

As secondary measures of trust, I also included attitudinal measures of benevolence-based and competence-based trust (adapted from Brooks, Dai & Schweitzer, 2014; Twyman, Harvey, & Harries, 2008; Levin & Cross, 2004). The attitudinal trust measure from Study 1 consisted only of a single item, and so the attitudinal measures from Study 2 aimed to be more reliable and to capture two distinct constructs of trust.

I measured warmth and competence using the traits established by Fiske, Cuddy and Glick (2002). In addition to warmth and competence, I measured several other candidate mechanisms capturing socially valuable responses. The perceived perspective-taking scale was adapted from the Interpersonal Reactivity Index (Davis, 1980). The perceived empathy measure was adapted from prior work investigating perceived empathy in third-party judgments of moral decision-making (Uhlmann et al., 2013), prior work on the role of harm aversion in moral judgments (Rom et al., 2017; Cushman et al., 2012; Gray et al., 2012), and the "empathic concern" subscale of the Interpersonal Reactivity Index (Davis, 1980), as used by prior research in investigating moral judgment (Crockett et al., 2010; Kahane et al., 2015). Finally, the perceived humanization measure was adapted from measures of dehumanization – objectification of people as inert or instrumental, lacking individuality and likened to inanimate objects (Haslam

& Loughnan, 2014), and based on the impression formation process of understanding others as individuals (Neuberg & Fiske, 1987). The attitudinal measures of trust and the person perception measures were all presented in randomized order. (See Appendix A for all measures.)

All participants completed comprehension checks about the dilemma, the partner's decision, and the partner's justification. At the end of the study, participants completed a further comprehension check related to the partner's justification. Thus, this study employed more stringent comprehension checks than in Study 1. (See Appendix C.)

Analysis Plan. I preregistered to analyze the results of the trust game using a linear regression predicting the outcome variable (amount of money transferred in the trust game) with a dummy predictor variable indicating the condition (deontological, utilitarian, or VOI justification). This was the primary analysis.

I also preregistered to analyze the dependent variables of perceived empathy, perceived humanization, perceived perspective-taking, and perceived warmth and competence, in order to investigate possible multiple mechanisms for the effect of condition on transfers in the trust game. All of these analyses were conducted with linear regressions with condition as the predictor variable.

Across all dependent measures, the primary contrast of interest was between the VOI condition and the utilitarian condition. I tested differences between the VOI condition and the deontological condition as secondary analyses.

Results

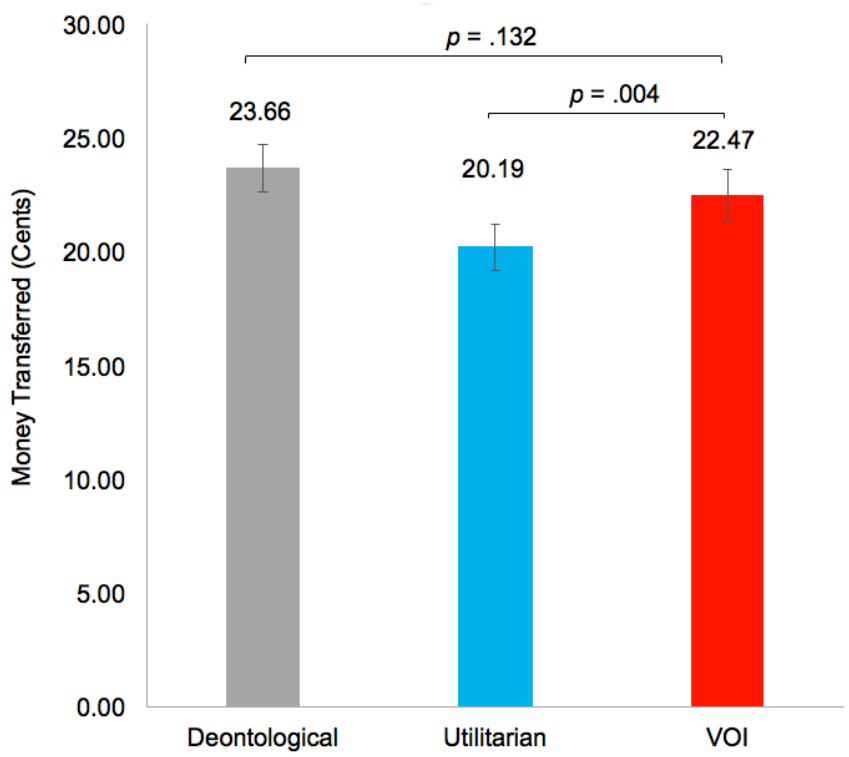
Participant Moral Judgment. In response to the footbridge case, 73.70% of participants said that it was not morally acceptable to push the person off the footbridge (deontological judgment), and 26.30% of participants said it was morally acceptable to push (utilitarian

judgment). Participants responded with a mean of 2.91 (*SD* = 1.68) on the scale measure of moral judgment.

Trust Game.

Money Transferred. As predicted, participants transferred more money to the partner who gave the VOI justification (*M*=22.47, *95% CI:* [21.32, 23.63]), compared to the partner who gave the utilitarian justification (*M*=20.19, *95% CI:* [19.16, 21.22]; *ß*=2.29 (*95% CI:* [.74, 3.83], *t*(940)=2.90, *p*=.004). There were no differences in money transferred between the VOI condition and the deontological condition (*M*=23.66, *95% CI:* [22.63, 24.70], *ß*= -1.19 (*95% CI:* -2.74, .36), *t*(940)=-1.51, *p*=.132). (See Figure 2.5.)

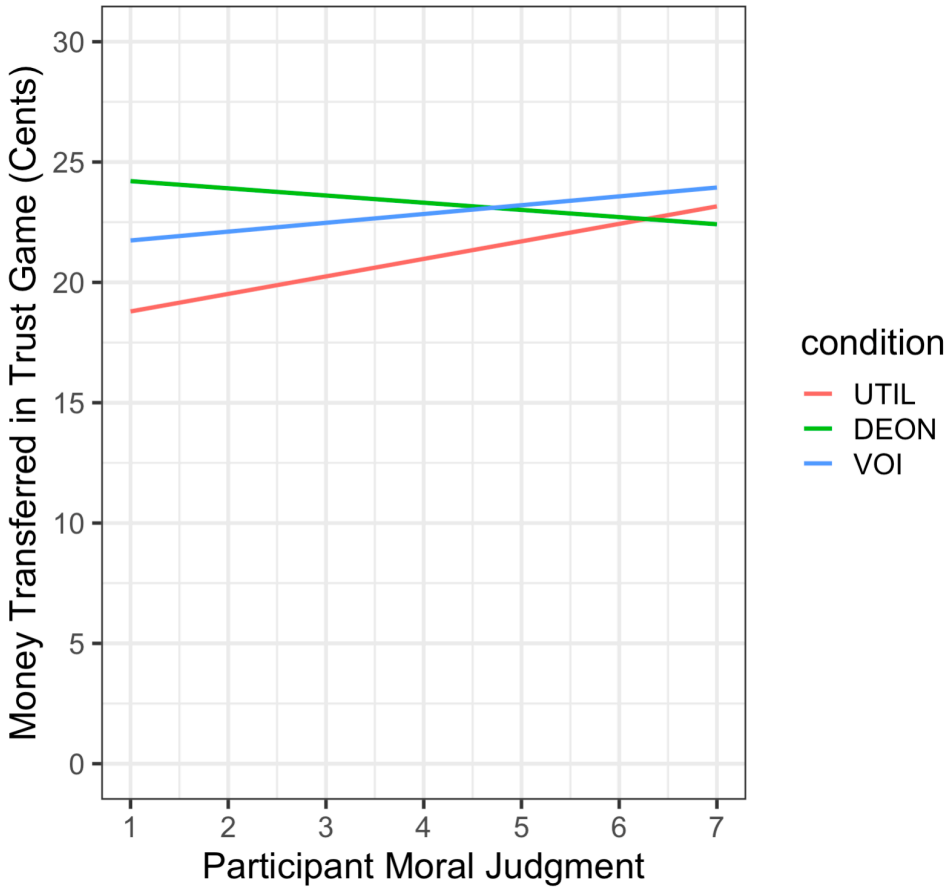
FIGURE 2.5. Results for trust game transfers from Study 2.



Analyses employ linear regression. Bracketed values show 95% CI.

For completeness, I also ran a linear model predicting how condition influences transfers in the trust game, controlling for the interaction between condition and participants' own moral judgments in the footbridge case. There remained a significant main effect of VOI justification, compared to utilitarian justification, in transfers in the trust game (β= 3.31, *95% CI:* [.19, 6.43], $t(937) = 2.08$, $p = .038$). There were no detected differences in transfers in the trust game to the partner who gave the VOI compared to utilitarian justification, based on participants' own moral judgments (β= -.36, *95% CI:* [-1.28, .55], $t(937) = -.77$, $p = .439$). That is, the effect of the VOI compared to utilitarian justification on trust did not depend on whether participants responded more in the deontological or utilitarian direction. (See Figure 2.6.)

FIGURE 2.6. Interaction model from Study 2.

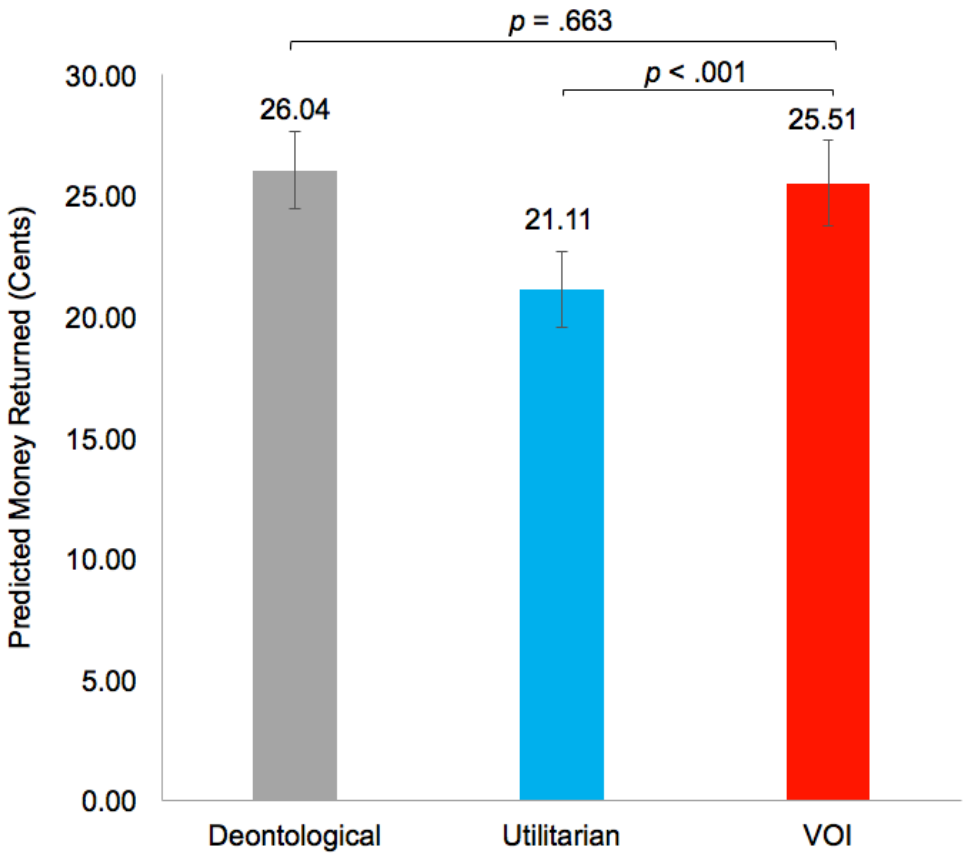


There was a significant main effect of deontological vs. utilitarian judgment (β=6.44, *95% CI:* [3.56, 9.33], $t(937)=4.38$, $p < .001$), aligning with prior research that deontologists are preferred generally. Furthermore, there was a significant interaction effect between deontological vs. utilitarian justification and participant judgment, such as participants responding more in the utilitarian direction transferred less money to the deontological decision-maker compared to utilitarian (β=-1.03, *95% CI:* [-1.90, -.16], $t(937)=-2.32$, $p = .021$). In addition, there was a significant main effect of participant judgment in the utilitarian direction on transfers to the utilitarian decision-maker (β=.73, *95% CI:* [.12, 1.33], $t(937)=2.35$, $p = .019$).

Predicted Return. As hypothesized, participants predicted to receive more money back in the trust game from the partner who gave the VOI justification ($M = 25.51$, *95% CI:* [23.74, 27.28]), compared to the partner who gave the utilitarian justification ($M = 21.11$, *95% CI:* [19.53, 22.68]; $\beta = 4.40$ (*95% CI:* [2.03, 6.77], $t(940) = 3.65$, $p <.001$).

There were no differences in predicted money received between the VOI condition and the deontological condition ($M = 26.04$, *95% CI:* [24.46, 27.62]; $\beta = -.53$ (*95% CI:* [-2.90, 1.85], $t(940) = -.44$, $p = .663$. (See Figure 2.7.)

FIGURE 2.7. Results for predicted trust game returns from Study 2.



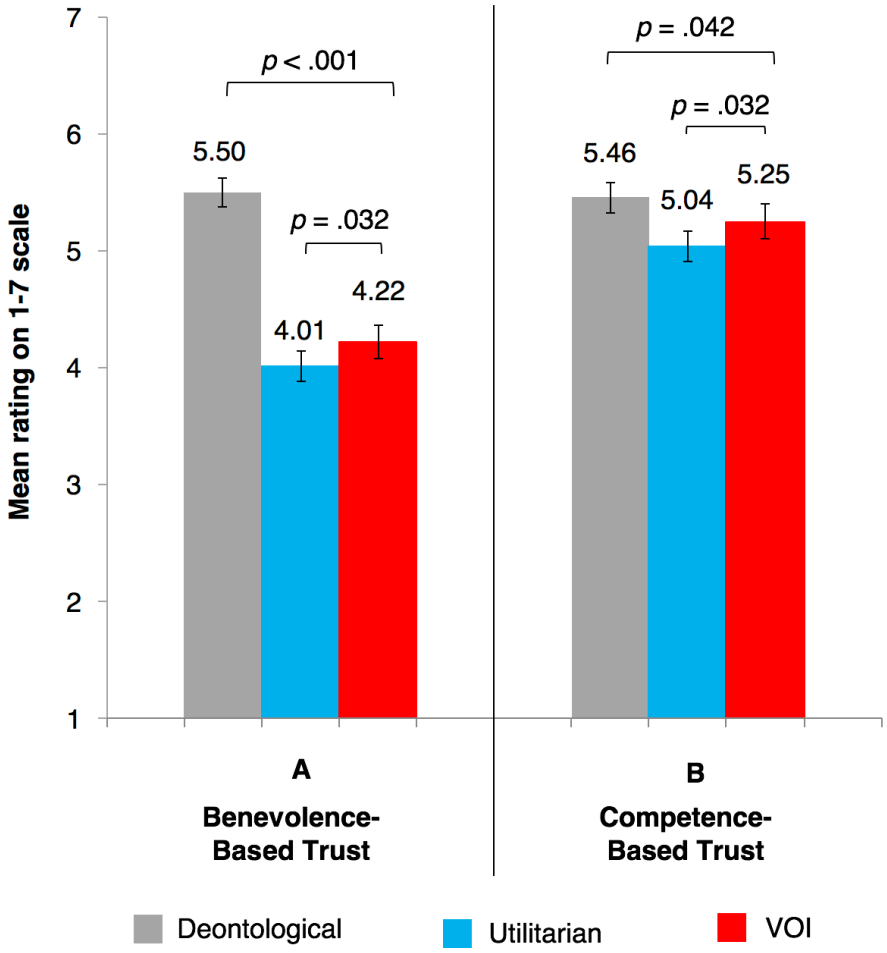
Analyses employ linear regression. Bracketed values show 95% CI.

Attitudinal Measures of Trust.

Benevolence-Based Trust. Participants reported higher benevolence-based trust toward the partner who gave the VOI justification ($M = 4.22$, *95% CI:* [4.08, 4.37]), compared to the partner who gave the utilitarian justification ($M = 4.01$, *95% CI:* [3.89, 4.14]; $\beta = .21$ (95% CI: [.02, .40]), $t(940) = 2.14$, $p = .032$).

Participants reported lower benevolence-based trust toward the partner who gave the VOI justification, compared to the partner who gave the deontological justification ($M = 5.50$, *95% CI:* [5.37, 5.63]; $\beta = 1.28$ (*95% CI:* [1.08, 1.47]), $t(940) = -12.96$, $p < .001$). (See Figure 2.8A.)

FIGURE 2.8. Results for benevolence-based and competence-based trust (attitudinal measures) from Study 2.



Analyses employ linear regression. Bracketed values show 95% CI.

Competence-Based Trust. Participants reported higher competence-based trust toward the partner who gave the VOI justification (M = 5.25, *95% CI:* [5.10, 5.40]), compared to the partner who gave the utilitarian justification (M = 5.04, *95% CI:* [4.91, 5.17]; β = .22 (*95% CI:* [.02, .41]), $t(940) = 2.15$, $p = .032$).

Participants reported lower competence-based trust toward the partner who gave the VOI justification, compared to the partner who gave the deontological justification (M = 5.46, *95% CI:* [5.33, 5.59]; β = .20 (*95% CI:* [.01, .40]), $t(940) = -2.04$, $p = .042$). (See Figure 2.8B.)

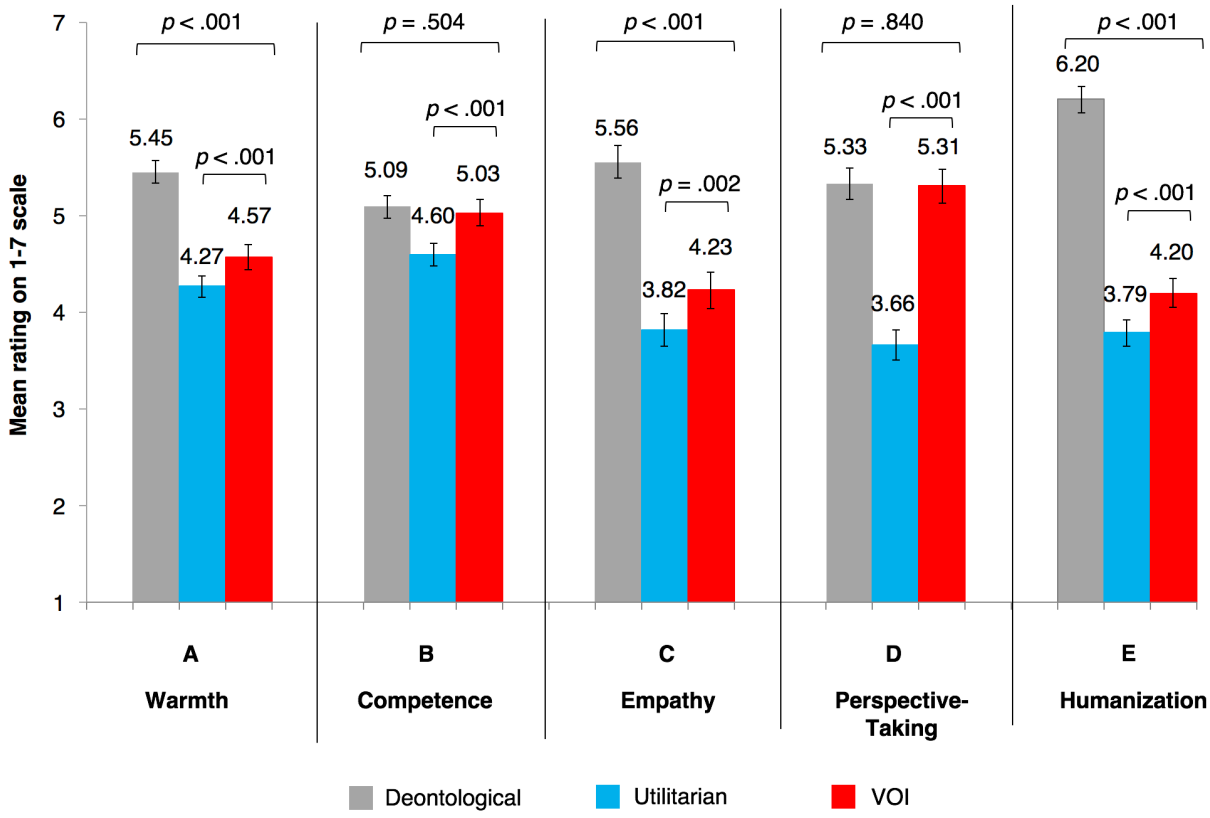
Person Perception Measures.

Warmth. Participants perceived the partner who gave the VOI justification as warmer (M = 4.57, *95% CI:* [4.44, 4.70]), compared to the partner who gave the utilitarian justification (M = 4.27, *95% CI:* [4.15, 4.38]; β = .30 (*95% CI:* [.13, .47]), $t(940) = 3.42$, $p < .001$).

To test warmth as the mechanism of the VOI justification, compared to utilitarian justification, on transfers in the trust game, I conducted a mediation analysis. I tested for and bootstrapped the indirect effect over 10,000 simulations and found an indirect effect that was significantly different from zero (indirect effect = .80, *95% CI:* [.30, 1.40]). This replicates the finding from Study 1 that perceived warmth plays an important role in explaining why providing a VOI justification for a utilitarian decision increases trust from observers.

Participants perceived the partner who gave the VOI justification as less warm compared to the partner who gave the deontological justification (M = 5.45, *95% CI:* [5.34, 5.57]; β = .88 (*95% CI:* [.71, 1.05]), $t(940) = -9.97$, $p < .001$). (See Figure 2.9A.)

FIGURE 2.9. Results for person perception attitudinal measures from Study 2.



Analyses employ linear regression. Bracketed values show 95% CI.

Competence. Participants perceived the partner who gave the VOI justification as more competent (M = 5.03, *95% CI:* [4.89, 5.17]), compared to the partner who gave the utilitarian justification (M = 4.60, *95% CI:* [4.48, 4.72]; β = .43 (*95% CI:* [.25, .61]), $t(940) = 4.67$, $p <$ .001).

Fitting a mediation model with perceived competence as the mediator variable, I tested for and bootstrapped the indirect effect over 10,000 simulations, and found an indirect effect that was significantly different from zero (indirect effect = 1.02, *95% CI:* [.55, 1.64]). This result provides further evidence that perceived competence also plays an important role in explaining

why providing a VOI compared to utilitarian justification, for the same utilitarian decision, increases trust from observers.

There were no differences in perceived competence between the partner who gave the VOI justification and the partner who gave the deontological justification (M = 5.09, *95% CI:* [4.97, 5.21]; β = -.06 (*95% CI:* [-.24, .12]), $t(940) = -.67$, $p = .504$). (See Figure 2.9B.)

Perceived Empathy. Participants perceived the partner who gave the VOI justification (M = 4.23, *95% CI:* [4.04, 4.42] to be more empathic, compared to the partner who gave the utilitarian justification (M = 3.82, *95% CI:* [3.65, 3.99]; β = .41 (*95% CI:* [.16, .67]), $t(940) = 3.16$, $p = .002$).

To test perceived empathy as a mechanism of the VOI justification on transfers in the trust game, I tested a mediation model. Across 10,000 simulations, this procedure estimated an indirect effect that was significantly different from zero (indirect effect = .41, *95% CI:* [.13, .86]). This suggests that perceived empathy plays a role in explaining the effect of VOI justification on increased trust from observers.

Participants perceived the partner who gave the VOI justification to be less empathic compared to the partner who gave the deontological justification (M = 5.56, *95% CI:* [5.38, 5.73]; β = 1.32 (*95% CI:* [1.07, 1.58]), $t(940) = -10.17$, $p < .001$). (See Figure 2.9C.)

Perceived Perspective-Taking. Participants perceived higher perspective-taking in the partner who gave the VOI justification (M = 5.31, *95% CI:* [5.13, 5.49]), compared to the partner who gave the utilitarian justification (M = 3.66, *95% CI:* [3.50, 3.82]; β = 1.65 (*95% CI:* [1.41, 1.89], $t(940) = 13.49$, $p < .001$).

I ran a mediation model testing perceived perspective-taking as a further mechanism explaining the effect of VOI justification on increased trust from observers. Across 10,000

simulations, this procedure estimated an indirect effect that was significantly different from zero (indirect effect = 1.61, *95% CI:* [.74, 2.58]). This suggests that perceived perspective-taking plays a role in explaining the effect of VOI justification on increased trust from observers.

There were no differences in perceived perspective-taking between the VOI condition and the deontological condition (M = 5.33, *95% CI:* [5.17, 5.49]; β = -.02 (*95% CI:* [-.26, .22]), $t(940) = -.20$, $p = .840$). (See Figure 2.9D.)

Perceived Humanization. Participants perceived the partner who gave the VOI justification to be more humanizing (M = 4.20, *95% CI:* [4.05, 4.36]), compared to the partner who gave the utilitarian justification (M = 3.79, *95% CI:* [3.66, 3.93]; β = .41 (*95% CI:* [.21, .62], $t(940) = 3.94$, $p < .001$).

I ran a mediation model testing perceived humanization as a mechanism explaining the effect of VOI compared to utilitarian justification on increased trust from observers. Across 10,000 simulations, this procedure estimated an indirect effect that was significantly different from zero (indirect effect = .48, *95% CI:* [.19, .97]). This suggests that perceived humanization plays an additional role in explaining the effect of VOI compared to utilitarian justification on increased trust from observers.

Participants perceived the partner who gave the VOI justification as less humanizing compared to the partner who gave the deontological justification (M = 6.20, *95% CI:* [6.06, 6.34]; β = 2.00, *95% CI:* [1.79, 2.21], $t(940) = 19.02$, $p < .001$). (See Figure 2.9E.)

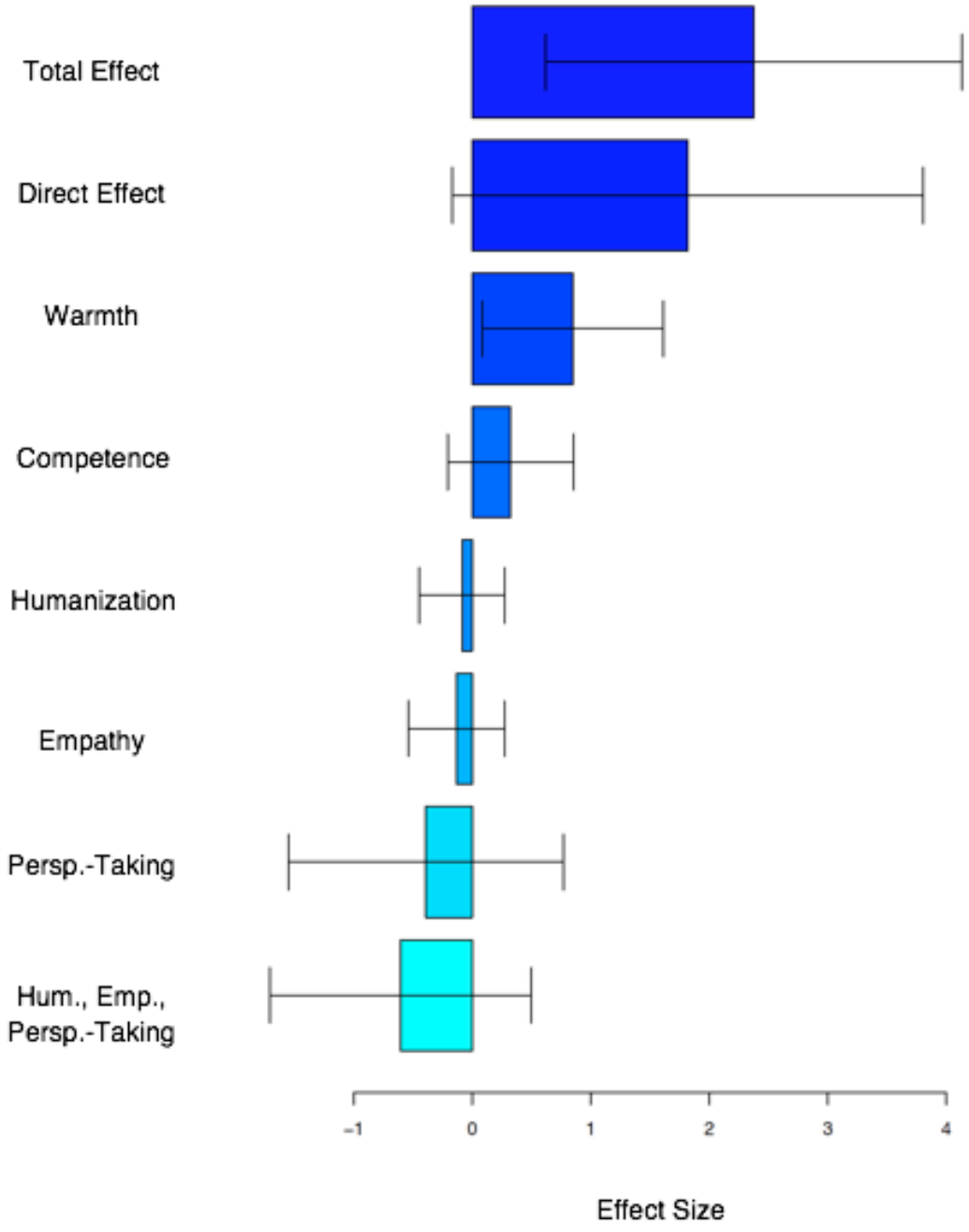
Multiple Mediation Analysis.

To differentiate the individual effects from multiple mediators, which measure different constructs, on increased trust toward a partner who gives a VOI compared to utilitarian justification, I conducted a mediation analysis using the "mma" R package (Yu & Li, 2017; Yu

et al., 2014). This mediation analysis parses the total effect into the direct effect of experimental condition (VOI compared to utilitarian justification) on trust, and the indirect effects through multiple individual mediators.

I first ran a mediation analysis with all of the mediator variables, and included the joint mediation effects of perceived empathy, perceived perspective-taking, and perceived humanization. I used a generalized linear model to estimate the mediation effects – including the total effect, direct effect, and indirect effects -- and used the bootstrap method to calculate the estimated 95% CI, with 500 iterations of bootstrap resampling. The model estimated an indirect effect of warmth significantly different from zero (indirect effect = .88, *95% CI:* [.13, 1.46]). All other indirect effects of mediators were non-significant. (See Figure 2.10.) The model estimated a non-significant direct effect of VOI vs. utilitarian justification (direct effect = 1.70, *95% CI:* [-.11, 3.46]), suggesting "full mediation", and a significant total effect of all the predictors (total effect = 2.32, *95% CI:* [.67, 3.93]). Calculating the proportion mediated, 37.93% of the increase in transfers in the trust game to the partner with the VOI vs. utilitarian justification can be explained by the effect of increased perceived warmth.

FIGURE 2.10. Estimated mediation effects on transfer results in trust game from Study 2.



Plots show estimated effect sizes (distance from 0), and bracketed values show 95% CI of the effect sizes. Mediation analysis with multiple mediators, using generalized linear model.

Discussion

Results from Studies 1-2 show that people are more likely to trust VOI decision-makers compared to utilitarian decision-makers. This result was replicated using a behavioral measure of trust in Study 2. Furthermore, the results from Study 2 suggest that this effect is explained primarily by warmth, out of the possible set of socially desirable emotional responses.

Furthermore, the results show that although people are more likely to trust the deontological decision-maker compared to the VOI decision-maker, and generally view the deontological decision-maker more positively compared to the VOI decision-maker, people are nevertheless more likely to trust the VOI decision-maker compared to the utilitarian decision-maker. Although providing a utilitarian justification for a utilitarian decision is perceived negatively, using a VOI justification for the same utilitarian decision boosts trust from observers.

Study 3

The primary motivation for Study 3 (preregistered on AsPredicted.org, #36594) is to investigate generalizability of the findings from Studies 1-2. In Studies 1-2, participants were presented with the moral judgments of a partner responding to the footbridge dilemma, which reliably elicits non-utilitarian responses. Study 3 tests whether participants exhibit the same trusting behavior toward a partner who gives a VOI justification in response to a different moral dilemma, one regarding the distribution of resources in the domain of bioethics. In this dilemma, an ethics committee must decide whether to take oxygen away from one patient, thus killing that patient, in order to provide for 9 life-saving surgeries (Robichaud, 2015). Prior research has found that 42.60% of participants make the utilitarian judgment in this case (Huang et al., 2019). The utilitarian moral judgment in the hospital case appears to be less controversial than the moral judgment in the footbridge case (i.e., where 71.57% of participants in Study 1 and 73.70% in

Study 2 give the utilitarian judgment to push). Thus, Study 3 investigates a possible boundary condition of the effect of VOI justification on trust game transfers. Does the VOI justification increase trust from observers when the moral dilemma at hand is less controversial?

Furthermore, rather than recruiting from Mechanical Turk, in Study 3 I recruit participants from a different online platform, Prolific, to investigate generalizability across samples. As in Study 2, the primary dependent variable in Study 3 is still trust game behavior. Based on the findings from Studies 1-2, the measures of mechanism in Study 3 only include warmth and competence.

Method

Sample. The sample size was determined *a priori* by a power analysis for a linear regression (ANOVA with 3 groups) capable of detecting an effect size $f = .148$ (determined by the effect size for increased transfers in the trust game found in Study 2) at power = .80. According to the power analysis, the targeted final sample size was 444. Taking into account an exclusion rate of 29% of the total recruitment sample size (determined by the exclusion rate found in Study 2), the total recruitment sample size would be 628 participants.

I recruited 631 participants from Prolific. Each participant was paid $2.00 with a $0.30 bonus. All participants were U.S. residents. All exclusion criteria were decided *a priori.* I excluded 8 participants with duplicate Prolific IDs, and 114 participants who did not pass at least one comprehension check. This left a final sample of 509 participants (235 male, 261 female, $M_{age} = 37.18$, $SD_{age} = 12.27$) for analysis.

Procedure. The design was identical to the design of Study 2, with the following differences. The study tested a moral dilemma different from those of Studies 1-2. Participants in Study 3 first gave their moral judgment to a bioethical dilemma, the Liberty Hospital case, and

then read justifications from their "partner" regarding their moral judgment to the same dilemma, before playing the trust game with the partner.

As in Study 2, the primary dependent measure in Study 3 was the amount of money transferred in the trust game to one's partner. A secondary dependent measure of trust was participants' predictions of how much money their partner would return to them. Because the results from Study 2 suggested that warmth was the primary causal mechanism underlying the effect of VOI vs. utilitarian justification on increased transfers in the trust game, Study 2 measures warmth as the mechanism. A secondary measure was competence, since the mediation analysis from Study 1 suggested competence as a mediator as well. (See Appendix A for all measures and Appendix B for study stimuli.) Finally, participants completed comprehension checks, as in Study 2. (See Appendix C.)

Analysis Plan.

I preregistered to analyze the trust game results using a linear regression predicting the amount of money transferred in the trust game with a dummy predictor variable indicating the condition (deontological, utilitarian, or VOI justification). I also planned to conduct mediation analyses both with warmth as the single mediator, and with warmth and competence as possible multiple mediators.

Furthermore, I preregistered that for exploratory purposes, I would also run linear regressions predicting the outcome variable of trust game transfer, with the predictor variables of condition and the participant's judgment in the hospital case, to see if participant's judgment has an interaction effect or main effect on trust game transfer.

Across all dependent measures, the primary contrast of interest was between the VOI condition and the utilitarian condition. I tested differences between the VOI condition and the deontological condition as secondary analyses.

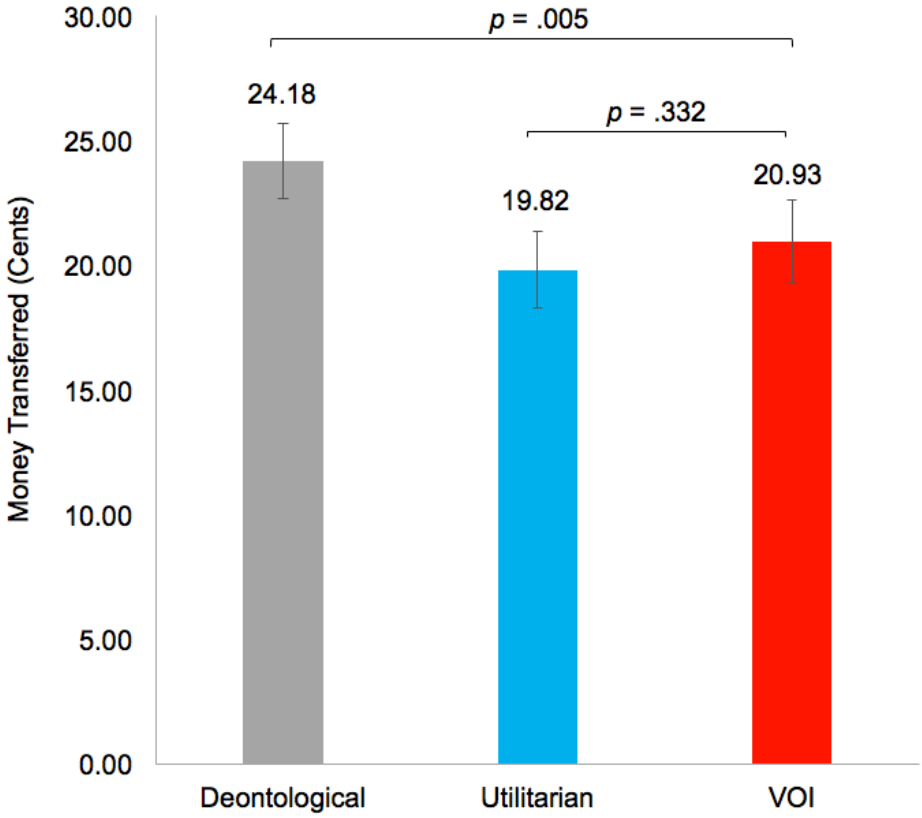
Results

Participant Moral Judgment. In response to the hospital case, 57.56% of participants said that it was not morally acceptable to take the patient off oxygen (deontological response), and 42.44% of participants said it was morally acceptable to take the patient of oxygen (utilitarian response). Participants responded with a mean of 3.60 on the scale measure (*SD* = 1.68). These results align with prior research showing the distribution of participant responses to this moral dilemma (Huang et al., 2019).

Trust Game.

Money Transferred. As in Study 2, I ran a linear regression predicting how condition influences transfers in the trust game. Contrary to my hypothesis, there were no differences in trust game transfers between the VOI-justification condition (M = 20.93, *95% CI:* [19.26, 22.60]) and the utilitarian-justification condition (M = 19.82, *95% CI:* [18.30, 21.34]; β = 1.12 (*95% CI:* [-1.14, 3.38]), $t(506) = .97$, $p = .332$). Participants transferred less in the trust game to the participant who gave the VOI justification, compared to the participant who gave the deontological justification (M = 24.18, *95% CI:* [22.66, 25.70]; β = -3.24 (*95% CI:* [-5.50, -.98]), $t(506) = -2.82$, $p = .005$). (See Figure 2.11.)

FIGURE 2.11. Results for trust game transfers from Study 3.



Analyses employ linear regression. Bracketed values show 95% CI.

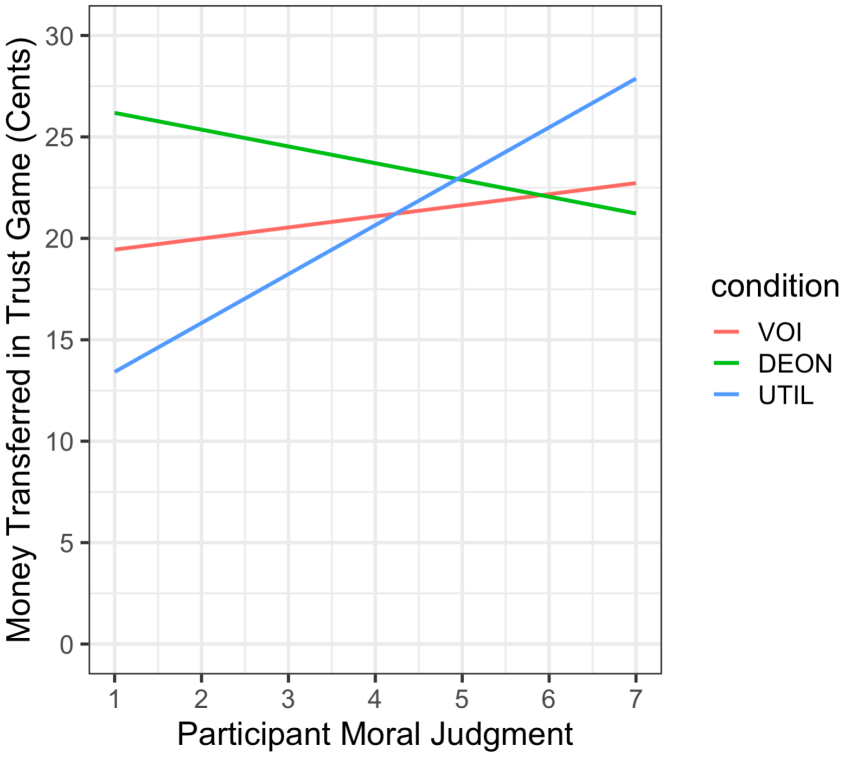
Since people are more likely to give utilitarian responses in the hospital case compared to the footbridge case, participants' own judgments may influence how the VOI justification, compared to utilitarian, influences their transfers in the trust game. I ran a linear model predicting how condition influences transfers in the trust game, controlling for the interaction between condition and participants' own moral judgments in the hospital case.

Overall, the results show that the relationship between transfers in the trust game and condition (decision justification) depends on the participants' own judgments in the hospital case. There is a main effect of VOI justification, compared to utilitarian, on trust game transfers,

controlling for the effect of participants' own moral judgment, and controlling for the effect of the interaction between participant judgment and partner justification. Participants transferred more money in the trust game to the partner who gave the VOI justification (β = 7.89, *95% CI:* [2.60, 13.18], $t(503) = 2.93$, $p = .004$), compared to the utilitarian justification.

Results show a significant interaction between the VOI vs. utilitarian justification and participant judgment. The effect of the VOI vs. utilitarian justification on trust game transfers differs depending on participant judgment. The more deontological the participant judgment, the more the VOI justification increases trust compared to the utilitarian justification (β = 1.86, *95% CI:* [-3.17, -.56], $t(503) = 2.81$, $p = .005$). This significant interaction term suggests that for deontologists, even though they prefer the deontological justification the most, the VOI justification comes in second, while the utilitarian justification comes in last. For utilitarians, on the other hand, the utilitarian justification is preferred over the VOI and deontological justifications. (See Figure 2.12.)

FIGURE 2.12. Interaction model from Study 3.



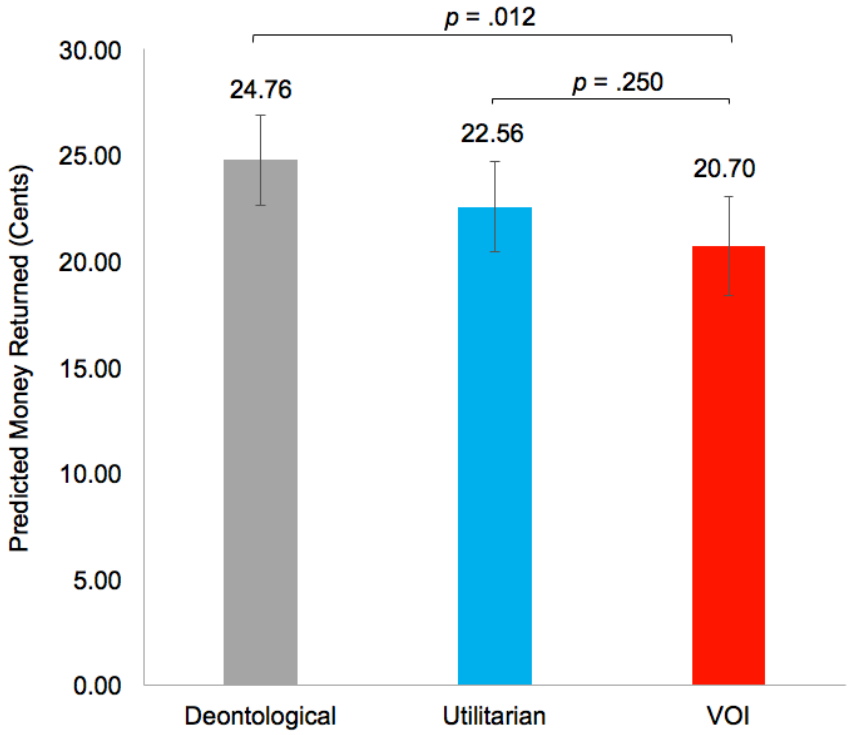
Transfers in trust game as a function of participant moral judgment and condition.

There was a significant main effect of deontological vs. utilitarian judgment on transfers in the trust game (β = 16.01, *95% CI:* [11.17, 20.85], $t(503)$=6.50, $p < .001$), aligning with prior research showing that deontologists are more trusted than utilitarians. Results also show a significant interaction term between the deontological vs. utilitarian justification and participant judgment. Results show that the more utilitarian the participants' own moral judgment, the less money they transferred to the partner who gave the deontological explanation compared to the utilitarian explanation (β = -3.24, *95% CI:* [-4.47, -2.00], $t(503)$ = -5.15, $p < .001$). Furthermore, participants who gave more utilitarian responses in the hospital case also transferred more money in the trust game to the partner who gave the utilitarian justification (β = 2.41, *95% CI:* [1.58, 3.25], $t(503)$ = 5.68, $p < .001$). These results align with a large body of evidence showing that

people are more likely to show a bias toward people with whom they agree (Lydon, Jamieson, & Zanna, 1988).

Predicted Return. Results on predicted returns in the trust game found no differences between the VOI-justification condition (*M* = 20.70, *95% CI:* [18.35, 23.04]) and the utilitarian-justification condition (*M* = 22.56, *95% CI:* [20.42, 24.69]); *ß* = -1.86 (*95% CI:* [-5.03, 1.31]), *t*(506)= -1.51, *p* = .250). Participants predicted to receive higher returns from the partner who gave the deontological justification (*M* = 24.76, *95% CI:* [22.62, 26.89]), compared to the VOI justification (*ß* = 4.06, *95% CI:* [.89, 7.23], *t*(506) = 2.51, *p* = .012). (See Figure 2.13.)

FIGURE 2.13. Results for predicted trust game returns from Study 3.

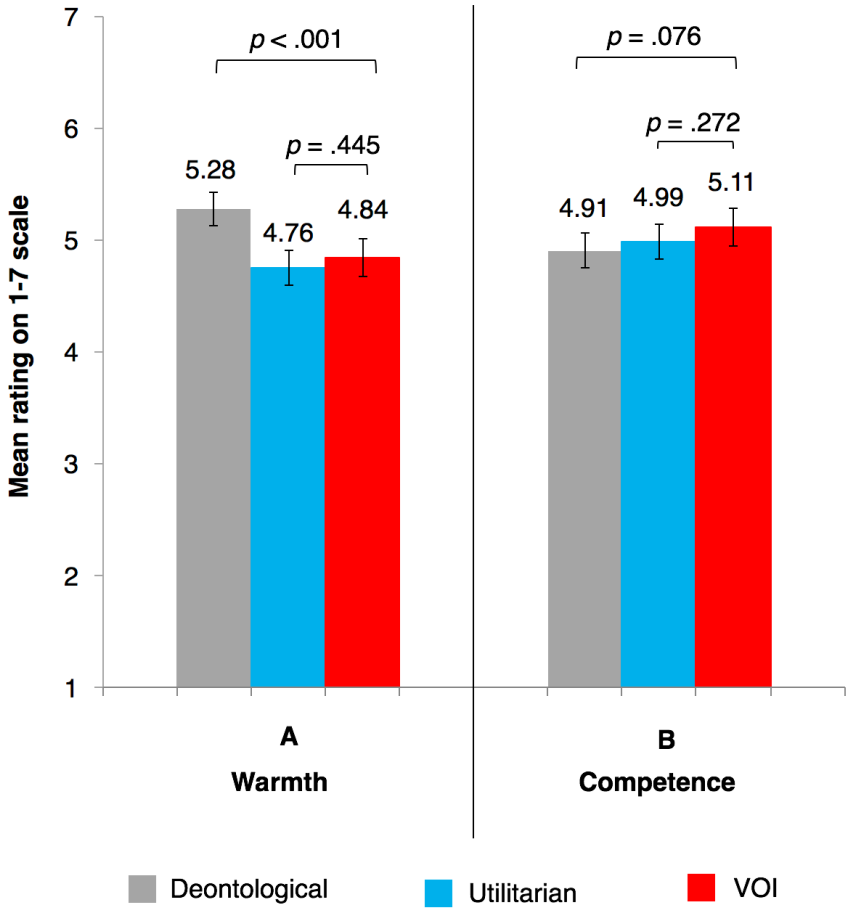


Analyses employ linear regression. Bracketed values show 95% CI.

Warmth and Competence.

Warmth. Contrary to my hypothesis, I did not find a difference in perceived warmth of the partner between the VOI-justification condition (M = 4.84, *95% CI:* [4.68, 5.01]) and the utilitarian-justification condition (M = 4.76, *95% CI:* [4.61, 4.91]; β = .09 (*95% CI:* [-.14, .31], $t(506) = .77$, $p = .445$). Participants perceived the partner who gave the VOI justification as less warm compared to the partner who gave the deontological justification (M = 5.28, *95% CI:* [5.12, 5.43]; β = -.43 (*95% CI:* [-.66, -.21]), $t(506) = -3.77$, $p < .001$). (See Figure 2.14A.)

FIGURE 2.14. Results for warmth and competence measures from Study 3.



Analyses employ linear regression. Bracketed values show 95% CI.

A mediation model shows that the decrease in perceived warmth mediates the decrease in transfers in the trust game to the partner who gave the VOI justification compared to the partner who gave the deontological justification (indirect effect = -1.18, *95% CI:* [-2.14, -.52]; bootstrapped indirect effect with 10,000 iterations).

Competence. There was no detected difference in perceived competence of the partner between the VOI-justification condition (M = 5.11, *95% CI:* [4.95, 5.28]) and the utilitarian-justification condition (M = 4.99, *95% CI:* [4.83, 5.14]; β = .13 (95% CI: [-.10, .35]), $t(506)$ = 1.10, p = .272). There was also no detected difference in perceived competence of the partner between the VOI-justification condition and the deontological-justification condition (M = 4.91, *95% CI:* [4.76, 5.06]; β = .20 (*95% CI:* [-.02, .43]), $t(506)$ = 1.78, p = .076). (See Figure 2.14B.)

Discussion

Study 3 shows a boundary condition of the effect of the VOI justification intervention on observer trust. For a dilemma that is less controversial (e.g., generating relatively more utilitarian responses) such as the hospital case, people who respond more in the deontological direction show more trust toward the VOI decision-maker compared to the utilitarian decision-maker, compared to those who respond more in the utilitarian direction. For those who respond more in the utilitarian direction in the hospital case, the utilitarian decision may not seem aversive or cold enough to require a VOI rather than utilitarian justification.

By contrast, for a dilemma that is more controversial (e.g., generating relatively less utilitarian responses) such as the footbridge case, the VOI justification for the utilitarian decision works particularly well in increasing observer trust from both deontologists and utilitarians. Taken together, the results from Studies 1-3 show that when a decision-maker is dealing with a

situation where the utilitarian response is unpopular, the decision-maker is better off defending that decision using a VOI justification.

General Discussion

Across three studies, I investigate how third-party observers perceive decision-makers based on their justifications of their moral decisions. In Study 1, I find that people report more trust toward a decision-maker who expresses a VOI compared to utilitarian justification for the same utilitarian decision, and that this increase in trust is explained by a boost in perceived warmth and competence of the decision-maker. Study 2 shows that people transfer more money in the trust game to the partner who expresses a VOI compared to utilitarian justification. Study 2 also shows that the primary mechanism driving this increase in behavioral trust is an increase in perceived warmth of the partner. Study 3 shows that for moral dilemmas where the utilitarian decision is relatively more popular, there is an interaction effect between participants' own moral judgments and trust of the partner. Taken together, these results show that when a decision-maker is dealing with a situation where the utilitarian response is unpopular, the decision-maker is better off defending the utilitarian response using VOI justification. When a decision-maker is dealing with a situation where people are generally split between the utilitarian and deontological responses, the use of VOI justification primarily increases trust from deontologists.

Theoretical Contributions

These findings contribute to the partner choice account illustrating the fundamentally social role of moral judgments. VOI reasoning signals being a cooperative partner, primarily by signaling warmth. Even though a decision-maker using a VOI justification may not be as trusted compared to a decision-maker using a deontological justification, the decision-maker using a VOI justification may be more trusted compared to the decision-maker using a utilitarian

justification. VOI reasoning mitigates the negative reputational consequences of making utilitarian judgments. Thus, VOI justifications, compared to utilitarian justifications, may increase the selection of utilitarian decision-makers as social partners. Indeed, VOI reasoning may serve the hybrid function of maximizing good consequences, while also signaling respect for the individuality of persons and thus promoting social cooperation.

Second, the finding that people trust VOI decision-makers more than utilitarian decision-makers, and that this is more pronounced for people responding more deontologically (Study 3), aligns with the findings on act-person dissociation (Uhlmann et al., 2013). Research on the distinction between act-centered and person-centered moral judgments suggests that third-party observers can distinguish between the moral acceptability of the act in question and the character of the decision-maker making the moral judgment (Pizarro & Tannenbaum, 2013; Uhlmann, et al., 2013). That is, people can disagree with the moral judgment of the decision-maker (i.e., deontologists can disagree with the utilitarian judgment of the decision-maker), but have a positive impression of that decision-maker's character. Justification of a moral judgment sends an important social signal in cooperative exchanges. The present results show that, through the lens of VOI reasoning, people can disagree with a utilitarian act, but nevertheless endorse the utilitarian decision-maker.

Third, these findings introduce a novel justification for utilitarian judgments. Extant research on people's judgments in moral dilemmas often investigate these judgments as either utilitarian or deontological. However, commonsense morality may be more pluralist (Everett et al., 2016). Moral judgments could be justified using different explanations or procedures, and the present research introduces plurality into procedures for arriving at moral positions. VOI-based utilitarianism may be more viable as a folk psychological theory compared to cost-benefit-based

utilitarianism. VOI-based utilitarianism could have the potential to take hold as commonsense morality, even if it does not surpass the popularity of deontological morality. Rawls' philosophy -- one of the most influential moral philosophies of the 20th century -- is highly favored within academic communities. If Rawlsian morality is also perceived as socially valuable by the general public, then it's possible that veil-of-ignorance reasoning could be used extensively and effectively in the public sphere.

Practical Contributions

The present research shows how VOI reasoning could be used to help justify utilitarian decisions. Decision-makers may fear making utilitarian judgments for fear of being viewed as less trustworthy by others. However, VOI justification of utilitarian decisions may increase trust and perceptions of warmth from third-party observers. VOI justifications could benefit decision-makers and organizations whose genuine aim is to promote the greater good. For example, effective altruists may be perceived by others as lacking warmth and concern for individuals. If effective altruists wish to increase public support for their movement, they could use VOI reasoning to justify making effective donation decisions, and thereby increase observer trust and potentially donations to effective altruist organizations.

VOI reasoning, as a tool for justifying utilitarian decisions, would be particularly relevant for policymakers and leaders. In the context of democratic decision-making, a policymaker's decisions must align with the court of public opinion for these policies to be adopted widely. To illustrate an example of how VOI reasoning could be used to help justify utilitarian decisions, consider policies for universal healthcare coverage. When politicians put forward arguments for universal healthcare, they often appeal to utilitarian justifications. We want to provide health insurance for the greatest number of people. There are sacrifices we would need to make –

namely, higher taxes – but these sacrifices would be justified because the benefits outweigh the costs. Such policies may garner public criticism because nobody wants to pay higher taxes. However, consider another way to justify such a policy: Imagine you have an equal chance of being any citizen in the U.S. Which healthcare policy would you want if you didn't know who you could be? If there is no universal healthcare in place, and you get severely sick in the future, you have a low probability of being among the wealthy few, and thus a high probability of ending up uninsured. If there is universal healthcare in place and you get sick, you will be insured and treated. When people reason about policies from the vantage point of each citizen potentially affected by the policies, without any information that could bias their decisions, people could arrive at and endorse policies that maximizes public welfare.

Future Directions

The present research opens room for several future directions. First, it would be important to investigate how the VOI intervention compares to other interventions in increasing perceived warmth of the decision-maker. Future research could investigate the effects of manipulating warmth independent of the decision-maker's justification. For example, although gender of the decision-maker could increase perceived warmth, prior research has found no effect of target gender on observer liking or trust, such that third-party observers perceive both male and female deontological decision-makers as similarly more likeable and trustworthy, compared to utilitarian decision-makers (Sacco et al., 2017). Another way to manipulate warmth could be for the decision-maker to express emotional conflict, such as communicating the difficulty in making the utilitarian judgment (Everett et al., 2016). It would be useful to investigate whether expressing VOI reasoning is more effective than expressing emotional conflict in increasing perceived warmth.

Second, it would be important to investigate the relationship between VOI reasoning and demographic variables such as socioeconomic status, religiosity, and political affiliation. In the current set of studies, I did not find differences in trust or the person perception measures based on political interest or affiliation. This is likely because the dilemmas used in this set of studies were hypothetical rather than real-life policy dilemmas. Future research could use a real-world policy dilemma, such as universal healthcare or gun control, and in such cases how people perceive a VOI compared to utilitarian justification may depend on their political leanings or party affiliations. Prior research has found greater endorsement of instrumental harm among Republicans and Libertarians compared to Democrats (Iyer et al., 2012; Kahane et al., 2017), suggesting that Democrats may respond more in the deontological direction. As such, the VOI justification, as an intervention in explaining utilitarian decisions, may be more effective for Democrats compared to Republicans.

Third, following up on the importance of VOI justifications for leaders, future research could investigate how the public perceives leaders when they use VOI compared to utilitarian justifications. Such studies could employ specific, real-world policy dilemmas. While the studies in this paper show that a VOI compared to utilitarian justification increases perceived warmth, in the context of policy dilemmas, using a VOI justification may also boost competence compared to utilitarian and deontological justifications. Prior research shows that people view utilitarians as more competent than deontologists (Rom et al., 2017). People may prefer leaders who make more utilitarian decisions, since leaders often need to take an impartial stance by considering the overall positive and negative outcomes of each potential policy (Molinsky & Margolis, 2005; Sunstein, 2005). And since leaders are responsible for the aggregate outcomes of society, they may be judged based on the overall consequences of their actions. People view administrators

who make utilitarian decisions as better leaders, even if they also view these administrators as lacking empathy (Uhlmann et al., 2013). It is likely that voters care about a leader's warmth – for example, in expressing outrage at moral violations – as well as a leader's competence in handling decisions effectively. VOI reasoning could increase perceived competence compared to deontological reasoning, while increasing perceived warmth compared to utilitarian reasoning – thereby increasing overall combined positive perceptions compared to both utilitarian and deontological reasoning. While VOI reasoning could increase behavioral trust through the mechanism of perceived warmth, VOI reasoning may also increase endorsement of the decision-maker as a leader or manager (e.g. in hiring or promotion decisions) via an increase in perceived competence. This would be a fruitful avenue for future research.

References

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition,* 121(1), 154–161.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences,* 36(01), 59–78.

Bentham, J. (1983). The collected works of Jeremy Bentham: Deontology, together with a table of the springs of action; and the article on utilitarianism. Oxford, England: Oxford University Press (Original work published 1789).

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000295

Brooks, A. W., Dai, H., & Schweitzer, M. E. (2014). I'm sorry about the rain! Superfluous apologies demonstrate empathic concern and increase trust. *Social Psychological and Personality Science*, *5*(4), 467-474.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences,* 107, 17433–17438.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion,* 12(1), 2-7.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy.

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200-216.

Everett, Jim A. C. and Pizarro, David A. and Crockett, M. J. (2016) Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, 145 (6). pp. 772-787.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Fried, C. (1978). Right and wrong. Cambridge, MA: Harvard University Press.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*(2), 206-215.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124.

Greene, J. D. (2014). Moral Tribes: Emotion, Reason and the Gap Between Us and Them. London, England: Atlantic Books Ltd.

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), 814–834.

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399-423.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences*, *116*(48), 23989-23995.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE,* 7, e42366. http://dx.doi.org/10.1371/journal.pone.0042366

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2017.) Beyond Sacrificial Harm: A Two-Dimensional Model of Utilitarian Psychology. *Psychological Review*, *125*(2), 131-164.

Kahane, G., Everett, J. A. C., 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.

Kant, I. (2002). Groundwork for the metaphysics of morals. New Haven, CT: Yale University Press (Original work published 1797).

Kogut, T., & Ritov, I. (2005). The ''identified victim'' effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, 18, 157–167.

Krebs, D. (2008). Morality: An Evolutionary Account. *Perspectives on Psychological Science*, 3(3), 149–172.

Levin, Daniel Z., and Rob Cross. "The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer." *Management Science* 50, no. 11 (2004): 1477-1490.

Lydon, J. E., Jamieson, D. W., & Zanna, M. P. (1988). Interpersonal Similarity and the Social and Intellectual Dimensions of First Impressions. *Social Cognition*, 6(4), 269–286.

Mill, J. S. (1863). Utilitarianism. London, England: Parker, Son, and Bourne.

Molinsky, A. L., & Margolis, J. D. (2005). Necessary evils and interpersonal sensitivity in organizations. *Academy of Management Review*, 30, 245–268.

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431-444.

Noë, R., & Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, 35(1), 1–11.

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, *2*(3), 511-527.

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are 'counter-intuitive'deontological judgments really counter-intuitive? An empirical reply to. *Social Cognitive and Affective Neuroscience*, *9*(9), 1368-1371.

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

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In P. Shaver & M. Mikulincer (Eds.), The social psychology of morality: Exploring the causes of good and evil (pp. 91–108). Washington, DC: American Psychological Association.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653-660.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879-891.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods,* 16, 93–115.

Rawls, J. (1971). A theory of justice. Cambridge, MA: Belknap Press of Harvard University Press.

Robichaud, C. (2015). Liberty hospital simulation. Classroom exercise.

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology,* 69, 44–58.

Ross, W. D. (1930). The right and the good. Oxford, England: Oxford University Press.

Sacco, D. F., Brown, M., Lustgraaf, C. J., & Hugenberg, K. (2017). The adaptive utility of deontology: Deontological moral decision-making fosters perceptions of trust and likeability. *Evolutionary Psychological Science*, 3(2), 125–132.

Scanlon, T. M. (1998). What we owe to each other. Vol. 66. Cambridge, MA: Belknap Press of Harvard University Press.

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

Sunstein, C. (2005). Moral heuristics. *Behavioral and Brain Sciences*, 28, 531–542.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395-1415.

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46(1), 35–57.

Twyman, M., Harvey, N., & Harries, C. (2008). Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making*, *3*(1), 111-120.

Uhlmann, E. L., Zhu, L.(. L.)., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, 126(2), 326–334.

Yu, Q and Li, B 2017 mma: An R Package for Mediation Analysis with Multiple Mediators. *Journal of Open Research Software*, 5: 11, DOI: https://doi.org/10.5334/jors.160

Yu, Q., Fan, Y., & Wu, X. (2014). General multiple mediation analysis with an application to explore racial disparities in breast cancer survival. *Journal of Biometrics & Biostatistics*, *5*(2), 1-9.

Supporting Information Appendix:

Chapter 2 Samples and Additional Results

Study 1

Sample. We recruited 402 participants from Amazon's Mechanical Turk (MTurk) in exchange for $2.00 per participant. All participants were U.S. residents. All exclusion criteria were determined *a priori*. We excluded 9 participants with duplicate MTurk IDs, 1 participant who didn't pass the attention check at the beginning of the study, and 128 participants who didn't pass the comprehension checks. These exclusions left a sample of 264 participants (146 male, 118 female; M_{age} = 35.10, SD_{age} = 10.79).

Analyses and Results. Consistent with our predictions, we found that participants in the VOI condition gave more utilitarian responses to the standard footbridge dilemma, as compared to those in the control condition. We estimated the effect of condition on utilitarian judgment (dichotomous response of yes vs. no) using logistic regression. The percentage of participants who made the utilitarian judgment was 37.84% in the VOI condition (*95% CI:* 29.32%, 47.18%) versus 24.18% in the control condition (*95% CI:* 18.05%, 31.59%), which constitutes a 90.83% increase in the odds of making the utilitarian judgment in the VOI condition compared to the control condition, *95% CI:* (12.15%, 226.40%), p = .018. (See Fig. 1.1A.) Likewise, participants rated the utilitarian response as more morally acceptable in the VOI condition (M = 3.32, SD = 2.05) as compared to the control condition (M = 2.70, SD = 1.65), β = .63, *95% CI:* (.18, 1.07), $t(262) = 2.74$, p = .007. (See Fig. 1.2A.)

As expected, participants in the VOI condition tended to give utilitarian responses to the VOI version of the footbridge dilemma: 63.96% preferred that the decision-maker push (*95% CI:* 54.79%, 72.50%), and participants tended to rate the utilitarian response as morally acceptable

(M = 4.46, SD = 2.33).

Study 2

Sample. We recruited 1,506 participants from MTurk in exchange for $2.00 per participant. All participants were U.S. residents. All exclusion criteria were determined *a priori*. We excluded 18 participants with duplicate MTurk IDs, 6 participants who did not pass the attention check, and 588 participants who did not pass the comprehension checks. The exclusions left a sample of 894 participants (420 male, 474 female; M_{age} = 35.44, SD_{age} = 11.01).

Analyses and Results. Consistent with our predictions, we found that participants who responded to the VOI dilemmas before responding to the standard dilemmas gave more utilitarian responses to the standard dilemmas, as compared to those who only responded to the standard dilemmas. We estimated the effect of condition on utilitarian judgment using logistic regression.

For the hospital dilemma, the percentage of participants who made the utilitarian judgment was 53.89% in the VOI condition (*95% CI:* 48.62%, 59.07%) versus 42.60% in the control condition (*95% CI:* 38.51%, 46.78%), which constitutes a 57.51% increase in the odds of making the utilitarian judgment in the VOI condition compared to the control condition, *95% CI:* (20.22%, 106.62%), p = .001. (See Fig. 1.1B.) For the scale measure, we detected an interaction between condition and order, and we therefore first report on the two orders separately. Among participants who saw the AV case first, participants in the VOI condition reported taking the patient off oxygen as more morally acceptable (M = 4.27, SD = 1.86) compared to participants in the control condition (M = 3.23, SD = 1.74), β = 1.05, *95% CI:* (.71, 1.39), $t(890) = 6.03$, $p <$ .001. Among participants who saw the hospital case first, participants in the VOI condition also

reported taking the patient off oxygen as more morally acceptable (M = 3.94, SD = 1.93) compared to participants in the control condition (M = 3.59, SD = 1.75), β = .35, *95% CI:* (.002, .699), $t(890) = 1.97$, $p = .049$. Second, we present the results combined across order: Participants in the VOI condition reported taking the patient off oxygen as more morally acceptable (M = 4.11, SD = 1.90) compared to participants in the control condition (M = 3.40, SD = 1.75), β = .71, *95% CI:* (.47, .95), $t(892) = 5.71$, $p < .001$. (See Fig. 1.2B for presentation of combined results.)

For the AV dilemma, the percentage of participants who made the utilitarian judgment was 83.00% in the VOI condition (*95% CI:* 78.67%, 86.59%) versus 58.32% in the control condition (*95% CI:* 54.14%, 62.38%), which constitutes a 248.89% increase in the odds of making the utilitarian judgment in the VOI condition compared to the control condition, *95% CI:* (152.85%, 387.33%), $p < .001$. (See Fig. 1.1C.) For the scale measure, we detected an interaction between condition and order, and we therefore first report on the two orders separately. Among participants who saw the AV case first, participants in the VOI condition reported swerving as more morally acceptable (M = 5.39, SD = 1.64) compared to participants in the control condition (M = 4.05, SD = 1.83), β = 1.34, *95% CI:* (1.00, 1.68), $t(890) = 7.77$, $p < .001$. For participants who saw the hospital case first, participants in the VOI condition also reported swerving as more morally acceptable (M = 5.12, SD = 1.83) compared to participants in the control condition (M = 4.29, SD = 1.83), β = .83, *95% CI:* (.48, 1.18), $t(890) = 4.70$, $p < .001$. Second, we present the results combined across order: Participants in the VOI condition reported swerving as more morally acceptable (M = 5.25, SD = 1.74) compared to participants in the control condition (M = 4.16, SD = 1.83), β = 1.09, *95% CI:* (.85, 1.33), $t(892) = 8.86$, $p < .001$. (See Fig. 1.2C for presentation of combined results.)

As expected, participants tended to give utilitarian responses to the VOI versions of the dilemmas. For the VOI version of the AV case, 87.23% preferred to be in a state in which the AV is required to swerve to save more lives (*95% CI:* 83.55%, 90.54%), and participants tended to rate this policy requiring swerving as morally acceptable ($M = 5.47$, $SD = 1.63$). Likewise, for the VOI version of the hospital case, 79.83% preferred that the ethics committee use the oxygen for the nine surgeries rather than the single patient (*95% CI:* 75.39%, 83.82%), and participants tended to rate this decision as morally acceptable ($M = 4.77$, $SD = 1.86$).

Study 3

Sample. We recruited 1,409 participants from MTurk, in exchange for $1.00 per participant. All participants were U.S. residents. As in the previous studies, all exclusion criteria were determined *a priori*. We excluded 16 participants with duplicate MTurk IDs, 4 participants who didn't pass the attention check, and 556 participants who didn't pass the comprehension checks. This left a sample of 833 participants (378 male, 455 female; $M_{age} = 35.82$; $SD_{age} = 13.92$).

Analyses and Results. Consistent with our predictions, we found that participants who responded to the VOI charity dilemma before responding to the standard charity dilemma gave more utilitarian responses to the standard charity dilemma, as compared to those who only responded to the standard charity dilemma. Because we detected an effect of order of the charities shown, we estimated the effect of condition in a logistic regression that controls for order. The percentage of participants who made the utilitarian judgment was 62.78% in the VOI condition (*95% CI:* 57.24%, 68.00%), versus 53.87% in the control condition (*95% CI:* 49.55%, 58.12%). Testing for a difference between conditions, there was a 44.46% increase in the odds of

making the utilitarian choice in the VOI condition compared to the control condition, *95% CI:* (8.36%, 93.08%), $p = .013$. (See Fig. 1.1D.) Thus, we find that veil-of-ignorance reasoning increases the likelihood of donating to the more effective charity.

As expected, participants in the VOI condition tended to give utilitarian responses to the VOI version of the charity dilemma, with 80.39% preferring that the decision-maker donate to the charity that would fund the cataract surgeries in India (*95% CI:* 75.73%, 84.54%).

Study 4

Sample. In Study 4, the sample size was determined *a priori* based on a power analysis for a logistic regression capable of detecting an effect size of odds ratio = 1.7 (estimated from Study 2), at power = .90 and using a two-tailed test. According to the power analysis, the targeted final sample size was 1,398. We also took into account an exclusion rate of approximately 33% of the recruitment sample size, determined by exclusion rates in Studies 1-3. Therefore, we aimed to recruit 2,097 participants in order to reach the targeted final sample size of 1,398.

We recruited 2,117 participants from MTurk. Participants completed the study in exchange for $2.00. All participants were U.S. residents. As in the previous studies, all exclusion criteria were determined *a priori*. We excluded 32 participants with duplicate MTurk IDs, 11 participants who did not pass the attention check, and 500 participants who did not pass the comprehension checks. This left a final sample size of 1,574 (676 male, 898 female; M_{age} = 35.31, SD_{age} = 11.24).

Analyses and Results. Consistent with our predictions, we found that participants in the VOI condition gave more utilitarian responses to the standard AV dilemma, as compared to

those who only responded to the standard AV dilemma, and compared to those who responded to the sculpture dilemma prior to the standard AV dilemma. We estimated the effect of condition on utilitarian judgment using logistic regression. The percentage of participants who made the utilitarian judgment was 74.70% in the VOI condition (*95% CI:* 70.32%, 78.63%), versus 50.43% in the simple control condition (*95% CI:* 46.37%, 54.48%), and versus 55.40% in the anchoring control condition (*95% CI:* 51.31%, 59.42%). This constitutes a 190.24% increase in the odds, compared to the simple control condition (*95% CI:* [121.26%, 283.71%], $p < .001$), and a 137.71% increase in the odds compared to the anchoring control condition (*95% CI:* [80.97%, 213.75%], $p < .001$). (See Fig. 1.1E.) Likewise, participants rated the utilitarian policy as more morally acceptable in the VOI condition ($M = 4.91$, $SD = 1.71$), as compared to those in the simple control condition ($M = 3.82$, $SD = 1.86$; $t(1569) = 9.60$, $\beta = 1.09$, *95% CI:* [.87, 1.32], $t(1571) = 9.65$, $p < .001$), and as compared to those in the anchoring control condition ($M = 4.10$, $SD = 1.72$; $\beta = .81$, *95% CI:* [.59, 1.03], $t(1571) = 7.13$, $p < .001$). (See Fig. 1.2D.)

As expected, participants in the VOI condition tended to give utilitarian responses to the VOI version of the AV dilemma, with 84.45% preferring the utilitarian policy (*95% CI:* 80.70%, 87.65%). Also as expected, participants in the anchoring control condition tended to give utilitarian responses to the sculpture dilemma: 98.61% preferred that the decision-maker push the sculpture (*95% CI:* 97.24%, 99.30%). Testing between the VOI condition and anchoring control condition, participants were more likely to make the utilitarian judgment in the sculpture dilemma than in the VOI version of the AV dilemma (1199.08% increase in the odds, *95% CI:* [553.43%, 2864.81%], $p < .001$). Participants were also more likely to rate the utilitarian response in the sculpture dilemma as more morally acceptable ($M = 6.67$, $SD = .78$), compared to the VOI version of the AV dilemma ($M = 5.37$, $SD = 1.58$; $\beta = 1.30$, *95% CI:* [1.15, 1.45], $t(991)$

= 17.04, $p < .001$). Because the sculpture dilemma elicited an even higher utilitarian response than the VOI version of the AV dilemma, this served as a conservative test of the anchoring/generic consistency explanation for our primary results.

Study 5

Sample. In Study 5, the sample size was determined *a priori* by a power analysis as in Study 4, but adjusted for two conditions. According to the power analysis, the targeted final sample size was 932. We also took into account an exclusion rate of approximately 33% of the recruitment sample size, determined by exclusion rates in Studies 1-4. Therefore, we aimed to recruit 1,398 participants in order to reach the targeted final sample size of 932.

We recruited 1,400 participants from MTurk. Participants completed the study in exchange for $2.00. All participants were U.S. residents. As in the previous studies, all exclusion criteria were determined *a priori*. We excluded 31 participants with duplicate MTurk IDs, 35 participants who did not pass the attention check, and 599 participants who did not pass the comprehension checks. This left a final sample size of 735 (354 male, 381 female; M_{age} = 34.58, SD_{age} = 11.33).

Analyses and Results. Consistent with our predictions, we found that participants who responded to the VOI AV dilemma before responding to the standard AV dilemma gave more utilitarian responses to the standard AV dilemma, as compared to those who responded to the reversed-VOI AV dilemma before responding to the standard AV dilemma. We estimated the effect of condition on utilitarian judgment using logistic regression. The percentage of participants who made the utilitarian judgment was 73.43% in the VOI condition (*95% CI:* 68.88%, 77.54%), versus 63.99% in the reversed-VOI control condition (*95% CI:* 58.71%,

68.95%), which constitutes a 55.56% (*95% CI:* 13.64%, 113.29%) increase in the odds, $p = .006$. (See Fig. 1.1F.) Likewise, participants rated the utilitarian policy as more morally acceptable in the VOI condition ($M = 4.89$, $SD = 1.77$), as compared to those in the reversed-VOI control condition ($M = 4.49$, $SD = 1.80$, $\beta = .40$, *95% CI:* (.14, .66), $t(733) = 3.03$, $p = .003$). (See Fig. 1.2E.)

As expected, far more participants tended to give utilitarian responses to the VOI AV dilemma, with 91.48% (*95% CI*: 88.31%, 93.85%) preferring the utilitarian policy requiring swerving, compared to 48.81% (*95% CI:* 43.50%, 54.15%) favoring the utilitarian policy in the reversed-VOI AV dilemma (1025.90% increase in the odds, *95% CI:* [654.83%, 1622.16%], $p < .001$). Likewise, participants tended to rate the utilitarian policy as more morally acceptable in the VOI AV dilemma ($M = 5.60$, $SD = 1.40$), compared to in the reversed-VOI AV dilemma ($M = 4.44$, $SD = 1.91$; $\beta = 1.15$, *95% CI:* [.91, 1.39], $t(733) = 9.39$, $p < .001$).

In Study 5, participants who engaged in standard veil-of-ignorance reasoning, reflecting a principle of impartiality, were more likely to make subsequent utilitarian judgments, compared to those who engaged in a modified form of veil-of-ignorance reasoning bearing no special relation to impartiality. Critically, both conditions involved numerical/probabilistic reasoning and a limited kind of perspective-taking, indicating that these factors, among others, cannot explain the observed effect of the VOI exercise on subsequent moral judgment.

Study 6

Sample. Sample size was determined *a priori* by a (two-tailed) power analysis for a logistic regression to detect an estimated effect size of odds ratio = 1.9, with 24% probability in the control condition, at power = .90. We estimated to detect an effect size of odds ratio=1.9

since this study design was similar to the design of Study 1, where we found an effect size of odds ratio=1.9. We assumed a 24% probability in the control condition due to the 24.18% probability of a utilitarian response to the footbridge case from the simple control condition in Study 1. According to the power analysis, the targeted final sample size was 492 (or 246 per condition). We took into account an exclusion rate of approximately 34% of the recruitment sample size, determined by the exclusion rate of Study 1, which was similar in study design. Therefore, we aimed to recruit 749 participants in order to reach the targeted final sample size of 492.

We recruited 744 participants from MTurk. Participants completed the study in exchange for $2.00. All participants were U.S. residents. We excluded 8 participants with duplicate MTurk IDs, 3 participants who did not pass the first attention check, 16 participants who did not pass the second attention check, and 146 participants who did not pass at least one comprehension check. This left a final sample size of 571 (241 male, 324 female, 3 non-binary, 3 preferred not to answer; M_{age} = 34.15, SD_{age} = 10.97).

Analyses and Results. Consistent with our predictions, we found that participants who responded to the VOI footbridge dilemma before responding to the standard footbridge dilemma gave more utilitarian responses to the standard footbridge dilemma, as compared to those who responded to the utilitarian dilemma before responding to the standard footbridge dilemma. We estimated the effect of condition on utilitarian judgment using a logistic regression. The percentage of participants who made the utilitarian judgment in the standard footbridge dilemma was 36.74% in the VOI condition (*95% CI:* 31.11%, 42.73%), versus 20.52% in the utilitarian-perspective control condition (*95% CI:* 16.37%, 25.41%), which constitutes a 124.96% (*95% CI:* 55.27%, 227.85%) increase in the odds, $p < .001$. (See Fig. 1.1G.) Likewise, participants rated

pushing in the standard footbridge dilemma as more morally acceptable in the VOI condition (M = 3.33, SD = 1.86), as compared to those in the utilitarian-perspective control condition (M = 2.89, SD = 1.83, β = .44, *95% CI:* (.13, .74), $t(569) = 2.83$, $p = .005$). (See Fig. 1.2F.)

Far more participants tended to give utilitarian responses to the utilitarian dilemma, with 95.11% (*95% CI*: 92.05%, 97.03%) indicating that from a utilitarian perspective they would want the person to push, compared to 68.18% (*95% CI:* 62.32%, 73.52%) favoring pushing in the VOI footbridge dilemma (808.44% increase in the odds, *95% CI:* [423.17%, 1581.40%], $p < .001$). Likewise, participants instructed to adopt a utilitarian perspective tended to rate pushing as more morally acceptable in the utilitarian dilemma (M = 6.55, SD = 1.11), compared to in the VOI footbridge dilemma (M = 4.20, SD = 1.99; β = 2.35, *95% CI:* [2.09, 2.61], $t(569) = 17.76$, $p < .001$).[7]

[7] In the preregistration for Study 6 (#27268 on AsPredicted.org), we stated in the last section: "As secondary analyses, we will analyze the responses to the Utilitarian Footbridge case and to the VOI Footbridge case. We predict that participants are more likely to make the utilitarian judgment in the VOI Footbridge case, compared to the Utilitarian Footbridge case." The inclusion of the second sentence was a typo, resulting from a copying error using a preregistration from a prior study. In designing Study 6, our intention was that responses made from a utilitarian perspective in stage 1 of the control condition would be at least as utilitarian as responses made in response to the VOI version of the dilemma (stage 1 of the VOI condition). In the main section of the preregistration for Study 6 we accurately specified our secondary hypothesis about the responses during stage 1 of the control condition: "We predict that participants will tend to give utilitarian responses to the footbridge dilemma when asked to adopt

Thus, participants in the utilitarian-perspective control condition tended to give utilitarian responses to the footbridge dilemma when asked to adopt a utilitarian perspective, but did not tend to give utilitarian responses once they were simply responding to the footbridge dilemma in its standard form. These results provide evidence against the alternative explanation that participants in the VOI condition are simply giving a specific response to a specific dilemma in the first phase, and then giving the same response to the same dilemma in the second phase.

Study 7

Sample. In Study 7, the sample size was determined *a priori* by a power analysis, as in Studies 4-5, capable of detecting an effect size of odds ratio = 1.7. According to the power analysis, the targeted final sample size was 1398. We also took into account an exclusion rate of approximately 33% of the recruitment sample size. Therefore, we aimed to recruit 2,097 participants in order to reach the targeted final sample size of 1,398.

We recruited 2,141 participants from MTurk. Participants completed the study in exchange for $2.00. All participants were U.S. residents. As in the previous studies, all exclusion criteria were determined *a priori*. We excluded 34 participants with duplicate MTurk IDs, 6 participants who did not pass the attention check, and 711 participants who did not pass the comprehension checks. This left a final sample of 1390 (606 male, 784 female; M_{age} = 35.32, SD_{age} = 11.08).

Joe's utilitarian perspective, but we predict that participants will not tend to give utilitarian responses once they are no longer instructed to adopt Joe's perspective and are instead simply responding to the footbridge dilemma in its standard form.")

Analyses and Results. Contrary to our predictions, we found no significant differences in the percentages of participants who gave utilitarian responses to the AV case in the transfer-VOI condition, as compared to the simple control condition, and as compared to the anchoring control condition. We estimated the effect of condition on utilitarian judgment using logistic regression. The percentage of participants who made the utilitarian judgment was 57.38% in the transfer-VOI condition (*95% CI:* 51.70%, 62.88%), as compared to 53.33% in the simple control condition (*95% CI:* 49.38%, 57.25%), and as compared to 54.30% in the anchoring control condition (*95% CI:* 49.80%, 58.72%). This constitutes no detectable difference in the odds compared to the simple control condition (17.82% increase, *95% CI:* [-10.81%, 55.87%], p =.249), or to the anchoring control condition (13.33% increase*, 95% CI:* [-15.31%, 51.86%], p = .401). (See Fig. 1.1H.)

For the scale measure, we found that participants rated the utilitarian response as more morally acceptable in the transfer-VOI condition (M = 4.19, SD = 1.80), as compared to those in the simple control condition (M = 3.94, SD = 1.86), β = .25, *95% CI:* (.003, .507), t(1387) = 1.98, p = .048. However, there were no significant differences in participants' scale responses between the transfer-VOI condition and the anchoring control condition (M = 4.05, SD = 1.78), β = .15, *95% CI:* (-.12, .41), t(1387)= 1.08, p = .280. (See Fig. 1.2G.)

As expected, participants in the transfer-VOI condition tended to give utilitarian responses in the VOI versions of the dilemmas. 83.89% preferred that the decision-maker donate to the more effective charity (*95% CI:* 79.44%, 87.77%), and participants tended to rate donating to the more effective charity as morally acceptable (M = 5.22, SD = 1.69). In the VOI hospital dilemma, we detected an effect of the presentation order of the dilemmas, and therefore we present the utilitarian responses for each order separately: Among participants who saw the VOI

hospital dilemma first, 72.96% preferred that the ethics committee use the oxygen for the nine surgeries rather than the single patient (*95% CI:* 65.72%, 79.46%), and among participants who saw the VOI hospital dilemma second, 87.77% preferred that the ethics committee use the oxygen for the nine surgeries rather than the single patient (*95% CI:* 82.65%, 92.51%). Participants also tended to rate using the oxygen for the nine surgeries rather than the single patient as morally acceptable ($M = 4.68$, $SD = 1.92$).

Furthermore, as expected, participants tended to give utilitarian responses in the prior dilemmas of the anchoring control condition. For the sculpture dilemma, 97.90% preferred that the decision-maker push the sculpture to save two lives (*95% CI:* 96.35%, 98.94%). Likewise, participants tended to rate pushing the sculpture as morally acceptable (we detected an effect of the presentation order of the dilemmas: for those who saw the sculpture dilemma first, $M = 6.61$, $SD = .82$; for those who saw the sculpture dilemma second, $M = 6.41$, $SD = 1.15$). For the speedboat dilemma, 89.73% preferred that the decision-maker borrow the speedboat to save nine lives (*95% CI:* 86.79%, 92.24%), Likewise, participants tended to rate borrowing the speedboat as morally acceptable (we detected an effect of the presentation order of the dilemmas: for those who saw the speedboat dilemma first, $M = 5.89$, $SD = 1.44$; for those who saw the speedboat dilemma second, $M = 6.15$, $SD = 1.27$).

We also compared participants' responses to the prior dilemmas in the transfer-VOI condition and the anchoring control condition. Since the order of the dilemmas in each condition was counterbalanced, we compared participants' responses to the first dilemma they saw, and responses to the second dilemma they saw. For the first dilemma, participants were more likely to make the utilitarian choice in the anchoring control condition (94.13%, *95% CI:* [91.63%, 95.92%]), compared to in the transfer-VOI condition (76.85%, *95% CI:* [71.72%, 81.29%];

383.17% increase, *95% CI:* [206.09%, 681.63%], $p < .001$). For the second dilemma, participants were also more likely to make the utilitarian choice in the anchoring control condition (93.50%, *95% CI:* [90.91%, 95.39%]), compared to in the VOI condition (86.91%, *95% CI:* [82.59%, 90.29%]; 116.64% increase, *95% CI:* [32.22%, 257.79%], $p < .001$). Likewise, for the first dilemma, participants were more likely to rate the utilitarian choice as more morally acceptable in the anchoring control condition ($M = 6.26$, $SD = 1.22$), compared to in the transfer-VOI condition ($M = 4.86$, $SD = 1.95$, $\beta = 1.40$, *95% CI:* (1.18, 1.63), $t(773) = 12.35$, $p < .001$). For the second dilemma, participants were also more likely to rate the utilitarian choice as more morally acceptable in the anchoring control condition ($M = 6.27$, $SD = 1.22$), compared to in the transfer-VOI condition ($M = 5.04$, $SD = 1.69$, $\beta = 1.24$, *95% CI:* (1.03, 1.44), $t(773) = 11.81$, $p < .001$). Thus, as in Study 4, because the prior dilemmas in the anchoring control condition elicited even more utilitarian responses than the prior dilemmas in the transfer-VOI condition, this served as a conservative test of our hypothesis.

These results suggest an important boundary condition of our hypothesized mechanism and thus of the main effect. Although veil-of-ignorance reasoning about a specific case can influence subsequent responses to the standard version of that same case, we find no strong evidence that people spontaneously transfer the effects of veil-of-ignorance reasoning across cases.

TABLE S1. Results for Studies 1-3, both including and excluding participants who failed attention and/or comprehension checks.

	Study 1 Footbridge		**Study 2 Hospital**		**Study 2 AV**		**Study 3 Charity***	
Inclusion criteria	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All
Proportion of participants favoring the utilitarian response in VOI condition	37.84% [29.32%, 47.18%]	44.62% [37.79%, 51.65%]	53.89% [48.62%, 59.07%]	50.07% [46.46%, 53.68%]	83.00% [78.67%, 86.59%]	76.73% [73.54%, 79.65%]	62.78% [57.24%, 68.00%)	65.42% [61.77%, 68.89%]
Proportion of participants favoring the utilitarian response in control condition	24.18% [18.05%, 31.59%]	23.74% [18.32%, 30.16%]	42.60% [38.51%, 46.78%]	39.31% [35.88%, 42.85%]	58.32% [54.14%, 62.38%]	58.30% [54.74%, 61.77%]	53.87% [49.55%, 58.12%]	55.40% [51.70%, 59.05%]
Control condition type	Simple Control	Simple Control	Simple Control	Simple Control	Simple Control	Simple Control	Simple Control	Simple Control
Effect size (odds ratio)	1.91 [1.12, 3.26]	2.59 [1.69, 4.01]	1.58 [1.20, 2.07]	1.55 [1.26, 1.90]	3.49 [2.53, 4.87]	2.36 [1.89, 2.96]	1.44 [1.08, 1.93]	1.52 [1.23, 1.89]
p-value	$p = .018$	$p < .001$	$p = .001$	$p < .001$	$p < .001$	$p < .001$	$p = .013$	$p < .001$
Dependent variable (dichotomous choice)	Standard Footbridge	Standard Footbridge	Standard Hospital	Standard Hospital	Standard AV	Standard AV	Standard Charity	Standard Charity
Total sample size	264	393	894	1488	894	1488	833	1393
Sample size in VOI condition	111	195	347	735	347	735	311	689
Sample size in control condition	153	198	547	753	547	753	522	704
Inclusion rate in VOI condition	56.92%		47.21%		47.21%		45.14%	
Inclusion rate in control condition	77.27%		72.64%		72.64%		74.15%	

*Model controls for order of charities shown

All analyses employ logistic regression. Bracketed values show 95% CI.

TABLE S2. Results for Studies 4-7, both including and excluding participants who failed attention and/or comprehension checks.

	Study 4 AV				**Study 5 AV**		**Study 6 Footbridge**		**Study 7 AV**			
Inclusion criteria	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All
Proportion of participants favoring the utilitarian response in VOI condition	74.70% [70.32%, 78.63%]	71.37% [67.87%, 74.62%]	74.70% [70.32%, 78.63%]	71.37% [67.87%, 74.62%]	73.43% [68.88%, 77.54%]	73.20% [69.78%, 76.36%]	36.74% [31.14%, 42.73%]	36.83% [32.08%, 41.85%]	57.38% [51.70%, 62.88%]	58.64% [54.97%, 62.22%]	57.38% [51.70%, 62.88%]	58.64% [54.97%, 62.22%]
Proportion of participants favoring the utilitarian response in control condition	50.43% [46.37%, 54.48%]	49.71% [46.02%, 53.41%]	55.4% [51.31%, 59.42%]	56.38% [52.67%, 60.02%]	63.99% [58.71%, 68.95%]	70.07% [66.51%, 73.41%]	20.52% [16.37%, 25.41%]	21.70% [17.77%, 26.23%]	53.33% [49.38%, 57.25%]	52.33% [48.65%, 56.00%]	54.30% [49.80%, 58.72%]	54.76% [51.03%, 58.43%]
Control condition type	Simple Control	Simple Control	Anch-oring Control	Anch-oring Control	Reversed-VOI Control	Reversed-VOI Control	Util. Persp. Control	Util. Persp. Control	Simple Control	Simple Control	Anch-oring Control	Anch-oring Control
Effect size (odds ratio)	2.90 [2.21, 3.83]	2.52 [2.02, 3.15]	2.38 [1.81, 3.14]	1.93 [1.54, 2.41]	1.56 [1.14, 2.13]	1.17 [0.92, 1.48]	2.24 [1.55, 3.28]	2.10 [1.52, 2.92]	1.18 [0.89, 1.56]	1.29 [1.05, 1.59]	1.13 [0.85, 1.52]	1.17 [0.95, 1.45]
p-value	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p = .006$	$p = .200$	$p < .001$	$p < .001$	$p = .249$	$p = .017$	$p = .401$	$p = .143$

All analyses employ logistic regression. Bracketed values show 95% CI.

TABLE S2 (Continued). Results for Studies 4-7, both including and excluding participants who failed attention and/or comprehension checks.

Dependent variable (dichotomous choice)	Standard AV	Standard AV	Standard AV	Standard AV	Standard AV	Standard AV	Standard Foot-bridge	Standard Foot-bridge	Standard AV	Standard AV	Standard AV	Standard AV
Total sample size	1,574	2085	1,574	2085	735	1369	571	736	1390	2107	1390	2107
Sample size in VOI condition	419	688	419	688	399	694	264	372	298	706	298	706
Sample size in control condition	581	700	574	697	336	675	307	364	615	707	477	694
Inclusion rate in VOI condition	60.90%		60.90%		57.49%		70.97%		42.21%		42.21%	
Inclusion rate in control condition	83.00%		82.35%		49.78%		84.34%		86.99%		68.73%	

TABLE S3. Results for Studies 1-2, both including and excluding participants who failed attention and/or comprehension checks. All analyses employ linear regression. Bracketed values show 95% CI.

	Study 1 Footbridge		**Study 2 Hospital***		**Study 2 AV***	
Inclusion criteria	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All
Mean rating of moral acceptability of the utilitarian response in VOI condition	3.32 [2.98, 3.67]	3.51 [3.24, 3.77]	4.11 [3.92, 4.30]	3.92 [3.78, 4.06]	5.25 [5.06, 5.44]	5.09 [4.96, 5.23]
Mean rating of moral acceptability of the utilitarian response in control condition	2.70 [2.41, 2.99]	2.70 [2.44, 2.96]	3.40 [3.24, 3.55]	3.34 [3.21, 3.48]	4.16 [4.01, 4.31]	4.19 [4.06, 4.32]
Control condition type	Simple Control	Simple Control	Simple Control	Simple Control	Simple Control	Simple Control
Effect size ß	.63 [.18, 1.07]	.81 [.44, 1.18]	.71 [.47, .95]	.58 [.39, .77]	1.09 [.85, 1.33]	.90 [.71, 1.09]
p-value	$p = .007$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
Dependent variable (scale response)	Standard Footbridge	Standard Footbridge	Standard Hospital	Standard Hospital	Standard AV	Standard AV
Total sample size	264	393	894	1488	894	1488
Sample size in VOI condition	111	195	347	735	347	735
Sample size in control condition	153	198	547	753	547	753
Inclusion rate in VOI condition	56.92%		47.21%		47.21%	
Inclusion rate in control condition	77.27%		72.64%		72.64%	

* Results from combined model. See Study 2 results in SI Appendix for model showing interaction between condition and order of cases presented.

TABLE S4. Results for Studies 4-7, both including and excluding participants who failed attention and/or comprehension checks. All analyses employ linear regression. Bracketed values show 95% CI.

	Study 4 AV				**Study 5 AV**		**Study 6 Footbridge**		**Study 7 AV**			
Inclusion criteria	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All	Passed attention & comp. checks	All
Mean rating of moral acceptability of the utilitarian response in VOI condition	4.91 [4.74, 5.08]	4.83 [4.69, 4.96]	4.91 [4.74, 5.08]	4.83 [4.69, 4.96]	4.89 [4.71, 5.06]	4.86 [4.73, 4.99]	3.33 [3.11, 3.55]	3.32 [3.12, 3.51]	4.19 [3.98, 4.40]	4.26 [4.12, 4.39]	4.19 [3.98, 4.40]	4.26 [4.12, 4.39]
Mean rating of moral acceptability of the utilitarian response in control condition	3.82 [3.68, 3.96]	3.80 [3.67, 3.93]	4.10 [3.96, 4.25]	4.14 [4.01, 4.27]	4.49 [4.30, 4.68]	4.70 [4.57, 4.83]	2.89 [2.69, 3.10]	2.95 [2.75, 3.14]	3.94 [3.79, 4.08]	3.90 [3.77, 4.04]	4.05 [3.88, 4.21]	4.11 [3.97, 4.25]
Control condition type	Simple Control	Simple Control	Anch-oring Control	Anch-oring Control	Reversed-VOI Control	Reversed-VOI Control	Util. Persp. Control	Util. Persp. Control	Simple Control	Simple Control	Anch-oring Control	Anch-oring Control
Effect size ß	1.09 [.87, 1.32]	1.03 [.84, 1.21]	.81 [.59, 1.03]	.68 [.50, .87]	.40 [.14, .66]	.16 [-.02, .35]	.44 [.13, .74]	.37 [.09, .65]	.25 [.003, .51]	.35 [.16, .55]	.15 [-.12, .41]	.15 [-.05, .34]
p-value	$p <$ .001	$p <$ .001	$p <$ .001	$p <$ .001	$p =$ .003	$p =$ .089	$p =$ .005	$p =$ .009	$p =$ .048	$p <$ .001	$p =$ .280	$p =$ .140
Dependent variable (scale response)	Standard AV	Standard AV	Standard AV	Standard AV	Standard AV	Standard AV	Standard Foot-bridge	Standard Foot-bridge	Standard AV	Standard AV	Standard AV	Standard AV
Total sample size	1,574	2085	1,574	2085	735	1369	571	736	1390	2107	1390	2107
Sample size in VOI condition	419	688	419	688	399	694	264	372	298	706	298	706
Sample size in control condition	581	700	574	697	336	675	307	364	615	707	477	694

Inclusion rate in VOI condition	60.90 %		60.90 %		57.49 %		70.97 %		42.21 %		42.21 %	
Inclusion rate in control condition	83.00 %		82.35 %		49.78 %		84.34 %		86.99 %		68.73 %	

Appendix A:

Chapter 3 Measures

Study 1

Attitudinal Trust

How **trustworthy** do you think **Sam** is?

1			4			7
Extremely untrustworthy	2	3	Neutral; not very untrustworthy, but not very trustworthy either	5	6	Extremely trustworthy
○	○	○	○	○	○	○

Person Perception Measures

How good a person do you think **Sam** is? How **moral** are they?

1			4			7
Extremely immoral/bad	2	3	Neutral; not very immoral, but not very moral either	5	6	Extremely moral/good
○	○	○	○	○	○	○

How **competent** do you think **Sam** is?

1						7
Not at all competent	2	3	4 Neutral	5	6	Extremely competent

How **warm or cold** do you think **Sam** is?

1						7
Extremely cold	2	3	4 Neutral	5	6	Extremely warm

How **capable** do you think **Sam** is?

1						7
Not at all capable	2	3	4 Neutral	5	6	Extremely capable

How **sociable** do you think **Sam** is?

1						7
Not at all sociable	2	3	4 Neutral	5	6	Extremely sociable

How good a **friend** do you think **Sam** would make?

1						7
An extremely bad friend	2	3	4 Neutral	5	6	An extremely good friend
○	○	○	○	○	○	○

How good a **spouse (marriage partner)** do you think **Sam** would make?

1						7
An extremely bad spouse	2	3	4 Neutral	5	6	An extremely good spouse
○	○	○	○	○	○	○

How good a **boss** do you think **Sam** would make?

1						7
An extremely bad boss	2	3	4 Neutral	5	6	An extremely good boss
○	○	○	○	○	○	○

How good a **President (of the United States)** do you think **Sam** would make?

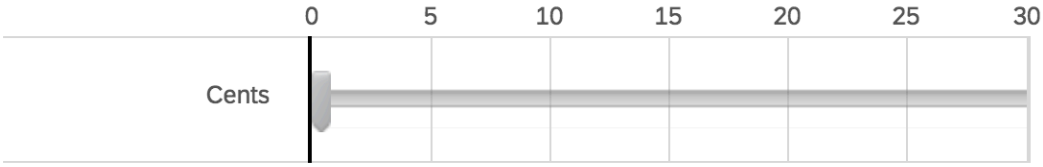
1						7
An extremely bad President	2	3	4 Neutral	5	6	An extremely good President
○	○	○	○	○	○	○

Studies 2-3

Transfer in Trust Game

How many cents (from $0.00 to $0.30) do you want to transfer to Person B?

Each cent you give is doubled, and your final bonus will depend on how much you transfer and how much **Person B** returns.

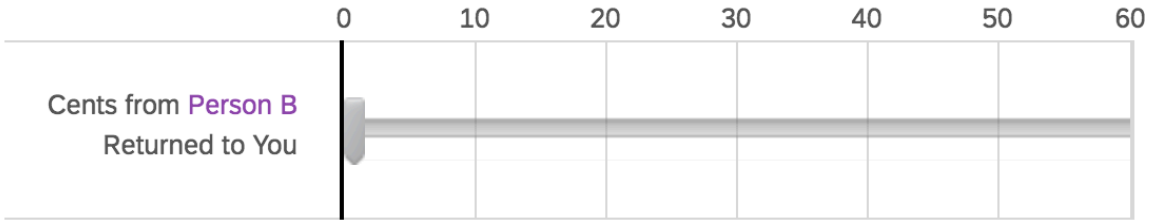


Predicted Return in Trust Game

You decided to transfer **${q://QID266/ChoiceNumericEntryValue/1} cents.**

This amount was doubled by us and added to **Person B**'s initial 30 cent bonus.

Of the total amount of money that Person B now has, how many cents do you predict Person B will return back to you?



Benevolence-Based Trust

Please provide your best guess for the following statements:

Person B...

	1 = Not at all	2	3	4	5	6	7 = Very much so
... would go out of their way to hurt me	○	○	○	○	○	○	○
... would look out for my best interests	○	○	○	○	○	○	○
... would make sure I was not harmed	○	○	○	○	○	○	○
... would care about what happened to me	○	○	○	○	○	○	○

Competence-Based Trust

Please provide your best guess for the following statements:

Person B...

	1 = Not at all	2	3	4	5	6	7 = Very much so
... intelligently explains their reasoning procedure	○	○	○	○	○	○	○
... is knowledgeable on how to make difficult decisions	○	○	○	○	○	○	○
... approaches decision-making with professionalism and expertise	○	○	○	○	○	○	○
... is well-qualified and prepared to make difficult decisions	○	○	○	○	○	○	○

Warmth

Please provide your best guess for the following traits describing **Person B:**

Person B is...

	1 = Not at all	2	3	4	5	6	7 = Very much so
Fair	○	○	○	○	○	○	○
Generous	○	○	○	○	○	○	○
Helpful	○	○	○	○	○	○	○
Honest	○	○	○	○	○	○	○
Righteous	○	○	○	○	○	○	○
Sincere	○	○	○	○	○	○	○
Tolerant	○	○	○	○	○	○	○
Understanding	○	○	○	○	○	○	○
Friendly	○	○	○	○	○	○	○
Warm	○	○	○	○	○	○	○

Competence

Please provide your best guess for the following traits describing **Person B:**

Person B is...

	1 = Not at all	2	3	4	5	6	7 = Very much so
Intelligent	○	○	○	○	○	○	○
Skillful	○	○	○	○	○	○	○
Capable	○	○	○	○	○	○	○
Clever	○	○	○	○	○	○	○
Competent	○	○	○	○	○	○	○
Foresighted	○	○	○	○	○	○	○
Innovative	○	○	○	○	○	○	○
Creative	○	○	○	○	○	○	○
Knowledgeable	○	○	○	○	○	○	○
Professional	○	○	○	○	○	○	○

Perceived Empathy

Please provide your best guess for the following statements:

Person B...

	1 = Not at all	2	3	4	5	6	7 = Very much so
... cares about each person's harm	○	○	○	○	○	○	○
... has empathy for each person	○	○	○	○	○	○	○
... considers the experience of each person	○	○	○	○	○	○	○
... considers the suffering of each person	○	○	○	○	○	○	○

Perceived Perspective-Taking

Please provide your best guess for the following statements:

Person B...

	1 = Not at all	2	3	4	5	6	7 = Very much so
... tries to imagine how s/he would feel if s/he were in someone else's place	○	○	○	○	○	○	○
... tries to imagine how things look from other people's points of view	○	○	○	○	○	○	○
... tries to look at each person's perspective before making a decision	○	○	○	○	○	○	○
... usually tries to put himself/herself in other people's shoes	○	○	○	○	○	○	○

Perceived Humanization

Please provide your best guess for the following statements:

Person B...

	1 = Not at all	2	3	4	5	6	7 = Very much so
... thinks about people as human beings rather than just as numbers	○	○	○	○	○	○	○
... treats people as a mere means to an end	○	○	○	○	○	○	○
... treats people as individuals with their own experiences	○	○	○	○	○	○	○
... treats people like inanimate objects	○	○	○	○	○	○	○

Appendix B:

Chapter 3 Stimuli of Decision Justifications

Study 1

VOI Condition

Sam thought about this moral dilemma.

Sam said that the decision-maker **should push the person wearing the backpack onto the tracks to save the five workers.**

Sam said:

"I believe it is most important to think about this decision from an impartial perspective. What would I want the decision-maker to do if I had an equal chance of being each of these six people affected by the decision? Imagine I had a 1 out of 6 chance of being the person wearing the backpack, and a 5 out of 6 chance of being one of the people on the tracks. If I didn't know who I was going to be, I would want the decision-maker to push the person wearing the backpack, since this would mean a greater chance of living. But when I think about pushing, it feels wrong. Nevertheless, if pushing is what I'd want from an impartial perspective, not knowing who I was going to be, then perhaps it really is the right thing to do, even if it feels wrong."

Utilitarian Condition

Sam thought about this moral dilemma.

Sam said that the decision-maker **should push the person wearing the backpack onto the tracks to save the five workers.**

Sam said:

"I believe that it is most important to think about the total costs and benefits associated with each option. If the decision-maker pushes the person onto the tracks, 1 person dies and 5 people live. If the decision-maker does not push the person onto the tracks, then 5 people die and 1 person lives. Pushing the person off the tracks produces the best overall balance of costs and benefits. I think that it is better to save many lives rather than just one."

Deontological Condition

Sam thought about this moral dilemma.

Sam said that the decision-maker **should not push the person wearing the backpack onto the tracks to save the five workers.**

Sam said:

"I believe that it is most important to think about the fundamental rights of persons. I think that it is wrong to sacrifice one person to save the five people on the tracks. I think that killing is just wrong regardless of the consequences."

Study 2

VOI Condition

When given this moral dilemma, **Person B** (the MTurker that you have been matched with) said that the decision-maker **should push the person wearing the backpack onto the tracks to save the five workers.**

Person B said:

> *"I believe it is most important to think about this decision from an impartial perspective. What would I want the decision-maker to do if I had an equal chance of being each of these six people affected by the decision? Imagine I had a 1 out of 6 chance of being the person wearing the backpack, and a 5 out of 6 chance of being one of the people on the tracks. If I didn't know who I was going to be, I would want the decision-maker to push the person wearing the backpack, since this would mean a greater chance of living. Imagining I could be in the shoes of each of these people equally, I would want the decision-maker to make the most impartial choice."*

Utilitarian Condition

When given this moral dilemma, **Person B** (the MTurker that you have been matched with) said that the decision-maker **should push the person wearing the backpack onto the tracks to save the five workers.**

Person B said:

> *"I believe that it is most important to think about the total costs and benefits associated with each option. If the decision-maker pushes the person onto the tracks, 1 person dies and 5 people live. If the decision-maker does not push the person onto the tracks, then 5 people die and 1 person lives. Pushing the person off the tracks produces the best overall balance of costs and benefits. I think that it is better to save many lives rather than just one."*

Deontological Condition

When given this moral dilemma, **Person B** (the MTurker that you have been matched with) said that the decision-maker **should not push the person wearing the backpack onto the tracks to save the five workers.**

Person B said:

> *"I believe that it is most important to think about the fundamental rights of persons. I think that it is wrong to sacrifice one person to save the five people on the tracks. I think that killing is just wrong regardless of the consequences."*

Study 3

VOI Condition

When given this moral dilemma, **Person B** (the other participant you have been matched with) said that the ethics committee **should take the patient off oxygen.**

Person B said:

> *"I believe that it is most important to think about this decision from an impartial perspective. What would I want the ethics committee to do if I had an equal chance of being each one of these ten patients affected by the decision? Imagine I had a 1 out of 10 chance of being the patient already at the hospital, and a 9 out of 10 chance of being an incoming patient from the earthquake. If I didn't know who I was going to be, I would want the ethics committee to take the patient off oxygen, since this would mean an overall greater chance of living. Imagining I could be in the shoes of each of these patients equally, I would want the ethics committee to make the most impartial choice."*

Utilitarian Condition

When given this moral dilemma, **Person B** (the other participant you have been matched with) said that the ethics committee **should take the patient off oxygen.**

Person B said:

> *"I believe that it is most important to think about the total costs and benefits associated with each option. If the ethics committee takes the patient off oxygen, 1 person dies and 9 people live. If the ethics committee keeps the patient on oxygen, then 9 people die and 1 person lives. Taking the patient off oxygen produces the best overall balance of costs and benefits. I think that it is better to save many lives rather than just one."*

Deontological Condition

When given this moral dilemma, **Person B** (the other participant you have been matched with) said that the ethics committee **should not take the patient off oxygen.**

Person B said:

> *"I believe that it is most important to think about the fundamental rights of patients. I think that it is wrong to sacrifice one patient to save the 9 incoming patients. I think that killing is just wrong regardless of the consequences."*

Appendix C:

Chapter 3 Comprehension Checks

Study 1

In the dilemma you encountered there were five people on the tracks and one person on the footbridge next to the decision-maker. Of these six people, what proportion of them will die if the decision-maker decides to push?

- 6 out of 6 (100%)
- 5 out of 6 (83%)
- 3 out of 6 (50%)
- 1 out of 6 (17%)
- 0 out of 6 (0%)

To confirm you've understood the information given about **Sam**, please indicate what **Sam** said that the decision-maker should do.

- Sam said that the decision-maker should sacrifice the person wearing the backpack to save the others
- Sam said that the decision-maker should NOT sacrifice the person wearing the backpack to save the others

Which of the following **best** describes the thought process of **Sam**?

- No information about this was given
- Thought about the total costs and benefits of each option
- Thought about what s/he would want if s/he had an equal chance of being each of the people affected by the decision
- Thought about the fundamental rights of persons
- None of the above

Study 2

All conditions

To confirm you've understood the information given about the person you're playing with, please indicate what **Person B** said that the decision-maker should do.

- Person B said that the decision-maker should sacrifice the person wearing the backpack to save the others
- Person B said that the decision-maker should NOT sacrifice the person wearing the backpack to save the others

Which of the following **best** describes the thought process of **Person B**?

- No information about this was given
- Thought about the total costs and benefits of each option
- Thought about what s/he would want if s/he had an equal chance of being each of the people affected by the decision
- Thought about the fundamental rights of persons
- None of the above (describe below)

In the moral dilemma presented at the start of the study, there were five people on the tracks and one person on the footbridge next to the decision-maker. Of these six people, what proportion of them will die if the decision-maker decides to push?

- 6 out of 6 (100%)
- 5 out of 6 (83%)
- 3 out of 6 (50%)
- 1 out of 6 (17%)
- 0 out of 6 (0%)

Deontological

Which of the following *most accurately characterizes* **Person B**'s reasoning in response to the dilemma presented at the beginning?

- "It is wrong to sacrifice one person to save the five people on the tracks"
- "It is wrong to sacrifice five people to save the one person on the tracks"
- "It is not wrong to sacrifice one person to save the five people on the tracks"
- "It is wrong to sacrifice one person to save the seven people on the tracks"

Utilitarian

Which of the following *most accurately characterizes* **Person B**'s reasoning in response to the dilemma presented at the beginning?

- "If the decision-maker pushes the person onto the tracks, 1 person dies and 5 people live"
- "If the decision-maker does not push the person onto the tracks, 1 person dies and 5 people live"
- "If the decision-maker does not push the person onto the tracks, 1 person dies and 6 people live"
- "If the decision-maker pushes the person onto the tracks, 5 people die and 1 person lives"

VOI

Which of the following *most accurately characterizes* **Person B**'s reasoning in response to the dilemma presented at the beginning?

- "Imagine I had a 5 out of 6 chance of being the person wearing the backpack, and a 1 out of 6 chance of being one of the people on the tracks"
- "Imagine I had a 1 out of 6 chance of being the person wearing the backpack, and a 5 out of 6 chance of being one of the people on the tracks"
- "Imagine I had a 1 out of 6 chance of being the decision-maker, and a 5 out of 6 chance of being one of the people on the tracks"
- "Imagine I had a 1 out of 7 chance of being the person wearing the backpack, and a 1 out of 7 chance of being the decision-maker"

Study 3

All Conditions

To confirm you've understood the information given about the person you're playing with, please indicate what **Person B** said that the ethics committee should do.

- Person B said that the ethics committee should take the patient off oxygen to save the nine incoming patients
- Person B said that the ethics committee should NOT take the patient off oxygen to save the nine incoming patients

Which of the following **best** describes the thought process of **Person B**?

- No information about this was given
- Thought about the total costs and benefits of each option
- Thought about what s/he would want if s/he had an equal chance of being each of the patients affected by the decision
- Thought about the fundamental rights of patients
- None of the above (describe below)

In the moral dilemma presented at the start of the study, there were nine incoming patients and one patient already at the hospital. Of these ten people, what proportion of them will die if the ethics committee decides to take the patient at the hospital off oxygen?

- 10 out of 10 (100%)
- 9 out of 10 (90%)
- 5 out of 10 (50%)
- 1 out of 10 (10%)
- 0 out of 10 (0%)

Deontological

Which of the following *most accurately characterizes* **Person B**'s reasoning in response to the dilemma presented at the beginning?

- "It is wrong to sacrifice one patient to save the nine incoming patients"
- "It is wrong to sacrifice nine patients to save the one incoming patient"
- "It is not wrong to sacrifice one patient to save the nine incoming patients"
- "It is wrong to sacrifice one patient to save the ten incoming patients"

Utilitarian

Which of the following *most accurately characterizes* **Person B**'s reasoning in response to the dilemma presented at the beginning?

- "If the ethics committee takes the patient off oxygen, 1 person dies and 9 people live"
- "If the ethics committee does not take the patient off oxygen, 1 person dies and 10 people live"
- "If the ethics committee does not take the patient off oxygen, 1 person dies and 9 people live"
- "If the ethics committee takes the patient off oxygen, 9 people die and 1 person lives"

VOI

Which of the following *most accurately characterizes* **Person B**'s reasoning in response to the dilemma presented at the beginning?

- "Imagine I had a 9 out of 10 chance of being the patient already at the hospital, and a 1 out of 10 chance of being one of the incoming patients"
- "Imagine I had a 1 out of 10 chance of being the patient already at the hospital, and a 9 out of 10 chance of being one of the incoming patients"
- "Imagine I had a 1 out of 10 chance of being on the ethics committee, and a 9 out of 10 chance of being one of the incoming patients"
- "Imagine I had a 1 out of 10 chance of being the patient already at the hospital, and a 9 out of 10 chance of being on the ethics committee"