Cognitive Reflection Drives Utilitarian Judgment in Self-Sacrificial and Other-Sacrificial Dilemmas: Applying Process Dissociation and Behavioral Validation to Moral Dilemmas in Multiple Relational Contexts

by

David Simpson

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology
University of Alberta

ABSTRACT

There is considerable evidence linking cognitive reflection with utilitarian judgments in dilemmas that involve sacrificing someone else for the greater good. However, the evidence is mixed on the question of whether cognitive reflection is associated with utilitarian judgments in *self-sacrificial* dilemmas. We employed process dissociation to extract a self-sacrificial utilitarian (SU) parameter, an altruism (A) parameter, an other-sacrificial (OU) utilitarian parameter, and a deontology (D) parameter. In Study 1, the cognitive reflection test (CRT) positively correlated with both SU and OU (replicated in Studies 2 and 4, pre-registered). In Study 2, we found that instructing participants to rely on reason increased SU and OU (replicated in Study 4, pre-registered). In Study 3, we found that SU and OU positively correlated with giving in the single-game version of the public goods game (replicated in Study 4, pre-registered), which provides behavioral validation that they are genuine moral tendencies. We then pooled the samples from Studies 1-4 together (N = 1,418) and performed a principal component analysis, which broke SU and OU each into two components. Each component was positively correlated with the CRT. Additionally, we identified three clusters of participants by applying k-means cluster analysis to the moral dilemmas. The cluster with significantly higher SU and OU scores than the other clusters also had significantly higher CRT scores. In Study 5, we found that SU was less influenced by relationship context than A. Together, these studies constitute strong cumulative evidence that SU and OU are both valid measures that are associated with reliance on cognitive reflection.

Keywords: Self-sacrifice, utilitarianism, dual-process, cognitive reflection.

PREFACE

The studies described in this dissertation were approved by the University of Alberta Ethical Review Board (Pro00117161). Additionally, the first four studies have been accepted for publication as a single paper. Kyle Nash, my supervisor, is listed as a second author in that paper. His role was supervising my research and offering input on early drafts. I came up with the hypotheses, designed the experiments, collected the data, analyzed the data, and wrote the paper.

DEDICATION

This dissertation is dedicated to my family. To my Dad, who never fails to make me laugh. To my Mom, who is my best encourager. To Daniel, who models what hard work looks like. To Nathan, who is my favorite person to discuss politics with. And to Grace, who will always be my baby sister.

ACKNOWLEDGEMENTS

I would like to thank Kyle Nash for his supervision over the last 3 years. I could not have asked for a better supervisor. I am very grateful to Shafa Mazidi for being better at coding than me, and to Paige Faulkner for being better at lab administration than me. I also have to give a shoutout to the late Albert Ellis, who has been my therapist via his writings. Lastly, I am very thankful to the students I taught in Fall 2023, Winter 2024, and Spring 2024. It was my first time teaching university courses, and you all gave me an exceedingly warm welcome.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Background

Moral judgment is ubiquitous in human life. It is with good reason that Pinker (2002) called us “the sanctimonious animal”. Given that we are products of a competitive process of natural selection, it initially seems surprising that humans would have any moral tendencies towards non-kin at all. The theory of reciprocal altruism is perhaps the most influential evolutionary explanation of where the basic ingredients of our moral tendencies came from. Trivers (1971) was the first to spell this theory out. He defined altruism as an action which is apparently costly to one organism (in terms of fitness), but benefits another (genetically unrelated) organism. If the organism receiving the benefit reciprocates in the future, this can be a net benefit to both organisms so long as the benefit to the receiver is greater than the cost to the giver. However, if an organism has a genetic predisposition to reap the benefits of the altruism of others without reciprocating in the future, that organism will leave more descendants than organisms with a genetic predisposition to be altruistic towards everyone. As a result, the evolutionary advantages of reciprocal altruism are only possible if mechanisms for cheater detection also evolve (Brown & Moore, 2000).

Trivers (1971) explains how selection pressures related to reciprocal altruism could give rise to various moralistic tendencies in our species. For example, moralistic aggression or anger serves the function of motivating people to punish those who receive altruism from others but do not reciprocate. Similarly, the function of the feeling of guilt is to motivate people to reciprocate the altruism of others. The theory of reciprocal altruism has been used to explain (along similar lines) the origin of a variety of other moral tendencies in humans, including a sense of fairness

(Trivers, 2006), norms for forgiveness (Krebs, 2008), and generosity[1] (Komter, 2010). There is evidence of simpler forms of moralistic tendencies (that seem to serve similar functions) in some of our primate relatives. For example, Capuchin monkeys display anger when they are given a less desirable reward compared to their neighbor who does the same task, which indicates that they have a rudimentary sense of fairness (Brosnan & de Waal, 2003). The fact that we share some moralistic emotions with our primate relatives is consistent with those tendencies having an evolutionary origin.

Given the likely evolutionary origins of our basic moralistic tendencies, Street (2017) argues that it is unlikely that our moral judgments track anything like objective truth. In order to be "objectively true", a statement must be true independently of whether anyone believes it to be true. The view that at least some moral statements are objectively true is known as moral realism (Pölzler, 2018). Street (2007) argues that there does not appear to be a way for objective moral facts to exert selection pressure. It is clear how various *physical* facts could constrain what genes propagate. For example, it is an objective physical fact that falling from heights causes death, and this is likely the evolutionary basis for our predisposition for fearing heights (Jackson & Cormack, 2007). Suppose it is objectively true that breaking a promise is morally wrong. How could this exert any selection pressure? There is no clear answer. However, there is a story to be told about why we have these intuitions (sketched in the previous paragraphs) in terms of reciprocal altruism and other evolutionary factors. That story does not need to appeal to objective moral truths.

[1] It must be noted that reciprocal altruism is not the only possible evolutionary explanation of morality. Other theorists have tried to explain the evolution of morality in terms of sexual selection (Miller, 2007), indirect reciprocity (Wedekind & Braithwaite, 2002), and, more controversially, group selection (Haidt, 2007).

Even though there is no good reason to think our moral intuitions track objective truth, humans still have the ability to reason about their moral intuitions. There have been various attempts at producing a philosophical theory which can systematize these intuitions. Rawls (1971, as cited in De Lazari-Radek & Singer, 2014) construes these as attempts to achieve reflective equilibrium. The process envisioned by Rawls goes as follows. We start with our intuitive moral judgments, which initially seem disparate and unrelated. For example, most people judge that physical assault, rape, and murder are wrong. We then try to develop a set of more basic principles which entail our various moral judgments. For example, one might posit the principle that "performing actions which cause suffering are wrong". This principle would explain the wrongfulness of actions like physical assault, because they cause suffering.

If any of our posited moral principles conflict with judgments about particular actions, we can reject or revise the principles. For example, most people have the judgment that performing surgery is morally permissible even though it causes suffering. This can lead to modifying the principle to say something like "performing actions which cause suffering are wrong, unless they lead to a greater overall reduction in suffering". This would entail that surgery is morally permissible since it decreases overall suffering (e.g. by removing a cancerous tumor), despite causing some suffering in the short term. If a moral principle explains a wide variety of judgments, it can be used to override or change our judgments, especially the ones we are less confident in. Many people would initially form the judgment that it is wrong to torture someone to get information about a ticking timebomb. But given that the amount of overall suffering is reduced by such an action, the principle "performing actions which cause suffering are wrong, unless they lead to a greater overall reduction in suffering" (which explains a wide variety of other judgments) means that we should abandon our initial judgment and say that torture would

be permissible in such a case. This back-and-forth process of negotiating between our initial moral judgments and our posited moral principles continues until we arrive at an internally coherent framework. The situation in which we are willing to, upon reflection, endorse our overall framework is called reflective equilibrium.

Different moral theories have different sets of moral principles. Consequentialist moral theories are those that use the consequences or results of an action as the criteria for determining the rightness or wrongness of said action. The most widely discussed version of consequentialism is utilitarianism, which was originally developed by Jeremy Bentham (1789). According to utilitarianism, the only relevant consequences are conscious states of happiness and suffering, both construed broadly as conscious states that are pleasant or unpleasant to experience. Actions which increase overall happiness are morally good, and those which increase overall suffering are morally bad (De Lazari-Radek & Singer, 2014).

Utilitarianism is often contrasted with deontological theories, which posit that some actions are intrinsically right or wrong, regardless of their consequences. One early proponent of a deontological approach to ethics was Immanuel Kant (1785). One of his proposals was that we should decide which principles to adopt on the basis of whether the principles are universalizable. To say that a principle is universalizable is to say that we could will that it be universalized without contradiction. For example, Kant argued that we could not consistently will that the principle "always lie" be followed. By his criteria, this implies that lying is morally wrong, regardless of the consequences. Additionally, he argued that humans (or any rational beings) exist as ends in themselves with intrinsic value. Because of this, it is always wrong to treat people as a mere means to an end. One ought always to treat people as ends in themselves. The key distinction between utilitarianism and deontology can be summarized as follows. In

utilitarianism, it is conscious states that have the highest importance; in deontology, it is the universalizability of maxims and the dignity of persons that have the highest importance.

One set of thought experiments that bring out the distinction between utilitarianism and deontology are the famous trolley dilemmas, which were originally proposed by Foot (1978). In Thomson's (1985) slightly modified version (which has become more famous), there are two dilemmas. The first version (which is known as the switch dilemma) involves deciding whether to pull a switch that will redirect a trolley from a track with five people to a track with one person. The second version (known as the footbridge dilemma) involves deciding whether to push a large man in front of a trolley to stop it from hitting five people. Most ordinary people have the intuition that killing one to save five is morally right in the switch case, but morally wrong in the footbridge case (Petrinovich & O'Neill, 1996). Utilitarianism implies that both decisions are morally right since they result in more lives saved. Foot (1978) argues that there is a morally relevant difference between the two cases, in terms that are broadly speaking deontological. In the switch case, the death is incidental (it is not the means by which the five are being saved), whereas in the footbridge case, the person's death is necessary for saving the five. In other words, the footbridge case involves using a person as a trolley-stopper. Foot (1978) cashes out this distinction in terms of the doctrine of double effect: it is permissible to do an action that foreseeably results in harm, as long as that harm is not an intended outcome of the action. On her account, someone pulling the switch is not aiming at the death of another person, it is just a regrettable (albeit foreseeable) side effect. By contrast, someone pushing someone in front of the trolley is aiming at their death. This can be seen as an extension of the deontological principle that we should not use people as a mere means, but always treat them as ends in themselves.

The standard philosophical approach to dilemmas like these is to put forward a priori arguments for the utilitarian or deontological answer. Greene (2015) took a different approach, and theorized that deontological and utilitarian responses to these dilemmas reflect different psychological processes. Greene (and various other researchers) have used batteries of dilemmas involving the decision of whether to harm someone else to save many. In this dissertation, these dilemmas will be referred to as "other-sacrificial dilemmas". The point of these dilemmas is not to predict real-world moral behavior, but rather to measure conflicting psychological processes (Plunkett & Greene, 2019). Greene (2008) conjectures that the reason that these two tendencies have become popular in philosophy is that they reflect two different natural psychological processes, which is consistent with evidence from a twin study finding that responses to these types of dilemmas are substantially heritable[2] (Smith & Hatemi, 2020).

Greene (2015) has tried to explain characteristically deontological and characteristically utilitarian responses in terms of the dual-process model. It is widely thought in cognitive psychology that humans reason using two systems: system 1, which is quick, automatic and intuitive, and system 2, which is slow, deliberative and rational[3] (Stanovich & West, 2000). Greene (2015) calls these two systems "automatic mode" and "manual mode", based on an analogy with cameras. Automatic mode is inflexible, not always conscious, and is rooted in emotion. This type of thinking is associated with activity in the medial frontal regions and the amygdala (Phan et al., 2002). Manual mode is flexible, conscious, slow, and rooted in

[2] Surprisingly, Graham and Haidt's (2009) five moral foundations (care, fairness, loyalty, obedience, and disgust) show little sign of being heritable (Smith et al., 2017).

[3] This division into two systems is clearly an oversimplification, but it is in line with the approach that Dennett (2007) calls "homuncular functionalism". The goal of this approach is to take the agency of a person and break it down into sub-agencies, without ascribing all of the competencies of the agent to the sub-agencies. The sub-agents can in turn be thought of as made of smaller sub-systems, and this process of reduction is meant to continue until the sub-systems are so simple that they could be replaced by a machine.

deliberation. Research indicates that activity in the dorsolateral prefrontal cortex (DLPFC) is correlated with a variety of types of abstract reasoning, and with controlled, conscious cognition (Braver et al, 2009; Yang et al., 2009), which fall under what Greene calls manual mode.

There is a large body of evidence indicating that utilitarian responses to other-sacrificial dilemmas are associated with reliance on manual mode[4] (Patil, 2021). It is on the basis of this evidence that Greene (2015) argues that utilitarian judgments are more reliable. He does this by appealing to what he calls the "no cognitive miracles principle". He argues that automatic mode (quick, intuitive thinking) can only be reliable insofar as it has been shaped by "trial and error experience", which can include biological evolution, cultural development, or individual trial-and-error experience. If automatic mode were reliable in the absence of one of these three sources of adaptation, that would be a kind of cognitive miracle. Therefore, we should only rely on automatic mode for what Greene calls "familiar moral problems", which are those with which we have either evolutionary, cultural, or personal experience. For unfamiliar problems, the ones we do not have trial and error experience with, we should rely on manual mode. Trolley dilemmas involve a novel technological situation that we did not evolve with, so there is prima facie reason to classify it as an unfamiliar moral problem that our intuitions will not be suited to solving. Additionally, Greene says that the existence of widely differing intuitions about a moral dilemma is an indication that a moral problem is unfamiliar. This condition is clearly satisfied by trolley-type dilemmas, and so Greene concludes that the outputs of manual mode will be more reliable for those cases.

If by "reliable" Greene meant something like "reliably tracking objective moral truth", this would be subject to Street's (2007) argument that was described earlier. However, Greene

[4] The body of this dissertation will review and build on this body of evidence.

(2002) has written in defense of the idea that moral statements are not capable of being objectively true or false, so he would not be subject to the above criticism. To a first approximation, when he says a certain type of moral judgment is more "reliable", he means that the judgment is capable of creating consensus. We see this in his appeal to the common currency argument, which he also uses to support utilitarianism. This argument goes as follows. Different societies have converged on different sets of moral norms. They do this to ensure cooperation within their particular societies. But now that we live in an increasingly globalized world, we need a common moral framework that we can use to adjudicate disagreements between societies. In other words, we need a "common currency". Greene (2013) argues that although different societies disagree on many things, they at least all agree that, all else being equal, happiness is better than suffering. He concludes that utilitarianism is a plausible candidate for being a common currency for adjudicating disagreements between societies, since it is based entirely on the value of happiness and the disvalue of suffering. This line of argument presupposes that the criterion of adequacy for a moral theory is its ability to create consensus. Thus, his argument that utilitarian judgments are more reliable should not be taken to imply any commitment to moral realism.

To sum up, Greene argues that a) manual mode is more reliable for unfamiliar moral problems, b) manual mode tends to produce utilitarian responses to unfamiliar problems, and concludes that c) utilitarian responses are more reliable for unfamiliar moral problems. Lott (2016) objects to Greene's argument on the grounds that the psychological origin of a judgment is irrelevant to whether that judgment is justified. He gives incest as an example. There is evidence that intuition, or what Greene calls automatic mode, leads to our judgment that incest is wrong (Haidt, Bjorklund, & Murphy, 2000). There is a fairly straightforward evolutionary story

to tell about why we tend to make this judgment, having to do with the deleterious effects of inbreeding (Lieberman et al., 2003). The moral question of what judgment to make about incest if birth control is available is, in Greene's sense, an unfamiliar moral problem. For this unfamiliar moral problem, people still have the intuitive sense that incest is wrong (Haidt, 2001). Lott (2016) says that the mere fact that this judgment is caused by automatic mode is not a good reason for rejecting the judgment. We may still, upon reflection, decide that incest is morally wrong even in an unfamiliar context that is different from the one in which our intuitions were shaped (i.e. a context with birth control). Similarly, there could be valid a priori arguments for characteristically deontological responses to trolley arguments, even if the initial judgment is produced by automatic mode.

Greene does have the resources to respond to Lott's objection. Elsewhere, he has made an argument that has come to be called the argument from morally irrelevant factors (Paulo, 2019). Deontologically oriented philosophers have tended to say that pushing someone in front of a trolley to save others is impermissible because it involves using people as a mere means to an end (Foot, 1978). However, there are studies in which participants are given a dilemma where the decision is whether to pull a switch that causes someone to fall through a trapdoor and in front of a trolley (Greene et al., 2009). Crucially, they are still being used as a mere means (i.e. as a trolley stopper). In that case, it was discovered that most people give the "utilitarian" option in this case, just as they do in the case where the switch redirects the trolley to a different track. This indicates that the difference in responding between the footbridge case (pushing someone in front of a trolley) and the standard switch case (redirecting the trolley to hit another person) is not due to the fact that the footbridge case involves using someone as a trolley stopper. The difference in responding is actually due to the use of personal force; pushing someone feels more

wrong than causing their death via pulling a switch. Crucially, this is not something that either utilitarians or deontologists think is morally relevant. So, intuition-driven response to the footbridge case is not caused by anything either side deems morally relevant. The appeal to deontological prohibitions against using people as mere means seems to be a post-hoc rationalization.

Now, even if it is a post-hoc rationalization, we may still decide upon reflection to endorse it. But the fact that it is a post-hoc rationalization should cause us to be skeptical. Here is an analogy I used in my masters thesis (Simpson, 2021). Suppose psychologists experimentally demonstrated that the color of film titles had a large effect on how good people perceived the film to be. They did so by randomly assigning participants to two conditions where they watched the same movie, just with different colored titles. Participants in the "red title" condition reported that the movie was well done, whereas participants in the "blue title" condition reported that the movie was poorly done. Further suppose that participants in the "red title" condition, when asked, said things like "I enjoyed the movie because of its good character development[5]". This would be very compelling evidence that their assertions that they liked the movie because of the character development were just post-hoc rationalizations. Now, as Lott (2016) points out, the post-hoc rationalization might turn out (upon reflection) to be something we endorse. We might decide upon reflection that the character development really was good. But the fact that the color of the title caused our initial judgment that the character development was good should give us pause. It should cause us to reflect more on our judgment, and to not trust it. Similarly, we

[5] If the "title" example is not believable enough to be taken seriously, we could imagine the two conditions having different levels of attractiveness in the actors (but being the same in every other way), or some other variable that people would not upon reflection endorse as a reason to say the movie was well done.

should not trust the appeal to deontological prohibitions against using people as mere means, since those appear to be post-hoc rationalizations.

After arguing that the empirical evidence that utilitarian judgments are caused by manual mode provides reason to rely more on utilitarian judgments (at least for unfamiliar moral problems), Greene (2013) applies this conclusion to several modern cases. For example, he uses the question of how much to give to foreign life-saving charities as an example of an unfamiliar moral problem. Since we should (according to him) rely on utilitarianism for unfamiliar problems, he concludes we should give much more of our money to these life-saving charities. However, all of the empirical work purporting to find a connection between manual mode and utilitarian judgment involves other-sacrificial dilemmas. There was, prior to my work, no research looking for a correlation between reliance on manual mode and utilitarian responses to self-sacrificial dilemmas.

One of the most famous arguments to come out of the utilitarian tradition directly concerns self-sacrifice. Singer (1972), himself a utilitarian, asks you to imagine walking by a drowning girl in a pond. You only have seconds to act, and in order to save her, you will have to ruin your new, costly, clothes by wading into the pond to stop her from drowning. We feel a very strong intuition that sacrificing our possessions to save a life is obligatory in this case. However, we do not feel the same sense of obligation when we see an opportunity to donate money to life-saving charities. Utilitarianism plainly implies that we should perform the life-saving act in both cases. Greene (2013) argues that we endorse the utilitarian response to a self-sacrificial dilemma in the drowning girl case because of a visceral reaction. However, he argues if we relied more on manual mode, we would see that self-sacrifice is equally obligatory when it comes to donations to life-saving charities.

Conceptually, the above account is quite compelling, but it is ultimately a testable empirical hypothesis whether reliance on manual mode is associated with utilitarian responses to self-sacrificial dilemmas. My masters thesis included a study that was, to my knowledge, the first study to directly test the hypothesis. There are different ways of operationalizing manual mode (sometimes called "system 2"; Stanovich & West, 2000). The operationalization I chose to focus on was the cognitive reflection test (Frederick, 2005; Finucane & Gullion, 2010). I failed to find a correlation between cognitive reflection and utilitarian responses to self-sacrificial dilemmas (Simpson, 2021). However, I did not make use of the most sophisticated method (in the moral psychology literature) of extracting a psychological process, namely process dissociation (Conway & Gawronski, 2013). In what follows, I describe the work I did under the supervision of Kyle Nash. We extracted a self-sacrificial utilitarian parameter using process dissociation, and found that multiple lines of evidence indicate that it is associated with reliance on cognitive reflection.

Introduction

Self-sacrifice is an important feature of our ethical traditions. It is associated with concepts of heroism, righteousness, and even divinity. For example, in Western culture the most famous religious narrative (the crucifixion of Jesus) is about divine self-sacrifice. However, self-sacrifice has been largely ignored by contemporary moral psychology. There is a large body of evidence showing that there is a connection between cognitive reflection and utilitarian judgments in dilemmas that involve sacrificing *someone else* for the greater good (Patil et al., 2021). However, self-sacrifice is also an important aspect of utilitarianism. Peter Singer is the most influential living utilitarian philosopher, and the thought experiment for which he is most

famous (the drowning girl) is about the obligation to engage in self-sacrifice for the greater good (Singer, 1972). It is therefore worth investigating whether cognitive reflection is associated with utilitarian judgment in self-sacrificial dilemmas. To address this gap, we developed a process dissociation model to measure self-sacrificial utilitarianism, and examined if it is connected with cognitive reflection.

One of the most influential models in moral psychology is the dual-process model (Greene, 2008). As was explained in the background, this model posits that deontological responses to other-sacrificial dilemmas (i.e. judging that it is morally wrong to harm one to save many) are produced by quick and intuitive thinking. By contrast, utilitarian responses to other-sacrificial dilemmas (i.e. judging that it is morally right to harm one to save many) are, according to the model, produced by cognitive reflection. Greene (2015) argues that cognitive reflection (or "manual mode", as he calls it) will be more reliable for unfamiliar moral problems, which is his term for problems that we do not have evolutionary, cultural, or individual trial-and-error experience with. He argues that these three types of learning are the only way for our intuitions to become reliable, so in the absence of those three sources of reliability, we should rely on cognitive reflection. If Greene's argument is correct, then evidence that cognitive reflection tends to produce utilitarian responses is very significant. Evidence that cognitive reflection produces utilitarian judgments, in conjunction with the premise that cognitive reflection is more reliable for unfamiliar moral problems, leads to the conclusion that we should rely on utilitarian judgments for unfamiliar moral problems.

There are multiple lines of evidence for a connection between cognitive reflection and utilitarian responses to other-sacrificial dilemmas. Utilitarian responses to other-sacrificial dilemmas are positively correlated with scores on the cognitive reflection test, a measure of

one's tendency to reflect on one's intuitions and correct them (Paxton et al., 2012). Additionally, utilitarian responses to these dilemmas are positively correlated with activity in the dorsolateral prefrontal cortex (Greene et al., 2004), and with the thickness of the dorsolateral prefrontal cortex (Patil, 2021). This evidence is relevant because the dorsolateral prefrontal cortex (DLPFC) is a brain region associated with controlled cognition (Glenn et al., 2009; Yang et al., 2009). Disruption of cognitive reflection leads to fewer utilitarian responses, including through the disruption of the DLPFC (Zheng et al., 2018), and putting participants under cognitive load (Greene et al., 2008; Trémolière et al., 2012; Białek & De Neys, 2017). Giving participants more time to think about other-sacrificial dilemmas and instructing them to be deliberative both increase their likelihood of giving utilitarian responses (Suter & Hertwig, 2011). A recent meta-analysis which included 52 studies of responses to moral dilemmas found that promoting intuition (or inhibiting deliberation) tended to decrease utilitarian responses (Capraro, 2024). Another recent meta-analysis which included 53 studies found that increasing reliance on controlled cognition tended to increase utilitarian responses (Klenk, 2022)

One criticism of this work is that there are at least two possible explanations for why people endorse sacrificing one to save many: 1) people respond this way because they care about the number of people saved (i.e. they have utilitarian tendencies), 2) they do not care about harming the person (i.e. they have low harm aversion). A number of studies support the second explanation as being at least partly correct. For example, "utilitarian" responses to these dilemmas are positively associated with psychopathy (Bartels & Pizarro, 2011; Koenigs et al., 2012), and negatively associated with self-reported social connection, donating money, and feeling obligated to help people in poverty (Kahane et al., 2015). Given that utilitarians should believe in an obligation to help people in need (De Lazari-Radek & Singer, 2014), this counts

strongly against the construct validity of other-sacrificial dilemmas as a measure of utilitarianism. However, the dual-process model has the resources to explain these effects: people with reduced affective reactions (like psychopaths) should be less utilitarian because intuitive or emotional reactions ("automatic mode", to use Greene's term) have less of an effect on their moral judgments (Greene, 2015).

In order to further defend against the criticism that other-sacrificial dilemmas are not good measures of utilitarian tendencies, researchers used process dissociation to separate other-sacrificial utilitarian tendencies from deontological tendencies (Conway & Gawronski, 2013; Conway et al., 2018). The technique also separates utilitarianism from low harm aversion. In other words, it allows researchers to extract a measure of the extent to which people favour using harm to maximize the greater good that is not confounded by the extent to which people just do not feel averse to using harm in general. Participants are given 10 dilemmas in which harming another person maximizes the greater good (in other words, the standard dilemmas that prior research used), and 10 parallel dilemmas that are identical except for the fact that the same harmful action *does not* maximize the greater good. For example, one dilemma pair is as follows: in the incongruent dilemma (where utilitarianism and deontology lead to different answers), one must decide whether to torture someone to find lethal bombs, whereas in the parallel congruent dilemma (where utilitarianism and deontology lead to the same answer) one must decide whether to torture someone to find paint bombs. If participants approve of harm in the incongruent dilemmas (ones like the lethal bomb case), then this increases their other-sacrificial utilitarianism (OU) score[6]. But if they approve of harm in the congruent dilemmas (ones like the paint bomb

[6] Conway et al. (2018) call it a U parameter score, but for consistency's sake we will be calling it an OU parameter score (other-sacrificial utilitarianism) throughout this paper. This is because all of the dilemmas involve harming someone else, and we will be distinguishing this from self-sacrificial utilitarianism.

case), then this decreases their other-sacrificial utilitarianism (OU) parameter score. The OU parameter thus controls for the tendency to harm people even if it does not increase overall happiness. Figure 1 shows a representation of this process dissociation model (Conway & Gawronski, 2013). For a full list of the dilemmas, see the Appendix.

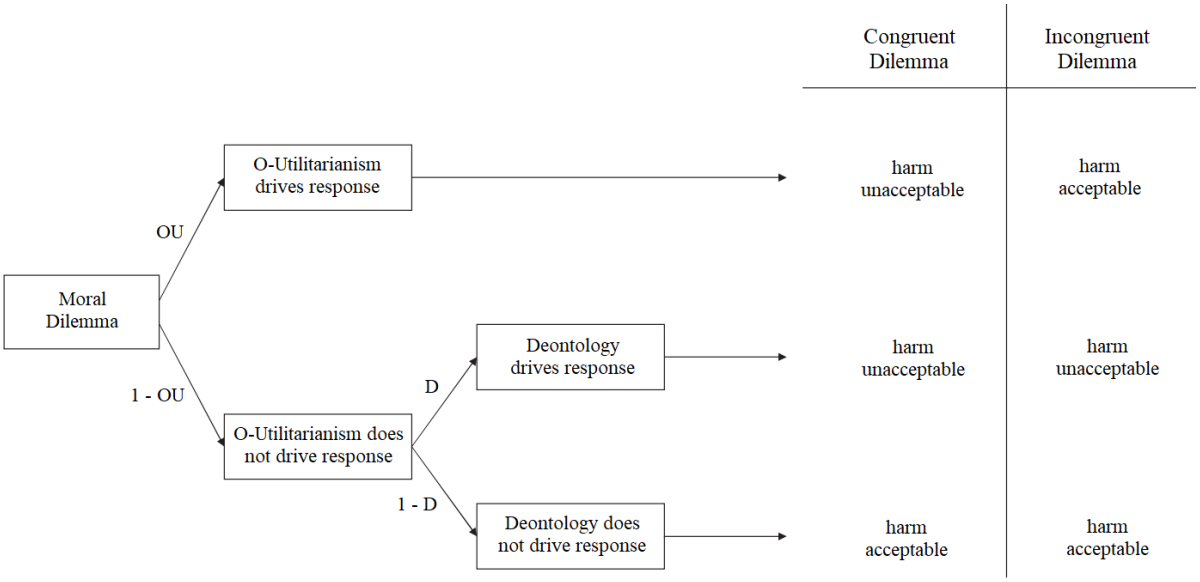

Figure 1. Processing tree modeling how Other-Sacrificial Utilitarianism (OU) and Deontology (D) lead to the judgments that harm (or other-sacrifice) is either acceptable or unacceptable in congruent and incongruent moral dilemmas.

The equations for other-sacrificial utilitarianism (OU, to distinguish it from self-sacrificial utilitarianism) and deontology (D) are as follows:

$$OU = p(\text{unacceptable} \mid \text{congruent}) - p(\text{unacceptable} \mid \text{incongruent})$$

$$D = p(\text{unacceptable} \mid \text{incongruent}) / (1 - OU)$$

To make these equations concrete, suppose that a participant says that harm is unacceptable in 9/10 congruent dilemmas, and says that harm is unacceptable in 3/10 incongruent dilemmas. Their OU parameter score would be calculated as follows: 0.9 – 0.3 = 0.6. The advantage of this parameter is that it controls for a mere lack of harm aversion. Crucially, the more a participant says that harm is acceptable in congruent dilemmas (i.e. the more willing they are to harm even if it does not minimize overall suffering), the *lower* their OU parameter score will be. This same participant's D parameter score would be calculated as follows: 0.3 / (1 - 0.6) = 0.75.

This other-sacrificial utilitarian (OU) parameter is positively correlated with cognitive reflection test scores (Byrd & Conway, 2019), reduced by cognitive load (Conway & Gawronski, 2013), and increased by an instruction to rely on analytical thinking (Li et al., 2018). So, there is still good evidence that it is associated with reliance on cognitive reflection. However, unlike utilitarian responses measured without process dissociation, the OU parameter is negatively correlated with psychopathy and positively correlated with caring about overall welfare, belief in the immorality of harm, and self-reported morality (Conway et al., 2018). Thus, other-sacrificial utilitarian (OU) parameter scores thus do seem to measure a genuine concern for the greater good, not a mere lack of harm aversion.

However, OU parameter scores do not correlate with charitable giving or with utilitarian responses to dilemmas that involve sacrificing one's own interests (or the interests of one's ingroup) for the greater good (Everett & Kahane, 2020). Thus, the OU parameter is still not an ideal utilitarian measure, because utilitarianism should include a willingness to engage in self sacrifice. Additionally, there is some initial evidence that intuition, not cognitive reflection, drives utilitarian self-sacrifice. For example, Conway et al. (2018) found that utilitarian

responses to self-sacrificial dilemmas positively correlated with the deontology (D) parameter. The D parameter is thought to be driven by empathy and emotion (Conway & Gawronski, 2013). There was also an unpublished study that found no correlation between utilitarian responses to self-sacrificial dilemmas and the cognitive reflection test (Simpson, 2021).

The above evidence tentatively suggests that there is no connection between cognitive reflection and utilitarian responses to self-sacrificial moral dilemmas. However, there is an ambiguity in the self-sacrificial dilemmas used in those studies: it is unclear whether people who endorse self-sacrifice are doing so because they have a pure altruistic tendency to sacrifice themselves to help others (whether or not it maximizes the greater good), or because they have a tendency to sacrifice themselves only when it maximizes the greater good. As with other-sacrificial dilemmas, the ambiguity between these two tendencies can be resolved using process dissociation. We will call these two tendencies altruism (A) and self-sacrificial utilitarianism (SU). We define altruism as the tendency to sacrifice oneself for others *whether or not* it maximizes the overall good, whereas we define self-sacrificial utilitarianism as the tendency to sacrifice oneself *only when* it maximizes the greater good. As is often the case in psychology, we are giving a definition of a term (in this case, "altruism") that is narrower than its ordinary language meaning. It is also different from Trivers' (1971) sense of the term which was discussed in the background section. Our definition of altruism is designed to distinguish it from self-sacrifice that is done only in cases where it maximizes the greater good. There are some cases of self-sacrificial behavior where the benefit to the other is less than the cost to the self. Someone who is altruistic (in our sense) will engage in self-sacrifice even in such cases where the cost to them is greater than the benefit it brings to someone else. By contrast, someone driven by self-sacrificial utilitarianism will only engage in self-sacrifice if the cost to them is less than

the benefit to the other person (or other people). Because both tendencies could be undergirding responses in favor of self-sacrifice in prior studies, the results of those studies are ambiguous.

We created two sets of 10 parallel dilemmas involving self-sacrifice. In these dilemmas, the question is not whether "harm" is acceptable, but whether "selfishness" is acceptable. In the model, "selfishness" is operationally defined as not endorsing self-sacrifice, it does not presuppose a selfish motivation. We could have just as easily used the term "non-self-sacrifice" instead of "selfishness". In the congruent dilemmas, altruism and self-sacrificial utilitarianism[7] both lead to the conclusion that selfishness is unacceptable. For example, one congruent dilemma involves the decision of whether to drive one's car off a cliff to one's certain death in order to avoid hitting five people. In this dilemma, utilitarians and altruists would both say that one should drive off the cliff. In the incongruent dilemmas, only altruism entails that selfishness is unacceptable. For example, the incongruent version of the cliff dilemma involves the decision of whether to drive one's car off a cliff to one's certain death in order to avoid hitting a single person more elderly than oneself. In this dilemma, only altruists would say that one should drive off the cliff. Figure 2 shows a representation of our process dissociation model[8].

[7] In these dilemmas (and the dilemmas from Conway & Gawronski, 2013), it is sometimes debatable what utilitarianism truly implies one morally should do. However, all that is required is that self-sacrifice be *more* justifiable in utilitarian terms in the congruent dilemmas than in the incongruent dilemmas.

[8] Conway and Gawronski (2013) point out that when making a process dissociation model, there is a choice for which process to make dominant. Following them, we made the utilitarian process the dominant one. Making altruism the dominant process leads to the absurd conclusion that people who are driven by neither process would endorse selfishness in the congruent dilemmas (i.e. you should *not* drive off the cliff to save five), but not in the incongruent dilemmas (i.e. you *should* drive off the cliff to save one elderly person).

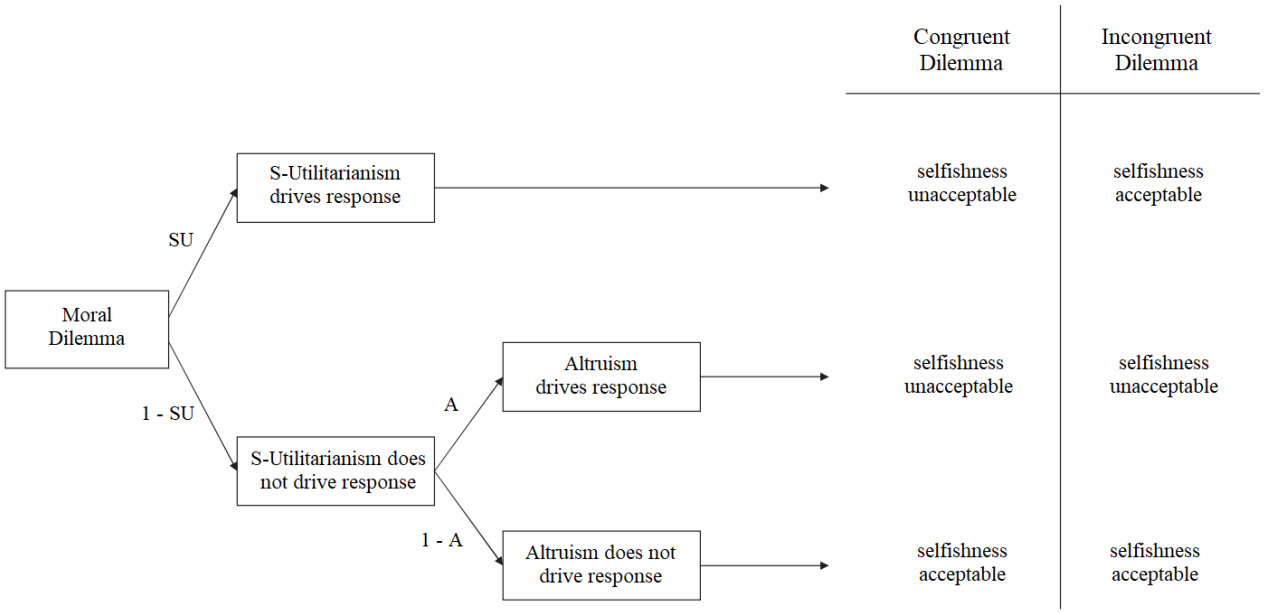

Figure 2. Processing tree modeling how Self-Sacrificial Utilitarianism (SU) and Altruism (A) lead to the judgments that selfishness is either acceptable or unacceptable in congruent and incongruent moral dilemmas.

The equations for the Self-Sacrificial Utilitarian (SU) and Altruism (A) parameters are as follows:

SU = p(unacceptable | congruent) - p(unacceptable | incongruent)

A = p(unacceptable | incongruent) / (1 - SU)

Again, we will make these equations concrete with an example. If a participant says that selfishness is unacceptable in 8/10 congruent dilemmas, and says that selfishness is unacceptable in 5/10 incongruent dilemmas. Their SU parameter score would be calculated as follows:

0.8 – 0.5 = 0.3. This same participant's A parameter score would be calculated as follows: 0.5 / (1 - 0.3) = 0.71.

These two process dissociation models were the basis for our studies. We had four moral measures to work with: other-sacrificial utilitarianism, deontology, self-sacrificial utilitarianism, and altruism. In Study 1, we measured whether there is a correlation between the cognitive reflection test and the four moral measures. In line with prior research, the cognitive reflection test positively correlated with other-sacrificial utilitarianism. Contrary to our initial expectations (but consistent with Greene's dual-process model), we found that cognitive reflection test scores also positively correlated with self-sacrificial utilitarianism. On this basis, for Studies 2-4 we derived our predictions from the hypothesis that cognitive reflection drives utilitarian judgment in self-sacrificial and other-sacrificial dilemmas. In Study 2, we replicated the results of Study 1, and added an experimental manipulation designed to increase reliance on cognitive reflection. In Study 3, we tested for a correlation between giving in the multi-game version of the public goods game (a behavior associated with reliance on cognitive reflection) and self-sacrificial utilitarianism. In Study 4, we did a high-powered replication of Study 2 and Study 3. In Study 5, we did an exploratory study of self-sacrificial utilitarianism and altruism that added a relational element. To be more specific, we created five versions of each dilemma; in each version the self-sacrifice is on behalf of someone with whom you have a different relationship.

Study 1

For a measure of reliance on cognitive reflection, we used the cognitive reflection test (CRT). Using composite measures of rationality and of heuristic-based reasoning as the criterion variables, Toplak et al. (2011) provided evidence that the CRT has predictive validity that is

independent of general cognitive ability. Additionally, giving cognitive load significantly impairs people's performance on the CRT, but not on comparably hard math problems without an intuition-reflection conflict (Johnson et al., 2014). This pattern indicates that it is a valid measure of one's tendency to override intuition, not merely a measure of cognitive ability. Prior studies have found that other-sacrificial utilitarianism correlates with the CRT[9] (Paxton et al., 2012; Byrd & Conway, 2019), but not with cognitive ability (Bostyn et al., 2020). Some have argued that the CRT and numeracy (mathematical ability as opposed to general cognitive ability) correlate so highly that they should be treated as the same construct (Erceg et al., 2020). However, Capraro et al. (2017) found that the CRT predicted efficient social decision making over and above numeracy, providing additional evidence for the incremental validity of the CRT. We initially predicted that the CRT would positively correlate with other-sacrificial utilitarianism, but not with self-sacrificial utilitarianism (see Simpson, 2021).

Critics of the use of other-sacrificial dilemmas have put forward a two-dimensional measure of utilitarianism which includes two scales, the impartial beneficence scale and the instrumental harm scale (Kahane et al., 2018). The former purports to measure impartial concern for the well-being of all. Notably, four out of the five items are about self-sacrifice on behalf of the greater good, so in terms of face validity it seems to be primarily about self-sacrifice. Additionally, the impartial beneficence scale, but not the instrumental harm scale, predicts kidney donation (Amormino et al., 2022), and therefore it also has predictive validity as a measure of self-sacrifice. However, it is subject to the same problem as other measures of self-sacrifice that are not based on process dissociation: it is unclear whether the endorsement of self-

[9] The standard cognitive reflection test includes only mathematical items. There are measures of verbal or logical reflection, but they do not uniquely correlate with utilitarian responses (as opposed to deontological responses), only mathematical reflection does (Byrd & Conway, 2019). This makes sense given that utilitarianism judgment is, at least in theory, about quantifying the amount of happiness and suffering that is produced by an action.

sacrifice is driven by utilitarianism or altruism. We included these scales to see how they would relate to the parameters we will extract using process dissociation.

Method

Participants. We recruited 198 participants[10]. They were recruited from the Psychology participant pool and received partial course credit for completing the study. After applying our pre-registered exclusions (https://osf.io/g6kvs), there were 168 participants left (123 females, 43 males, 2 other). There was a narrow age range (17-40), and the average age was 19.4. The ethnicity of the participants was as follows: 48.2% European (n = 81), 11.3% South Asian (n = 19), 9.5% East Asian (n = 16), 6.0% Black (n = 10), 4.8% Middle Eastern and North African (n = 8), 2.4% South East Asian (n = 4), 1.8% Latin American (n = 3), 0.6% Indigenous (n = 1), 7.7% mixed or other (n = 13), and 7.7% did not answer (n = 13). The relevant questions were all forced response, and anyone who did not complete the survey was excluded, in accordance with the pre-registered exclusions. We thus did not have to deal with missing data. We followed this practice for Studies 2-5 as well.

Cognitive reflection test. We used an updated version of the CRT (Finucane & Gullion, 2010) to avoid the problem of people being familiar with the questions. This version had three questions. One example is as follows: *If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200*

[10] In our pre-registration we said we would get a sample of 300. Due to oversight, we asked the research pool coordinator for 200 participants. In Study 2 and Study 4, we replicated the effects of Study 1 with larger sample sizes. All datasets and analysis syntax files are available here: https://zenodo.org/records/10064032. The raw data and data file with exclusions are both posted. Since many of the exclusions had to be done manually, there is not a syntax file with exclusions. However, they were all done in accordance with our preregistrations.

patients? In this question, the intuitive answer is “200 minutes”, but the correct answer (which requires reflection) is 2 minutes. This scale is used to measure reliance on system 2, or what Greene (2015) calls “manual mode”. For all the questions, see the Appendix. In this study (and in all subsequent studies with the CRT), the CRT was given before the dilemmas. This scale had relatively low reliability (Cronbach’s $\alpha = 0.464$).

Moral dilemmas. The moral dilemmas for deriving the Other-Sacrificial Utilitarian (OU) and Deontology (D) parameters were taken directly from Conway and Gawronski (2013). The OU parameter measures willingness to harm others for the greater good, whereas the D parameter measures unwillingness to harm others whether or not it maximizes the greater good. As was explained in the introduction, the equations for deriving these parameters are as follows, where “unacceptable” means harm is unacceptable:

$OU = p(\text{unacceptable} \mid \text{congruent}) - p(\text{unacceptable} \mid \text{incongruent})$

$D = p(\text{unacceptable} \mid \text{incongruent}) / (1 - OU)$

The moral dilemmas for deriving the Self-Sacrificial Utilitarian (SU) and Altruism (A) parameters were internally generated (with two exceptions, see the Appendix). The SU parameter measures willingness to sacrifice oneself for the greater good, and the A parameter measures willingness to sacrifice oneself whether or not it maximizes the greater good. As was explained in the introduction, the equations for deriving these parameters are as follows, where “unacceptable” means non-self-sacrifice (or selfishness) is unacceptable:

$SU = p(\text{unacceptable} \mid \text{congruent}) - p(\text{unacceptable} \mid \text{incongruent})$

$A = p(\text{unacceptable} \mid \text{incongruent}) / (1 - SU)$

For a complete list of the dilemmas, see the Appendix. Participants had to read 20 pairs of parallel dilemmas, or 40 dilemmas total. All 40 dilemmas were given in a fixed random order. We used https://www.random.org/lists/ to generate the order. We made the following manual adjustments to the order: we made sure that there was an item with a comprehension check question at the beginning, and we moved a dilemma any time there were two parallel dilemmas side by side (e.g. the vaccine congruent dilemma right after the vaccine incongruent dilemma). We did not use a different random order for each participant because then two parallel dilemmas would sometimes be given side by side. Participants were told there would be comprehension checks after some of the dilemmas to make sure they read them. We included five multiple choice comprehension checks, all of which asked, "What was the previous dilemma about?". We used the same 40 dilemmas and the same comprehension checks for Studies 2-4 as well.

For skewness, the OU, A, and D parameters all had scores in between -0.5 and 0.5, which is consistent with univariate normality. The SU parameter was slightly negatively skewed, with a score of -0.511. For kurtosis, the SU and D parameters had scores in between -0.5 and 0.5, which is consistent with univariate normality. The OU and A parameters had respective scores of -0.596 and -0.936.

Oxford utilitarianism scale. Kahane et al. (2018) applied factor analysis to a large number of utilitarian questions and extracted two factors, from which they made two scales: the impartial beneficence scale (five questions) and the instrumental harm scale (four questions). The former measures a tendency to endorse care for the greater good (four out of the five items are about self-sacrifice), and the latter measures a willingness to harm others for the greater good. Together these constitute the Oxford utilitarianism scale. This scale has typically been used by critics of the use of other-sacrificial dilemmas (Everett & Kahane, 2020).

The most commonly used lower limit for Cronbach alpha is 0.70 (Hair et al., 2010). Based on this conventional (albeit arbitrary) standard, reliability was minimally acceptable for the impartial beneficence scale (Cronbach's $\alpha = 0.707$), but not quite minimally acceptable for the instrumental harm scale (Cronbach's $\alpha = 0.653$).

Threat manipulation. For exploratory purposes, we included a threat manipulation. Half the participants were randomly assigned to read a graduate-level statistics passage that was designed to be incomprehensible, while the other half who were assigned to the control condition read a statistics passage that was very easy to read. The manipulation is designed to threaten people's sense of their own competence (McGregor et al., 2005). Threat manipulations of this kind have been shown to increase certainty and extremeness in people's identity-related beliefs, such as political or religious beliefs (McGregor & Jordan, 2007; McGregor et al., 2010). If people's moral beliefs are comparably important to their identity, it is plausible to expect threat to cause higher scores (more extremeness) in at least some of the four moral parameter scores.

Data analytical plan. In accordance with prior studies from this area in moral psychology (Paxton et al., 2012; Conway et al., 2018), we planned to calculate a Pearson's correlation coefficient between the cognitive reflection test and our four moral parameters: self-sacrificial utilitarianism (SU), other-sacrificial utilitarianism (OU), altruism (A), and deontology (D). For our threat manipulation, we planned to use four independent samples t-tests. We also planned to construct linear regression models of the impartial beneficence and instrumental harm scales. For the impartial beneficence scale, the SU parameter and A parameter were the predictors we chose. For the instrumental harm scale, the OU parameter and D parameter were the predictors we chose. Alpha was set at .05, and all of the tests were two-tailed.

In accordance with our preregistration (https://osf.io/g6kvs), the exclusion criteria were as follows: 1) anyone who took longer than 3 hours will be excluded, 2) five dilemmas were followed by a multiple choice comprehension question, and getting more than two wrong resulted in exclusion, 3) there are five questions at the end of the survey about how closely they paid attention, each of which resulted in exclusion if the answer indicated they were not paying attention. All of the attention questions are available in the above preregistration link.

Results

Descriptive statistics. The cognitive reflection test consisted of three questions. 18.5% of participants got none of the questions correct, 30.4% got one correct, 31.0% got two correct, and 20.2% got all three questions correct. The highest possible score for each of the moral parameters is 1. For the self-sacrificial utilitarian (SU) parameter, the average score was M = .461 (SD = .207). For the other-sacrificial utilitarian (OU) parameter, the average score was M = .357 (SD = .190). The average score for the altruism (A) parameter was M = .465 (SD = .291). The average score for the deontology (D) parameter was M = .566 (SD = .193).

Correlations. The cognitive reflection test (CRT) correlated positively with self-sacrificial utilitarianism (SU), r = .212, p = .006, and other-sacrificial utilitarianism (OU), r = .231, p = .003. It did not correlate with altruism (A), r = .005, p = .949, or deontology (D), r = -.018, p = .815. Additionally, the SU parameter and OU parameter were positively correlated with each other, r = .226, p = .003. The A parameter and D parameter also positively correlated with each other, r = .219, p = .004. Crucially, neither the SU parameter nor the OU parameter correlated with either the A parameter or the D parameter (all ps > .210). Because men scored

higher on the CRT than women[11], and were higher on the OU parameter[12], we also performed partial correlations between the CRT, SU, and OU while controlling for gender, and the correlations all remained significant (all *p*s < .010).

The impartial beneficence scale correlated positively with the self-sacrificial utilitarian (SU) parameter, $r = .290$, $p < .001$, and the altruism (A) parameter, $r = .336$, $p < .001$. It did not correlate with any of the other parameters (all *p*s > .293). The instrumental harm scale correlated positively with both the self-sacrificial utilitarian (SU) parameter, $r = .205$, $p = .008$, and the other-sacrificial utilitarian (OU) parameter, $r = .228$, $p = .003$. It was negatively correlated with the deontology (D) parameter, $r = -.342$, $p < .001$, and did not correlate with the A parameter ($p = .557$).

Table 1

Correlations Among Variables, Study 1

Variable	1	2	3	4	5	6
1. CRT	—					
2. SU Parameter	.212**	—				
3. OU Parameter	.231**	.226**	—			
4. A Parameter	.005	.097	-.004	—		
5. D Parameter	-.018	-.040	-.022	.219**	—	
6. Impartial Beneficence	-.051	.290***	.082	.336***	.027	—
7. Instrumental Harm	.108	.205**	.228**	-.046	-.342***	.188*

* $p < .05$, ** $p < .01$, *** $p < .001$

[11] Note: this is only true of the mathematical CRT. For the verbal CRT, there is no sex difference (Sirota et al., 2021). We used the mathematical CRT because it uniquely predicts utilitarian responses (Byrd & Conway, 2019).
[12] The male advantage on the CRT was significant ($p < .001$), but the sex difference for the OU parameter was not ($p = .108$). However, a meta-analysis that used 6,100 participants found a significant sex difference wherein males had higher OU parameter scores (Friesdorf et al., 2015).

As was mentioned, four out of five items in the impartial beneficence scale are about self-sacrifice. To see whether responses to this scale are driven more by altruism or by self-sacrificial utilitarianism, we constructed a linear regression model to predict impartial beneficence scale scores using self-sacrificial utilitarianism (SU) and altruism (A) as the predictors. Overall, the model was significant, $F(2, 165) = 18.126$, $p < .001$, $R^2 = .180$. The SU parameter significantly predicted impartial beneficence scores ($\beta = .260$, $p < .001$), as did the A parameter ($\beta = .311$, $p < .001$). We also constructed a linear regression model to predict instrumental harm scale scores using the other-sacrificial utilitarianism (OU) and deontology (D) as the predictors. Overall, the model was significant, $F(2, 165) = 16.336$, $p < .001$, $R^2 = .165$. The OU parameter significantly predicted instrumental harm scores ($\beta = .220$, $p = .002$), as did the D parameter ($\beta = -.337$, $p < .001$).

Threat manipulation. We used four independent samples t-tests to measure the effect of the threat manipulation on the four moral parameters. We did not find any significant effects (all ps > .645).

Discussion

Replicating prior research (Byrd & Conway, 2019), other-sacrificial utilitarianism was positively correlated with cognitive reflection test scores. Crucially, self-sacrificial utilitarianism was also positively correlated with cognitive reflection test scores[13]. This supports the idea that cognitive reflection is associated with self-sacrificial utilitarianism, at least when it is measured

[13] On the basis of Simpson (2021), we predicted in our pre-registration for Study 1 that there would be no significant correlation between CRT scores and self-sacrificial utilitarianism. Based on our findings in Study 1, our pre-registrations for Study 2 and Study 4 predicted a positive correlation between CRT scores and self-sacrificial utilitarianism.

via process dissociation. This result is consistent with Greene's (2013) claim that cognitive reflection causes utilitarian judgments. The result was also counter to our initial predictions, because we were expecting self-sacrificial utilitarianism to *not* be associated with cognitive reflection. However, since this study uses a better measurement technique than prior studies (e.g. Simpson, 2021), for Studies 2-4, we made predictions consistent with the hypothesis that self-sacrificial utilitarianism is associated with reliance on cognitive reflection.

Inducing a sense of threat (via making people read a difficult statistics passage) did not affect any of our moral measures. In prior studies, threat inductions increased certainty in beliefs that were important to personal identity, since these serve protective functions. For example, threat inductions increased zeal about various political topics (McGregor & Jordan, 2007) and religious beliefs (McGregor et al., 2010). It could be that lay people do not explicitly think of themselves as "utilitarian" or "deontological", whereas they do explicitly think of themselves as "liberal" or "Christian". If this were true, this would explain the fact that threat does not cause increased extremeness for moral beliefs, but it does for political and religious beliefs. Future studies could measure the extent to which people think of their identity in terms of their moral commitments.

Neither the deontology (D) nor the altruism (A) parameter were correlated with cognitive reflection. This result is consistent with prior studies finding that deontological responses are driven by quick intuitive thinking, not controlled cognition. For example, deontological responses are increased by inducing empathy (Conway & Gawronski, 2013), and are positively correlated with activity in the amygdala (Shenhav & Greene, 2014). Similarly, there is a large body of work indicating that emotional empathy causes altruism (Batson et al., 1989; Persson & Kajonius, 2016), albeit with a different operationalization of altruism than the one we used.

We included the Oxford utilitarianism scale, which breaks down into the instrumental harm scale and the impartial beneficence scale (Kahane et al., 2018). It has been used by critics of the use of trolley-like dilemmas as a measure of utilitarian tendencies (Everett & Kahane, 2020). However, the instrumental harm scale strongly correlated with higher other-sacrificial utilitarian (OU) parameter scores and lower deontology (D) parameter scores. Since OU and D parameter scores are derived from other-sacrificial dilemmas, these correlations could be a basis for convergence between critics and defenders of the use of those dilemmas as operationalizations of utilitarian tendencies. With regard to the impartial beneficence scale, it is an ambiguous measure because it could either be driven by a) a utilitarian tendency, or b) a purely altruistic tendency to self-sacrifice regardless of whether it maximizes the greater good. If high scores on the impartial beneficence scale are primarily driven by altruism (in our sense), that means that people who score highly would endorse self-sacrifice even if the cost to them was greater than the benefit conferred on the people being helped. By contrast, if high scores are driven by utilitarian tendencies, high scorers will only endorse self-sacrifice if there is more overall welfare produced. Our regression model suggests that both self-sacrificial utilitarian tendencies and altruistic tendencies account for a substantial amount of variance in scores on the impartial beneficence scale. This indicates that the impartial beneficence scale is not only measuring utilitarian tendencies, it is also measuring altruism (in our sense of the term).

Study 2

Study 1 showed that the cognitive reflection test is positively correlated with self-sacrificial utilitarian (SU) and other-sacrificial utilitarian (OU) tendencies. This provides evidence that reliance on cognitive reflection predicts utilitarian responses to moral dilemmas.

To test for a causal connection, we manipulated the use of cognitive reflection. In prior studies, instructing participants to rely on an analytic thinking style increased their OU parameter scores (Li et al., 2018; see also Capraro et al., 2019). We used a similar manipulation for this study. In one condition, participants were instructed to rely on reason when reading the dilemmas. In the other condition, they were instructed to rely on emotion. We pre-registered the following hypothesis (https://osf.io/c9x5j): reason drives utilitarian judgments, both other-sacrificial and self-sacrificial. We derived four predictions from this hypothesis[14]: 1) the cognitive reflection test would be positively correlated with other-sacrificial utilitarianism, 2) the cognitive reflection test would be positively correlated with self-sacrificial utilitarianism, 3) the reason prime would increase other-sacrificial utilitarianism relative to the emotion prime, and 4) the reason prime would increase self-sacrificial utilitarianism relative to the emotion prime.

Method

Participants. We recruited 371 participants[15] using Amazon mechanical turk (MTurk). They were paid $1.50 US for completing the survey. After applying our pre-registered exclusions (same as Study 1), there were 268 participants left (152 females, 115 males, 1 other). This sample size gave us more than 95% power to detect a correlation of $r = .212$ or greater (the smallest correlation that we were trying to replicate), with a false positive rate of 5% and a one-tailed test. Since this was not a student sample, there was a wide age range (18-96). The average age was 42.7. The ethnicity of the participants was as follows: 77.2% European (n = 207), 7.1%

[14] I made two other hypotheses and derived four predictions from them. I also made four miscellaneous predictions, for a total of eight additional predictions. Five out of these eight additional predictions were confirmed. The results are included in the Supplemental Analyses.
[15] In our pre-registration we said we would get a sample of 300.

Black (n = 19), 6.0% Latin American (n = 16), 4.1% East Asian (n = 11), 0.4% South East Asian (n = 1), 1.9% mixed or other (n = 5), and 3.4% did not answer (n = 9).

Cognitive reflection test. We used the same cognitive reflection test that was used in Study 1. This time, the reliability of the CRT was higher, but it still did not reach minimal acceptability (Cronbach's $\alpha = 0.665$).

Moral dilemmas. We used the same moral dilemmas that were used in Study 1. The only difference is that the moral dilemmas were given in a different fixed random order. For skewness, the SU, OU, A, and D parameters all had scores in between -0.5 and 0.5, which is consistent with univariate normality. For kurtosis, the SU and OU parameters had scores in between -0.5 and 0.5, which is consistent with univariate normality. The A and D parameters had respective scores of -0.763 and -0.529.

Reason/emotion manipulation. We randomly assigned participants to two conditions (inspired by Capraro et al., 2019). In the reason condition, participants were given instructions which included the following sentence: "*Please deal with the following dilemmas by relying on reason, rather than emotion.*" In the emotion condition, participants were given instructions which included the following sentence: "*Please deal with the following dilemmas by relying on emotion, rather than reason.*" This was given immediately before the moral dilemmas. Instructing the first group to rely on reason is a way of getting them to rely on slow, reflective thinking instead of automatic thinking. For the full instructions, see the Appendix.

Data analytical plan. To replicate the results of Study 1, we planned to calculate a Pearson's correlation coefficient between the cognitive reflection test and our four moral parameters: self-sacrificial utilitarianism (SU), other-sacrificial utilitarianism (OU), altruism (A),

and deontology (D). For our manipulation of whether to instruct participants to rely on reason or emotion, we planned to use independent samples t-tests with the condition as the independent variable, and the moral measures as the dependent variables. Alpha was set at .05, and all of our preregistered directional predictions were tested with one-tailed tests. All other tests were two-tailed.

In accordance with our preregistration (https://osf.io/c9x5j) the exclusion criteria were as follows: 1) anyone who took longer than 3 hours was excluded, 2) five dilemmas were followed by a multiple choice comprehension question, and getting more than two wrong resulted in exclusion, 3) there are five questions at the end of the survey about how closely they paid attention, each of which resulted in exclusion if the answer indicated they were not paying attention. All of the attention questions are available in the above preregistration link.

Results

Descriptive statistics. On the cognitive reflection test, 20.5% of participants got none of the questions correct, 18.7% got one correct, 25.4% got two correct, and 35.4% got all questions correct. For the self-sacrificial utilitarian (SU) parameter, the average score was $M = .268$ ($SD = .223$). For the other-sacrificial utilitarian (OU) parameter, the average score was $M = .292$ ($SD = .198$). The average score for the altruism (A) parameter was $M = .353$ ($SD = .254$). The average score for the deontology (D) parameter was $M = .571$ ($SD = .228$).

Pre-registered prediction 1. We predicted that the cognitive reflection test (CRT) would be positively correlated with the other-sacrificial utilitarianism (OU). The CRT correlated positively with the OU parameter, $r = .198$, $p = .001$. This prediction was confirmed.

Table 2

Correlations Among Variables, Study 2

Variable	1	2	3	4
1. CRT	—			
2. SU Parameter	.163**	—		
3. OU Parameter	.198***	.360***	—	
4. A Parameter	-.120*	.099	-.201***	—
5. D Parameter	.024	-.011	.027	.229***

* $p < .05$, ** $p < .01$, *** $p < .001$

Pre-registered prediction 2. We predicted that the cognitive reflection test (CRT) would be positively correlated with the self-sacrificial utilitarianism (SU). The CRT correlated positively with the SU parameter, $r = .163$, $p = .007$. This prediction was confirmed.

Pre-registered prediction 3. We predicted that the reason prime would increase the other-sacrificial utilitarian (OU) parameter relative to the emotion prime. Those in the reason condition had higher OU parameter scores ($M = .308$, $SD = .195$) than those in the emotion condition ($M = .278$, $SD = .200$), but this difference was not significant: $t(266) = 1.237$, $p = .109$, $d = .151$. The effect was in the right direction, but it was not significant, so the prediction was not confirmed.

Pre-registered prediction 4. We predicted that the reason prime would increase the self-sacrificial utilitarian parameter (SU) relative to the emotion prime. Those in the reason condition had higher SU parameter scores ($M = .287$, $SD = .231$) than those in the emotion condition ($M = .251$, $SD = .216$), but this difference was also not significant: $t(266) = 1.294$, $p = .098$, $d = .158$.

The effect was in the right direction, but it was not significant, so the prediction was not confirmed.

As in Study 1, we controlled for gender, and the positive correlations between the CRT, SU parameter and OU parameter remained significant (all *p*s < .004).

Discussion

Our predictions for Study 2 were based on Greene's (2015) hypothesis that cognitive reflection drives utilitarian judgment. Our pre-registered predictions about the positive correlation between the cognitive reflection test and utilitarian responses (to other-sacrificial and self-sacrificial dilemmas) was confirmed. We added an experimental manipulation wherein one group was instructed to rely on reason, the other on emotion. The instruction to rely on reason increased utilitarian responses to both types of dilemmas, but the effect was not significant. If the instruction to rely on reason had *significantly* increased utilitarian judgment, that would have provided convincing evidence of a causal connection between cognitive reflection and utilitarian judgment. In the absence of a causal connection, what else could explain the positive correlation between the two variables? One possibility is that a third variable causes both increases cognitive reflection and increases utilitarian judgment. There are many possible third variables, but one is autistic traits. Autism is associated with increased utilitarian judgment (Gleichgerrcht et al., 2013), and increased cognitive reflection (Brosnan & Ashwin, 2023). It could be that this explains why we detected a correlation between cognitive reflection and utilitarian judgment, but did not detect a causal connection.

However, we suspected that the reason we failed to detect a causal connection between an instruction to rely on reason and utilitarian judgment was that this study was underpowered (the p-values for the effect of the manipulation were both around .1, and the sample size was 268). We performed the same manipulation with a larger sample in Study 4. For Study 3, we sought additional evidence of a connection between cognitive reflection and utilitarian judgment using an economic game that has been found to be a proxy for reliance on reflection, namely, the public goods game (Rand, 2016).

Study 3

In a meta-analysis of experiments involving economic games, Rand (2016) found that pure cooperation (giving to others when there is no strategic benefit) is associated with reliance on intuition, but strategic cooperation (giving to others when it is economically rational to do so) is associated with reliance on deliberation. One example is the public goods game, where all participants get money (or tokens) and have the option to give to a common pool. The common pool is then multiplied by a constant and then divided evenly between the participants. There are two versions of the public goods game. In the single-game version, it is self-interested to not put anything in the pool, and donating is pure cooperation. In the multi-game version, the game is iterated and occurs many times. In this version, donating is construed as strategically rational by Rand (2016). This is because if you do not give, others will stop giving as well, and you will not benefit from the common pool.

Participants in this study were randomly assigned to the single-game version or the multi-game version. Since we found evidence that cognitive reflection predicts the self-sacrificial

utilitarian (SU) parameter and the other-sacrificial utilitarian (OU) parameter, and there is evidence that cognitive reflection is associated with strategic cooperation, we predicted that the SU parameter and OU parameter would each be positively correlated with strategic cooperation (i.e. giving in the multi-game version). By contrast, we predicted that the altruism (A) parameter and deontology (D) parameter would be positively correlated with pure cooperation (i.e. giving in the single-game version). These predictions were not confirmed. In fact, pure cooperation was marginally positively correlated with the SU parameter, and was significantly positively correlated with the OU parameter.

Method

Participants. We recruited 278 participants[16] from a social psychology course. After applying our pre-registered exclusions (https://osf.io/8utrk), there were 221 participants left (153 females, 65 males, 3 other). The age range was 18-46, and the average age was 20.9. The ethnicity of the participants was as follows: 29.9% European (n = 66), 20.4% East Asian (n = 45), 16.3% South Asian (n = 36), 7.7% South East Asian (n = 17), 4.5% Black (n = 10), 1.8% Latin American (n = 4), 1.4% Indigenous (n = 3), 5.4% mixed or other (n =12), and 2.7% did not answer (n = 6).

Public goods game. Participants were put into groups of four. Each member of the group was given 100 tokens that would be used in a raffle for a prize of 20 dollars. There was a common pool that participants could give tokens to. The amount in the common pool would be multiplied by two and then evenly divided between the four participants. Half of the participants

[16] In our pre-registration we said we would get a sample of 260.

were assigned to the single-game condition, and half were assigned to the multi-game version. In the single-game version, they only do one round of giving towards the common pool. In the multi-game version, they play the game 10 times in a row, and they know up front how many games will be played[17].

Moral dilemmas. The moral dilemmas were presented after the public goods game[18]. They were the same dilemmas that were used in Study 1 and Study 2, given in a different fixed random order. For skewness, the SU, OU, A, and D parameters all had scores in between -0.5 and 0.5, which is consistent with univariate normality. For kurtosis, the SU, OU, and D parameters had scores in between -0.5 and 0.5, which is consistent with univariate normality. The A parameter had a score of -0.841.

Data analytical plan. We planned to calculate a Pearson's correlation coefficient between the amount given in the public goods game and our four moral parameters: self-sacrificial utilitarianism (SU), other-sacrificial utilitarianism (OU), altruism (A), and deontology (D). In the single-game condition, we planned to use the total amount given. In the multi-game condition, we planned to use the average amount given. Alpha was be set at .05, and all of our preregistered directional predictions were tested with one-tailed tests. All other tests were two-tailed.

---

[17] If there are not enough people online at the right time to be assigned to a group of 4, 1-3 bots are assigned to respond. In the multi-game version, this is relevant because they could affect how much participants give. In the multi-game version, 16 participants were put in a group with at least one bot, 95 were put in a group with no bots. Participants in bot groups gave somewhat more (M = 49.144, SD = 17.601) than people in botless groups (M = 41.575, SD = 24.655), but this difference was not significant: $t(266) = 1.176$, $p = .242$, $d = .318$. In keeping with our preregistration, given that the difference was not significant, we did not exclude participants in bot groups from our analysis.

[18] This study was done in collaboration with another researcher interested in how social trust correlated with behavior in economic games. The other researcher added a set of questions on social trust and a short attachment scale. These items were not included in our pre-registration as they were not relevant to our hypotheses, so we did not include them here. The items are available upon request.

In accordance with our preregistration (https://osf.io/8utrk) the exclusion criteria were as follows: 1) if a group did not have four participants, a bot took their place; if participants playing with bots gave significantly more or less than participants without bots in the multi-game version, they were excluded, 2) anyone who took longer than 3 hours will be excluded, 3) five dilemmas were followed by a multiple choice comprehension question, and getting more than two wrong resulted in exclusion, 4) there were five questions at the end of the survey about how closely they paid attention, each of which resulted in exclusion if the answer indicates they were not paying attention. All of the attention questions are available in the above preregistration link.

Results

Descriptive statistics. In the single-game condition, the descriptives were as follows. The average amount given was $M = 50.100$ ($SD = 33.540$). For the self-sacrificial utilitarian (SU) parameter, the average score was $M = .401$ ($SD = .195$). For the other-sacrificial utilitarian (OU) parameter, the average score was $M = .342$ ($SD = .163$). The average score for the altruism (A) parameter was $M = .461$ ($SD = .256$). The average score for the deontology (D) parameter was $M = .589$ ($SD = .191$). In the multi-game condition, the descriptives were as follows. The average amount given (across the 10 games) was $M = 42.666$ ($SD = 23.850$). For the SU parameter, the average score was $M = .360$ ($SD = .193$). For the OU parameter, the average score was $M = .306$ ($SD = .178$). The average score for the A parameter was $M = .542$ ($SD = .273$). The average score for the D parameter was $M = .581$ ($SD = .189$).

Pre-registered Hypothesis 1. For participants in the multi-game condition, the total amount given did not correlate with any of the four parameters (all ps $> .534$). We were

predicting that both of the utilitarian parameters (SU and OU) would positively correlate with giving in this condition, so this prediction was not confirmed.

Pre-registered Hypothesis 2. The amount given in the single-game condition was not correlated with the altruism (A) parameter or the deontology (D) parameter (both *p*s > .283). We were predicting that these parameters would positively correlate with giving in this condition, so this prediction was not confirmed. However, for participants in the single-game condition, there was a marginal positive correlation between the total amount given and the self-sacrificial utilitarianism (SU), $r = .163$, $p = .098$, and a significant correlation between the total amount given and other-sacrificial utilitarianism (OU), $r = .293$, $p = .002$.

Table 3

Correlations Among Variables, Study 3 (Single Game Condition)

Variable	1	2	3	4
1. Donation	—			
2. SU Parameter	.163†	—		
3. OU Parameter	.293**	.324***	—	
4. A Parameter	.080	.240*	.055	—
5. D Parameter	.103	.072	.240*	.141

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Discussion

This study did not provide further evidence of an association between cognitive reflection and utilitarian judgment. Giving in the multi-game version is associated with reliance on reflection (Rand, 2016), and it did not correlate with self-sacrificial-utilitarianism or other-sacrificial utilitarianism. However, giving in the single-game condition marginally positively correlated with self-sacrificial utilitarianism, and significantly positively correlated with other-

sacrificial utilitarianism. This is surprising because giving in the single-game condition is associated with reliance on intuition (Rand, 2016). However, one additional difference between multi-game giving and single-game giving is that giving in the single-game is more genuinely self-sacrificial and prosocial (i.e. it is not the self-interested move). This could be what is driving its association with utilitarian judgment. The correlation we found could be thought of as behavioral validation of the utilitarian measures, since this correlation shows that utilitarian responses are associated with prosocial behavior. That is, this correlation could be taken to show that the utilitarian measures are not merely measures of psychopathy or coldness.

This is important because previous researchers have criticized the use of other-sacrificial dilemmas as measures of utilitarian tendencies. Several studies uncovered evidence that so-called "utilitarian" answers to other-sacrificial dilemmas (i.e. endorsing the use of harm to save many) are positively correlated with unsavoury traits, like Machiavellianism and psychopathy (Bartels & Pizarro, 2011; Kahane et al., 2015). The critics conclude that the endorsement of harm to save many (as an operational measure) does not actually reflect utilitarian tendencies, but a lack of harm aversion. This is one reason that defenders of the use of other-sacrificial dilemmas used process dissociation to extract the other-sacrificial utilitarian (OU) parameter. Conway et al. (2018) found that the OU parameter *negatively* correlated with psychopathy, and positively correlated with various moral traits. The fact that our study found that the OU parameter and the SU parameter both positively correlated with giving in an economic game is further evidence that these measures have construct validity as measures of moral tendencies.

The correlation between utilitarian judgment and giving in the single-game version of the public goods game is also significant because it constitutes behavioral validation, which previous studies have failed to find. In one prior study, utilitarian responses to hypothetical trolley-like

scenarios did not significantly predict utilitarian behavior in a real life situation (Bostyn et al., 2018). Giving in the public goods game (in either version) could be construed as a kind of utilitarian behavior, because giving to the common pool does contribute to the greater good (the amount is multiplied by a constant and spread evenly across the players). However, only the other-sacrificial (OU) parameter significantly positively correlated with giving in the single-game version of the public goods game. The self-sacrificial (SU) parameter also positively correlated with giving in that version, but the correlation was not significant. We suspected that this was due to insufficient power, so we sought to replicate the correlation with a larger sample in Study 4.

Study 4

In Study 2, the effect of the reason/emotion manipulation on self-sacrificial and other-sacrificial utilitarianism was not significant. The correlation between giving in the single-game version of the public goods game and self-sacrificial utilitarianism was also not significant. We aimed to replicate those findings with pre-registered predictions and a larger sample size. Because none of the correlations were even close to being significant in the case of the multi-game version of the public goods game, we did not include it in this replication.

We made six predictions: 1) the reason prime will increase the self-sacrificial utilitarian (SU) parameter relative to the emotion prime, 2) the reason prime will increase the other-sacrificial utilitarian (OU) parameter relative to the emotion prime, 3) giving in the public goods game will positively correlate with the self-sacrificial utilitarian parameter, 4) giving in the public goods game will positively correlate with the other-sacrificial utilitarian parameter, 5)

cognitive reflection test scores will be positively correlated with the self-sacrificial utilitarian parameter, 6) cognitive reflection test scores will be positively correlated with the other-sacrificial utilitarian parameter. All six of our predictions were confirmed.

We combined the participants from Studies 1-4 into one dataset (N = 1,418) and performed exploratory analyses. We did separate principal component analyses for self-sacrificial and other-sacrificial dilemmas. We broke the self-sacrificial utilitarian (SU) and other-sacrificial utilitarian (OU) parameters down into subcomponents, and found that each of the separate subcomponents positively correlated with cognitive reflection test scores. This constitutes additional corroboration of our hypothesis that utilitarian judgment is associated with reliance on cognitive reflection. We also divided participants into three clusters using K-means cluster analysis. The cluster with the highest SU and OU parameter scores (cluster 3) also had the highest scores on the cognitive reflection test. We also looked at whether these clusters differed by political orientation, given that cognitive reflection is positively correlated with being politically liberal (Lane & Sulikowski, 2017). Participants in cluster 3 were also the most liberal.

Method

Participants. We recruited 1,025 participants for this study from the Psychology subject pool and gave them all partial course credit for participating. After applying our pre-registered exclusions (https://osf.io/3yvn7), there were 696 participants left[19]. This sample size gave us more than 95% power to detect a correlation of $r = .163$ or greater (the smallest correlation we were trying to replicate), with a false positive rate of 5% and a one-tailed test. The ethnicity of

[19] We forgot to include questions for gender and age in this survey. In our pre-registration we said we would get a sample of 700.

the participants was as follows: 44.3% European (n = 308), 15.5% South Asian (n = 108), 11.9% East Asian (n = 83), 6.2% Black (n = 43), 4.2% South East Asian (n = 29), 1.9% Indigenous (n = 13), 1.6% Latin American (n = 11), 6.2% mixed or other (n = 43), and 1.3% did not answer (n = 9).

Procedure and materials. We used the same cognitive reflection test that was used in Studies 1-2. The reliability was again below minimal acceptability (Cronbach's $\alpha = 0.471$). We also used the same moral dilemmas that were used in Studies 1-3. For skewness, the SU, OU, A, and D parameters all had scores in between -0.5 and 0.5, which is consistent with univariate normality. For kurtosis, the OU and D parameters had scores in between -0.5 and 0.5, which is consistent with univariate normality. The SU and A parameters had respective scores of -0.655 and -0.870.

We used the same reason/emotion manipulation that was used in Study 2. We used the same public goods game as the one in Study 3, but with two differences: a) we only used the single-game version, and b) we randomized the order so that half the participants played the game at the beginning of the study (*before* the cognitive reflection test, reason/emotion manipulation, and dilemmas), and half the participants played the game at the end of the study (*after* the cognitive reflection test, reason/emotion manipulation, and dilemmas).

Data analytical plan. To replicate the results of Study 1 and Study 2, we planned to calculate a Pearson's correlation coefficient between the cognitive reflection test and our four moral parameters: self-sacrificial utilitarianism (SU), other-sacrificial utilitarianism (OU), altruism (A), and deontology (D). To replicate the results of Study 3, we also planned to calculate a Pearson's correlation coefficient between the total amount given in the single-game version of the public goods game and our four moral parameters. For our manipulation of

whether to instruct participants to rely on reason or emotion, we planned to use independent samples t-tests. Alpha was set at .05, and all of our preregistered directional predictions were tested with one-tailed tests. All other tests were two-tailed.

In accordance with our preregistration (https://osf.io/3yvn7) the exclusion criteria were as follows: 1) if a group did not have four participants, a bot took their place; if participants playing with bots gave significantly more or less than participants without bots in the multi-game version, they were excluded, 2) anyone who took longer than 3 hours was excluded, 3) five dilemmas were followed by a multiple choice comprehension question, and getting more than two wrong resulted in exclusion, 4) there were five questions at the end of the survey about how closely they paid attention, each of which resulted in exclusion if the answer indicates they were not paying attention. All of the attention questions are available in the above preregistration link.

Results

Descriptive statistics. On the cognitive reflection test, 16.1% of participants got none of the questions correct, 26.6% got one correct, 32.2% got two correct, and 25.1% got all questions correct. The average amount given was $M = 53.028$ ($SD = 31.510$). For the self-sacrificial utilitarian (SU) parameter, the average score was $M = .450$ ($SD = .238$). For the other-sacrificial utilitarian (OU) parameter, the average score was $M = .335$ ($SD = .186$). The average score for the altruism (A) parameter was $M = .453$ ($SD = .307$). The average score for the deontology (D) parameter was $M = .538$ ($SD = .204$).

Pre-registered Hypothesis 1. Those in the reason condition had higher self-sacrificial utilitarian (SU) parameter scores ($M = .484$, $SD = .236$) than those in the emotion condition ($M =$

.417, *SD* = .236), and this difference was significant: $t(694) = 3.733$, $p < .001$, $d = .283$. This hypothesis was confirmed.

Pre-registered Hypothesis 2. Those in the reason condition had higher other-sacrificial utilitarian (OU) parameter scores ($M = .348$, $SD = .185$) than those in the emotion condition ($M = .321$, $SD = .186$), and this difference was also significant: $t(694) = 1.951$, $p = .026$, $d = .148$. This hypothesis was confirmed.

The reason/emotion manipulation did not have any significant effect on the altruism (A) parameter, $t(694) = .036$, $p = .971$, but the reason manipulation significantly decreased scores on the deontology (D) parameter, $t(694) = 3.502$, $p < .001$, $d = -.265$.

Pre-registered Hypothesis 3. The amount donated in the single-game version of the public goods game positively correlated with the self-sacrificial utilitarian (SU) parameter, $r = .138$, $p < .001$. This hypothesis was confirmed.

Pre-registered Hypothesis 4. The amount donated in the single-game version of the public goods game positively correlated with the other-sacrificial (OU) parameter, $r = .130$, $p < .001$. This hypothesis was confirmed.

The amount donated in the single-game version of the public goods game also positively correlated with the altruism (A) parameter, $r = .123$, $p < .001$. However, it did not correlate with the deontology (D) parameter, $r = .040$, $p = .289$.

Pre-registered Hypothesis 5. The cognitive reflection test (CRT) positively correlated with the self-sacrificial utilitarian (SU) parameter, $r = .091$, $p = .008$. This hypothesis was confirmed.

Pre-registered Hypothesis 6. The cognitive reflection test (CRT) positively correlated with the other-sacrificial (OU) parameter, $r = .238$, $p < .001$. This hypothesis was confirmed.

The CRT marginally negatively correlated with the altruism (A) parameter, $r = -.062$, $p = .051$, and did not correlate with the deontology (D) parameter, $r = -.037$, $p = .162$.

Table 4

Correlations Among Variables, Study 4

Variable	1	2	3	4	5
1. Donation	—				
2. CRT	.062	—			
3. SU Parameter	.138***	.091**	—		
4. OU Parameter	.130***	.238***	.346***	—	
5. A Parameter	.123***	-.062†	.171***	-.019	—
6. D Parameter	.040	-.037	.049	.060	.156***

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

ANCOVA. To determine whether the effect of being in the reason condition was due to a failure of random assignment, we did two ANCOVAs, both of which had experimental condition as the independent variable and cognitive reflection test scores as the covariate. The effect of the reason manipulation was still significant for the self-sacrificial utilitarian (SU) parameter, $F(1, 693) = 14.606$, $p < .001$, and the other-sacrificial utilitarian (OU) parameter, $F(1, 693) = 4.854$, $p = .028$.

Principal component analysis. We combined the participants from Studies 1-4 into one dataset[20] (N = 1,418) in order to perform an exploratory principal component analysis on the self-sacrificial and other-sacrificial dilemmas. There is an asymmetry between how the

[20] This pooled sample also includes 65 participants who responded to the dilemmas in a separate study. In that study, participants had their brains scanned with EEG while they responded to the dilemmas. We chose to not include that study in this dissertation, as the data are still being analyzed.

congruent and incongruent dilemmas work in the self-sacrificial and the other-sacrificial process dissociation models. For self-sacrificial dilemmas, utilitarians endorse self-sacrifice in the *congruent* dilemmas, whereas for the other-sacrificial dilemmas they endorse other-sacrifice in the *incongruent* dilemmas. We ran a principal component analysis on those two types of dilemmas. We did this to break our measures of self-sacrificial and other-sacrificial utilitarianism down into factors. If only a subset of these factors positively correlated with cognitive reflection, then the association between cognitive reflection and utilitarian judgment is limited. If the correlation remains significant for each of the factors, this would constitute additional corroboration of our hypothesis.

For the self-sacrificial congruent dilemmas, we ran a principal component analysis with 25 iterations, and used the scree plot elbow criterion to decide on the number of factors (the elbow criterion gave us two factors). The two factors explain 42.650% of the variance.

Table 5

Factor loadings from principal component analysis, varimax rotation applied.

Dilemma	1	2
Trolley Congruent	.782	—
Cliff Congruent	.749	—
War Congruent	.735	—
Grenade Congruent	.721	—
Malaria Congruent	—	.749
Tsunami Congruent	—	.734
Vacation Congruent	—	.460
Organ Congruent	—	.422
Piano Congruent	—	.411
Family Congruent	—	.377

Note. — indicates that the factor loading was < .30.

For the other-sacrificial congruent dilemmas, we also did a principal component analysis with 25 iterations. However, the choice of how many factors to use was more difficult. If we used the scree plot elbow criterion, we got three factors, but they were uninterpretable. We decided to go with two factors because a) the pattern was uninterpretable, and b) to retain continuity with the previous analysis. The two factors explained 30.649% of the variance.

Table 6

Factor loadings from principal component analysis, varimax rotation applied.

Dilemma	1	2
Car Incongruent	.608	—
Baby Incongruent	.585	—
Vaccine Incongruent	.537	—
Abortion Incongruent	.477	—
Animal Incongruent	.440	—
Time Incongruent	.388	—
Relationship Incongruent	—	.618
Border Incongruent	—	.603
Torture Incongruent	—	.597
Hard Times Incongruent	—	.518

Note. — indicates that the factor loading was < .30.

To test for the robustness of the correlation between the cognitive reflection test (CRT) and the self-sacrificial utilitarian (SU) parameter, we broke down the SU parameter into two measures: one derived from the four dilemmas that loaded on component 1 (and their incongruent parallels), and one derived from the six dilemmas that loaded on component 2 (and their incongruent parallels). We did the same for the other-sacrificial utilitarian (OU) parameter, with one measure derived from the six dilemmas that loaded on component 1 (and their congruent parallels), and one derived from the four dilemmas that loaded on component 2 (and

their congruent parallels). All four resulting measures still significantly positively correlated with the CRT (all *p*s < .007). This indicates that cognitive reflection is consistently positively correlated with self-sacrificial utilitarianism and other-sacrificial utilitarianism, even when those two measures are broken down into factors.

K-means cluster analysis. We again used the pooled sample (N = 1,418), this time to extract clusters using K-means cluster analysis. We applied K-means cluster analysis to the set of all 40 dilemmas, and extracted three clusters. If the cluster that had the highest scores on the measures of utilitarian judgment also had the highest cognitive reflection (CRT) scores, that would constitute further corroboration of our hypothesis. We compared members of the three clusters by their mean scores on the four moral parameters: self-sacrificial utilitarianism (SU), other-sacrificial utilitarianism (OU), altruism (A), and deontology (D). Even though the CRT and self-reported political attitudes were not used to extract the three clusters, we also calculated the means for those measures by cluster. Cluster 1 included 377 participants, cluster 2 included 568 participants, and cluster 3 included 473 participants[21].

Table 7

Means (and standard deviations in parentheses) for the 3 clusters, which were derived from K-means cluster analysis applied to moral dilemma responses.

	SU	OU	A	D	CRT	Politics
Cluster 1 "Amoralists"	.230	.289	.203	.434	1.639	-.231
	(.181)	(.175)	(.154)	(.194)	(1.108)	(1.529)
Cluster 2 "Bleeding Hearts"	.423	.306	.707	.620	1.552	-.702
	(.215)	(.196)	(.179)	(.191)	(1.042)	(1.558)
Cluster 3 "Utilitarians"	.518	.382	.310	.588	1.822	-.934
	(.201)	(.169)	(.219)	(.188)	(1.002)	(1.495)

Note. Political ideology on a scale of -3 (very liberal) to +3 (very conservative).

[21] The CRT and the political attitudes question were only in some of the studies that were pooled, so the sample size is as follows for those measures. For the CRT: cluster 1 includes 325 participants, cluster 2 includes 426 participants, and cluster 3 includes 381 participants. For the political attitudes question: cluster 1 includes 303 participants, cluster 2 includes 403 participants, and cluster 3 includes 349 participants.

Members of cluster 1 were the lowest[22] on all four moral parameters measures. Interestingly, they were higher on the deontology parameter than the other parameters. We refer to this cluster as the "amoralists". They were also significantly less reflective than those in cluster 3 (while not being significantly different from cluster 2), and they were the least liberal. Members of cluster 2 were highest in both deontology and altruism, and were lowest in reflection. They were significantly less reflective than cluster 3, but not significantly different from cluster 1. We refer to this cluster as the "bleeding hearts". The most theoretically significant cluster is cluster 3. Members of this cluster were significantly higher in both utilitarian parameters, and they were significantly higher in cognitive reflection (compared to the other two clusters). They were also significantly more liberal than the other two clusters. We refer to this cluster as the "utilitarians".

Discussion

All six of our pre-registered predictions were confirmed. We replicated the positive correlation between single-game giving in the public goods game and both of the utilitarian parameters (SU and OU). With our larger sample, we also found that giving positively correlated with the altruism (A) parameter, so that the only parameter that was not positively correlated with giving was the deontology (D) parameter. The positive correlation between the cognitive reflection test (CRT) and self-sacrificial utilitarianism (SU) and other-sacrificial utilitarianism (OU) was also replicated, as was the fact that being in the reason prime condition increased SU

[22] We used independent samples t-tests to compare clusters 1, 2, and 3 on their moral parameter scores, CRT scores, and political attitudes. The differences in their responses to the moral parameters were all significant (all ps < .007). For CRT scores, cluster 3 had significantly higher scores than 1 and 2 (all ps < .018), but cluster 1 and cluster 2 were not significantly different from each other. For political attitudes, both comparisons were statistically significant (both ps < .038).

and OU parameter scores. This constitutes strong evidence for the hypothesis that cognitive reflection drives utilitarian responses to self-sacrificial and other-sacrificial dilemmas.

When we pooled the samples together, the correlation between the cognitive reflection test (CRT) and the SU and OU parameters remains significant even when the SU and OU parameters are each broken down into separate factors via principal component analysis. For self-sacrificial dilemmas, the dilemmas that load heavily on component 1 are all about self-sacrifice that involves dying, whereas all the dilemmas that load heavily on component 2 involve a form of giving other than dying. The one exception is the organ dilemma, which loads more heavily on component 2 despite the fact that it involves dying via giving your organs. This result would be consistent with organ donation feeling psychologically more like an actual donation, and less like stepping in front of a trolley (even though both involve self-sacrificial death). In the case of other-sacrificial dilemmas, the dilemmas that load heavily on component 1 are all about other-sacrifice that involves killing another person. The one exception the animal dilemma, which loads more heavily on component 1 despite the fact that it involves harming an animal. Perhaps people (or university students in particular) are sensitive to harming animals to such a high degree that it feels comparably wrong to killing. The data do not allow us to test this explanation. All the dilemmas that load heavily on component 2 involve a form of harm other than killing, except for the border dilemma, which involves killing a potential foreign terrorist. It seems initially plausible on this basis to conjecture that people are callous about killing in this case, and that they don't view it as genuine killing. However, this explanation is rendered unlikely by the fact that of all the incongruent dilemmas that involve killing, killing is endorsed at the lowest rate for this dilemma (see the Appendix).

It is also noteworthy that when participants are broken down into three clusters, the cluster with the highest self-sacrificial utilitarian (SU) and other-sacrificial utilitarian (OU) parameter scores (cluster 3, the utilitarians) also has the highest cognitive reflection test (CRT) scores. This is consistent with the association between cognitive reflection and utilitarian judgment being a robust pattern. Members of cluster 3 and cluster 2 (the bleeding hearts) were both significantly more liberal than members of cluster 1 (the amoralists), but they otherwise had very different psychological profiles. Members of cluster 3 were highly reflective and highly utilitarian, whereas members of cluster 2 were low on reflection and high on deontology and altruism. Prior studies have found that liberals are more reflective than conservatives (Lane & Sulikowski, 2017). The results of the cluster analysis are consistent with there being two distinct ways to be liberal: being a cold, calculating utilitarian (cluster 3), and being a "bleeding heart liberal" who is low in cognitive reflection (cluster 2).

Study 5

For this study[23], we focused on the self-sacrificial measures (i.e. self-sacrificial utilitarianism and altruism) and focus on how they relate to different human relationships. Earp et al. (2021) theorized that there are four main social relationship functions in our species: care, cooperation/reciprocal relationships, hierarchical relationships and mating relationships (see also Bugental, 2000). Care relationships are characterized by unconditional help and support, they would include, for example, parental relationships and sibling relationships. Cooperative relationships are characterized by equal status and reciprocity. Examples would include

[23] This study was actually done before Study 4, but we have placed it here because the order makes better logical sense.

friendships, teammates, and roommates. Hierarchical relationships are characterized by an inequality in status, such as employer-employee relationships. Finally, mating relationships are characterized by sexual contact or romantic commitment. We wrote five sets of versions of the self-sacrificial dilemmas, one corresponding to each of the relationship types ("sibling" for care, "friend" for cooperation, "subordinate at work" for hierarchical, and "significant other" for mating), and one corresponding to no relationship ("stranger"). Participants were randomly assigned to get one of these sets.

Additionally, we added the several constructs that been found to be related to moral measures in other studies for exploratory purposes. An anxious attachment style is negatively correlated with utilitarian responses to other-sacrificial dilemmas (Maranges et al., 2022), so we included the short form of the Experiences in Close Relationships (ECR) measure of attachment style (Wei et al., 2007). Psychopathy is positively correlated with utilitarian responses to other sacrificial dilemmas when process dissociation is not used (Kahane et al., 2015), but negatively correlated when process dissociation is used (Conway et al., 2018). We included the Triarchic Psychopathy Measure (TRiPM; Patrick, 2010) to test what (if any) relationship it has with self-sacrificial utilitarianism and altruism. Finally, we included the Religious Zeal Scale (McGregor et al., 2008). Religiosity negatively correlates with cognitive reflection (Stagnaro et al., 2018), so it is unsurprising that religiosity also negatively correlates with other-sacrificial utilitarianism (Szekely et al., 2015). This is a reason to predict that religiosity would also negatively correlate with self-sacrificial utilitarianism. Since there is a positive correlation between religiosity and prosocial behavior (Shariff, 2015), we would expect religiosity to positively correlate with altruism.

Method

Participants. We recruited 646 participants for this study from the University of Alberta subject pool and gave them all partial course credit for participating. After applying the same exclusion criteria as we did in Studies 1-3, there were 567 participants left (360 females, 195 males, 11 other, 1 did not answer). The age range was 17-35, and the average age was 19.0. The ethnicity of the participants was as follows: 42.3% European (n = 240), 16.6% South Asian (n = 94), 11.1% East Asian (n = 63), 6.0% South East Asian (n = 34), 3.2% Black (n = 18), 1.6% Indigenous (n = 9), 1.4% Latin American (n = 8), 9.3% mixed or other (n = 53), and 4.9% did not answer (n = 28).

Cognitive reflection test. We used the same version of the CRT that was used in Study 1 and Study 2.

Attachment style. We wanted to test whether attachment style has any relationship to utilitarian responses to self-sacrificial dilemmas. We used the short form of the Experiences in Close Relationships (ECR) measure of attachment style (Wei et al., 2007). This scale gives two dimensions of attachment style: the avoidance dimension, which measures one's avoidance of relational, and the anxiety dimension, which measures one's anxiety about relationships.

Psychopathy. We included the Triarchic Psychopathy Measure (TRiPM; Patrick, 2010). This measure breaks down into three scales: Boldness, Meanness, and Disinhibition. The Meanness scale also includes an Empathy facet. There is a large body of evidence for that the feeling of empathy drives altruism (Batson et al., 1989; Persson & Kajonius, 2016). Even though altruism is operationalized differently in this study, it would be theoretically significant if empathy positively correlated with it. It would constitute evidence of convergent validity

between my operationalization of altruism and prior operationalizations. The full psychopathy scale is available in the Appendix.

Religious zeal. This scale measures the degree of zealous commitment to one's religious beliefs (McGregor et al., 2008). As was mentioned in the introduction, there is some reason to expect it to negatively correlate with self-sacrificial utilitarianism, but positively correlate with altruism. It is available in the Appendix.

Moral dilemmas. The prior four scales were given in a random order, and then the moral dilemmas were given last. We wrote five sets of versions of the self-sacrificial dilemmas, one corresponding to each of Earp's (2021) relationship types ("sibling" for care, "friend" for cooperation, "subordinate at work" for hierarchical, and "significant other" for mating), and one corresponding to no relationship ("stranger"). Participants were randomly assigned into these five conditions, each participant only sees one type of relationship dilemma (e.g. if they were in the sibling condition, they read 20 dilemmas about self-sacrifice on behalf of a sibling). Other than the added relational content, the dilemmas are the same as those in the Appendix. For skewness, the SU and A parameters all had scores in between -0.5 and 0.5, which is consistent with univariate normality. For kurtosis, the SU parameter had a score in between -0.5 and 0.5, which is consistent with univariate normality. However, the A parameter had a score of -0.956.

Data analytical plan. We planned to calculate a Pearson's correlation coefficient between the scales we included and the two moral parameters: self-sacrificial utilitarianism (SU) and altruism (A). We also planned to analyze how likely people were to endorse self-sacrifice on behalf of different relationships: a sibling, a friend, a subordinate at work, a significant other, or a stranger. We also planned to use self-sacrifice in congruent dilemmas (i.e. dilemmas in which both altruists and self-sacrificial utilitarians would give self-sacrificial responses) as the basis for

ordering the five relationship types. Relationship type was then used as an ordinal variable which was correlated with self-sacrificial utilitarianism (SU) and altruism (A). Since relationship type was an ordinal variable, we calculated a Spearman's rank correlation. Alpha was set at .05, and all of the tests were two-tailed.

We did not preregister this study, but we used the same exclusion criteria as Study 1 and Study 2: 1) anyone who took longer than 3 hours was excluded, 2) five dilemmas were followed by a multiple choice comprehension question, and getting more than two wrong resulted in exclusion, 3) there were five questions at the end of the survey about how closely they paid attention, each of which resulted in exclusion if the answer indicated they were not paying attention.

Results

Descriptive statistics. For the self-sacrificial utilitarian (SU) parameter, the average score was $M = .390$ ($SD = .186$). The average score for the altruism (A) parameter was $M = .527$ ($SD = .281$). Both of these were obtained by averaging across relationship conditions. The range of possible scores on the ECR attachment scale is 6-42 for both the anxiety and avoidance dimensions. The average score on the anxiety dimension was $M = 24.841$ ($SD = 6.825$), and the average score on the avoidance dimension was $M = 17.572$ ($SD = 6.730$). The range of possible scores on the TriPM scale is 58-234. The average score was $M = 118.059$ ($SD = 18.617$). The range of possible scores on the Religious Zeal Scale is 19-95. The average score was $M = 50.663$ ($SD = 15.124$).

Relationship variable. We turned the five relationship types into an ordinal variable (ranging from 1 to 5). We used the decision to self-sacrifice in congruent dilemmas (i.e.

dilemmas in which both altruists and self-sacrificial utilitarians would give self-sacrificial responses) as the basis for ordering the five relationship types. Endorsement of self-sacrifice in every congruent dilemma would yield a mean score of 1, endorsement of self-sacrifice in no congruent dilemmas would yield a mean score of 0. The order (from the least self-sacrificial to the most self-sacrificial) was as follows: stranger (M = .503, SD = .222), subordinate (M = .623, SD = .190), friend (M = .732, SD = .175), significant other (M = .797, SD = .181), and sibling (M = .848, SD = .122). The pairwise comparisons were all significant (all ps < .034).

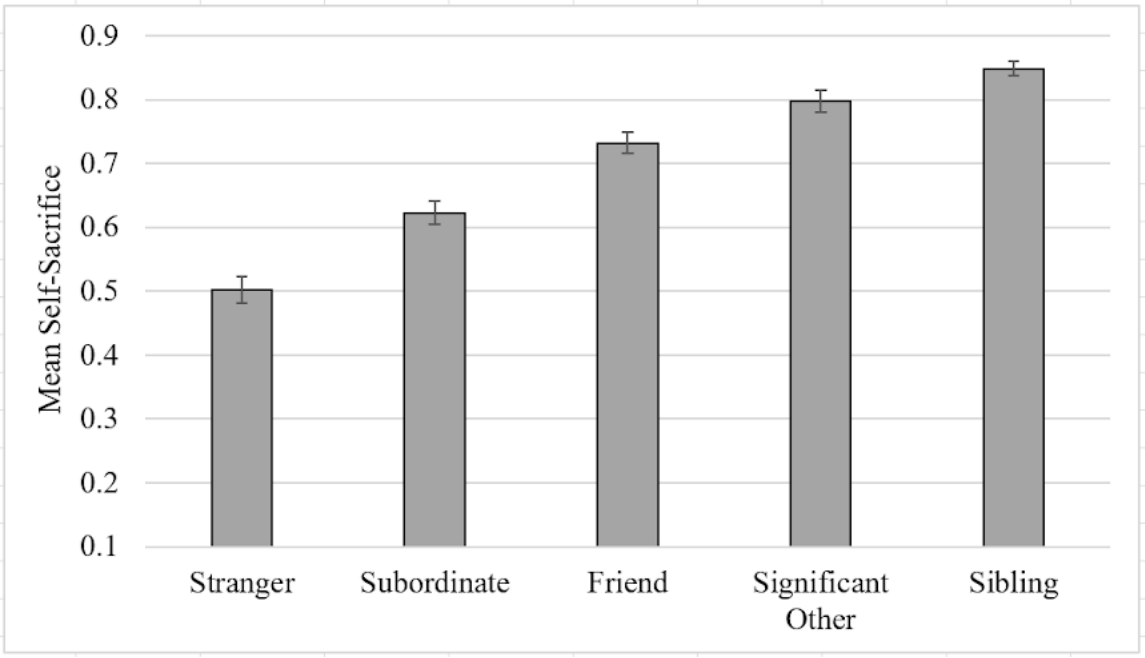

Figure 3. Mean endorsement of self-sacrifice in the congruent dilemmas by relationship type.

Exploratory analyses. The TRiPM breaks down into 3 components: boldness, disinhibition, and meanness. Boldness did not correlate with the self-sacrificial utilitarian (SU) parameter or the altruism (A) parameter (all ps > .524). Disinhibition negatively correlated with the SU parameter, r = -.099, p = .019, but did not correlate with the A parameter (p = .303).

Meanness was the best predictor; it negatively correlated with both the SU parameter, $r = -.123$, $p = .003$, and the A parameter, $r = -.132$, $p = .002$. Meanness further breaks down into four facets[24]. Of these four, empathy was the only one which significantly correlated with both the SU and the A parameter. We reverse-coded the empathy facet so that a high score corresponds to high empathy[25]. This measure of empathy positively correlated with the SU parameter, $r = .137$, $p = .001$, and with the A parameter, $r = .125$, $p = .003$. Religious zeal positively correlated with the A parameter, $r = .087$, $p = .039$, but not with the SU parameter, $r = -.048$, $p = .251$. Neither CRT scores, nor the two dimensions of attachment style were significantly correlated with the SU parameter or the A parameter (all ps $> .179$).

Since relationship type is an ordinal variable, we used Spearman's rank correlation. The relationship type variable correlated positively with the self-sacrificial utilitarian (SU) parameter, $r = .279$, $p < .001$, and the altruism (A) parameter, $r = .539$, $p < .001$. Using Fischer's Z transformation, we determined that the correlation for the A parameter was significantly larger, $p < .001$.

Discussion

For the congruent dilemmas (which are the ones where altruists and utilitarians should both endorse self-sacrifice), there was a clear order for which relationships people were more willing to self-sacrifice on behalf of. When we turned relationship type into an ordinal variable (ordered by mean self-sacrifice in congruent dilemmas), it correlated approximately twice as strongly with altruism than it did with self-sacrificial utilitarianism. This indicates that altruistic

[24] The four are: empathy, physical aggression, relational aggression, and honesty (Patrick, 2010).
[25] When it is a part of the TRiPM, a high score on the "empathy facet" means having less empathy (since it is counting towards an overall psychopathy score).

responding is more dependent on relationship type, but self-sacrificial utilitarian responding is more consistent across relationship types. Utilitarianism in theory involves impartiality because it means a commitment to maximizing happiness for everyone, regardless of what their relationship to you is (De Lazari-Radek & Singer, 2014). It is therefore a good feature of our operationalization of self-sacrificial utilitarianism that it is associated with more impartiality than altruism.

The empathy measure from the Triarchic Psychopathy Measure (reverse coded such that higher scores meant higher empathy) was positively correlated with both altruism and self-sacrificial utilitarianism. The connection between empathy and altruism is consistent with prior work done on the "empathy-altruism hypothesis" (Batson et al., 1989; Persson & Kajonius, 2016). The correlation between empathy and self-sacrificial utilitarianism may seem at first glance to be contrary to Greene's dual-process model. But Greene (2008) has conceded that in order for any moral judgment (including utilitarian judgment) to get off the ground, you have to have some affectively based care for others. In the case of utilitarian judgment, before you engage in the act of calculating the effect of your actions on happiness and suffering, you have to care about happiness and suffering. This study provides evidence that empathy could be the basis of caring about the happiness and suffering of others. A more surprising result of this study is that the cognitive reflection test was not positively correlated with the self-sacrificial utilitarianism in this study, unlike in Studies 1, 2, and 4. This could be because it included a relationship element, unlike our prior studies. It could also be because those studies had self-sacrificial and other-sacrificial dilemmas mixed together, and that responding in a consistently utilitarian manner requires more concentration and reflection because one has to keep track of what kind of dilemma one is reading.

General Discussion

There is a large body of studies looking at utilitarian responses to moral dilemmas. There are studies which find that utilitarian responses are associated with a variety of constructs that relate to cognitive reflection, including cognitive load (Greene et al., 2008), dorsolateral prefrontal cortex activity (Zheng et al., 2018), and instructions to rely on deliberation (Suter & Hertwig, 2011). However, virtually all of the relevant studies use moral dilemmas that involve harming someone else for the greater good. Even sophisticated process dissociation models only include dilemmas that fall under this paradigm (Conway & Gawronski, 2013; Gawronski et al., 2017). We put forward our own process dissociation model for self-sacrificial dilemmas. In Study 1 we found that the cognitive reflection test (CRT) positively correlated with self-sacrificial utilitarianism and other-sacrificial utilitarianism. In Study 2 we replicated this correlation, and found that instructing people to rely on reason increased self-sacrificial and other-sacrificial utilitarian judgments. However, this effect was not significant. In Study 4 we performed the same manipulation with a larger sample, and found that the effect was significant. The fact that the instruction to rely on reason significantly increased utilitarian judgment provides convincing evidence of a causal connection between cognitive reflection and utilitarian judgment.

With regard to the public goods game, prior studies have found that deliberation is associated with giving in the multi-game version, but not the single-game version (Rand, 2016). For this reason, we were initially expecting that giving in the multi-game version of the public goods game (but not in the single-game version) would positively correlate with both types of utilitarian responses. In Study 3, this prediction was not confirmed. In fact, giving in the single-

game version marginally positively correlated with self-sacrificial utilitarianism, and significantly positively with other-sacrificial utilitarianism. In Study 4 we detected a significant positive correlation for each utilitarian parameter (as well as the altruism parameter) when we used a larger sample.

A potentially important difference between the versions is that giving in the single-game version is more genuinely self-sacrificial, because it has no strategic benefit. It is possible that this explains its positive correlation with self-sacrificial utilitarianism (SU), other-sacrificial utilitarianism (OU), and altruism (A). We could think of this positive correlation as behavioral validation of those scales, since it shows that the scales predict genuinely self-sacrificial/prosocial behavior. Additionally, prosocial behavior in this context can be construed as utilitarian, because it maximizes benefits for the group overall. The deontology (D) parameter measures an unwillingness to harm others, and it was not correlated with giving in the single-game public goods game, even in Study 4 when we used a large sample. This is evidence that deontology is not related to a willingness to help others. It is significant that utilitarian responses to both self-sacrificial and other-sacrificial dilemmas predict prosocial behavior, because a prior study found that utilitarian responses had no correlation with behavior in response to real-life dilemmas (Bostyn et al., 2018).

Our results are also relevant to a longstanding debate between defenders and critics of other-sacrificial dilemmas (Greene, 2008; Kahane et al., 2015; Conway et al., 2018; Everett & Kahane, 2020). The critics favour measures like the Oxford utilitarianism scale, which includes a measure of impartial beneficence and a measure of instrumental harm (Kahane et al., 2018). With one exception, all the items in the impartial beneficence scale are about self-sacrifice. As was argued in the introduction, endorsement of self-sacrifice could be driven by utilitarianism, or

by altruism (in our sense of the term). We found that both self-sacrificial utilitarian and altruistic tendencies independently explain a substantial proportion of the variance in impartial beneficence scores. The instrumental harm scale was explained by participants having high other-sacrificial utilitarian scores and low deontology scores. This pattern suggests that it may be possible for the two camps (proponents and critics of sacrificial dilemmas) to integrate their theories. This is because the Oxford utilitarianism scale has convergent validity with the four parameters we extracted (via process dissociation) from sacrificial dilemmas. In light of our findings, future studies using the impartial beneficence scores should control for altruistic tendencies, at least if they want to claim that the scale is measuring utilitarian tendencies.

When we pooled the samples from Studies 1-4, we used factor analysis to uncover two factors undergirding self-sacrificial utilitarianism (SU) and two factors undergirding the other-sacrificial utilitarianism (OU). We found that the correlation between the cognitive reflection test and the SU and OU parameters remained significant even when we divided the SU and OU parameters each into two factors. This shows that the correlation between cognitive reflection and utilitarian judgment is not limited to a subset of dilemmas. Additionally, when we divided participants into three clusters using k-means cluster analysis, the cluster (cluster 3) with the highest SU and OU parameter scores also had the highest cognitive reflection test scores. These results speak to the robustness of the association between cognitive reflection and utilitarian judgment in self-sacrificial and other-sacrificial contexts. Interestingly, people in cluster 3 also were the most liberal on average, which is consistent with prior work finding that liberals are on average more reflective than conservatives (Lane & Sulikowski, 2017).

This dissertation contains the first study that combines Earp's (2021) relationship categories and measures of utilitarian tendencies. We only put the two together in the context of

self-sacrificial dilemmas, future studies could do the same with other-sacrificial dilemmas. When we rank order relationships by how much self-sacrifice people endorse (in the dilemmas that utilitarians and altruists should both endorse self-sacrifice), people are most willing to engage in self-sacrifice for a relative. This result is consistent with kin selection theory, which posits that people have a tendency to sacrifice their interests for kin because they share genes in common (West-Eberhard, 1957). Relationship type as a variable affects altruism much more than it does self-sacrificial utilitarianism, which suggests altruistic tendencies are more driven by relationship context, whereas utilitarian tendencies are more constant across relationship types.

One limitation is that our main measure of reliance on cognitive reflection was the cognitive reflection test (CRT), which did not have high internal reliability. This could be due to the fact that the test only has three items; tests with fewer items have lower Cronbach's alpha values (Taber, 2018). In the future, researchers could try to develop versions of the test with higher internal reliability. Thomson and Oppenheimer (2016) developed a version of the CRT with higher internal reliability. However, it includes non-mathematical items, and therefore would probably not positively correlate with utilitarian judgment (see Byrd & Conway, 2019). One issue with psychological tests in general, including the CRT, is that they do not solely measure the construct they purport to measure, and typically measure multiple constructs (Meehl, 1990). In the case of the CRT, the theoretical construct it purports to measure is the tendency to reflect on one's intuitions and correct them when necessary. However, the test also strongly correlates with general cognitive ability (Toplak et al., 2011) and numeracy (Erceg et al., 2020), which are different constructs. If the dual-process model is right, it should be cognitive reflection (as opposed to cognitive ability or numeracy some other construct) that is actually explaining the relationship between the CRT and our two utilitarian parameters. Future

studies could control for IQ scores or numeracy to test the possibility that a different construct explains the relationship.

The moral parameters also have some potential limitations. Given the way we constructed the process dissociation model, what distinguishes self-sacrificial utilitarians from altruists is that they are less willing to engage in self-sacrifice in certain situations (i.e. situations in which self-sacrifice does not maximize the greater good). However, in many situations, what distinguishes utilitarian self-sacrifice is that it involves a greater willingness to engage in self-sacrifice (Singer, 1972). This is a potential shortcoming of the model that future researchers could try to overcome. Another limitation is that our sample is not cross-culturally diverse. Since prior studies have found that there are cross-cultural differences in responses to other-sacrificial dilemmas (Ahlenius & Tännsjö, 2012; Qian et al., 2023), it is plausible that differences will exist when it comes to self-sacrificial dilemmas as well. Future studies should include cross-cultural replications. Additionally, when it comes to other-sacrificial dilemmas, a newer process dissociation model called the CNI model has added an inaction parameter (Gawronski et al., 2017). This parameter controls for a general preference for inaction, independent of whether the action accords with utilitarianism or deontology. Future studies could add an inaction parameter to our process dissociation model for self-sacrificial dilemmas.

To sum up, prior studies finding a correlation between reliance on cognitive reflection and utilitarian judgment are insufficient because they only include dilemmas that involve harming someone else for the greater good. We found evidence that there is also a connection between cognitive reflection and utilitarian judgment in self-sacrificial situations. Furthermore, utilitarian responses to these dilemmas predict pure cooperation in the single-game version of the public goods game, which indicates that utilitarian responses to moral dilemmas predict

utilitarian prosocial behavior. And finally, relationship context (what relationship you have with the person you are helping) affects self-sacrificial utilitarian judgments less than it affects altruistic judgments.

Conclusion

I will address whether this work has any practical implications. Greene (2015) argues that the dual-process model of moral cognition has practical implications. As explained earlier, he defends the reliability of cognitive reflection for unfamiliar moral dilemmas through an appeal to what he calls “the no cognitive miracles argument”. He claims that automatic mode should only be relied upon for familiar moral problems, which he defines as problems with which we have “trial-and-error experience”, either rooted in evolution, cultural development, or individual trial-and-error learning. Absent one of those three sources of reliability, it would be completely inexplicable (i.e. a miracle) if intuition were reliable in dealing with an unfamiliar moral problem. Since the trolley dilemmas center around a rare situation (which precludes individual experience) and involves recently invented technology (which precludes either biological evolution or cultural evolution giving rise to norms), it would fall under Greene’s definition of an unfamiliar moral problem. Furthermore, Greene says that disagreement about a moral dilemma is a good proxy for unfamiliarity, so long as it is not best explained in terms of disagreement about nonmoral facts. When conjoined with the premise that cognitive reflection causes utilitarian moral judgments, the conclusion is that utilitarianism should be relied upon for unfamiliar moral problems.

In his other work, Greene (2013) also includes distributing resources in a massively unequal global economy as an example of an unfamiliar moral problem. He infers that

utilitarianism should be relied on in this case as well, which has self-sacrificial implications. For those of us who live in the developed world, the utilitarian solution will be for all of us to give up significant amounts of our finances to aid those in the developing world. This is a practical consequence of Greene's argument. However, prior to our work, no studies have explicitly addressed the question of whether cognitive reflection is associated with self-sacrificial utilitarian judgments. Given that Greene's argument has been put forward as a reason to endorse utilitarian self-sacrifice, it is important to test his empirical premise that cognitive reflection undergirds utilitarian judgment in self-sacrificial cases. If it only undergirds utilitarian judgments in other-sacrificial dilemmas, then his argument in favor of utilitarian self-sacrifice is unsound. If, on the other hand, cognitive reflection is associated with self-sacrificial utilitarian judgments, this would provide better support for Greene's argument that we should accept utilitarianism for unfamiliar moral problems, including in contexts that involve self-sacrifice. If we accept Greene's argument that cognitive reflection is more reliable for unfamiliar moral problems (including problems having to do with income distribution in the modern world), then our results provide support for the idea that we should engage in more utilitarian self-sacrifice. For example, if we are confident that the money that we give to charity X will benefit people in the developing world more than the money would benefit us, the utilitarian decision would be to give to that charity. Given that our results indicate that this kind of utilitarian self-sacrifice is on average rooted in cognitive reflection, our results (in conjunction with Greene's argument that we should rely on cognitive reflection for evolutionarily unfamiliar moral problems) provides some support for engaging in more utilitarian self-sacrifice.

One objection to this dissertation relates to the kind of theory it is based on. Machado and Silva (2007) argue that there is a spectrum of what qualifies as a theory. At one end (the "strong"

end), there are theories like Newtonian theory which involve a set of mathematically precise principles which are logically connected, from which precise predictions can be logically deduced. These are hypothetico-deductive theories. On the other end (the "weak" end), there are theories which have verbally instantiated principles that are connected more loosely. Most theories in psychology fall more on the weak end of the spectrum. Since the theories are put forward in terms of statements like "utilitarian judgments are associated with reliance on manual mode", they typically only "entail" (in a weak sense) that there will be an effect in a particular direction (Meehl, 1967). In this case, Greene's (2015) dual-process model leads to the prediction that the cognitive reflection test will be positively correlated with the self-sacrificial and other-sacrificial utilitarian parameters. But it does not specify how large this correlation is expected to be. The model simply does not have the resources to entail a more precise prediction. So long as the relationship is in the predicted direction, and the p-value falls below .05, this is taken as corroboration of the model. For examples widely cited papers in this area that use statistical significance as the cutoff for a hypothesis being supported, see Kahane et al. (2015) and Conway et al. (2018). They use the same standard, despite the fact that these papers are on opposite sides of the debate over the usefulness of trolley-style dilemmas for research into utilitarian judgment.

Therefore, in saying that these studies support the dual-process model, I am taking for granted the standards of the research program that I am working within. I have stepped into a research program (the dual-process approach to moral judgment) and tried to move it forward. By the standards used within this research program, this set of studies constitutes corroboration of Greene's dual-process model. So long as there is enough of an effect that it is detectable (according to the criterion that effect is in the right direction and the p-value is below .05), the model is taken to be corroborated. The replication crisis has taught us that detecting a directional

effect that can be replicated is difficult, especially when one uses pre-registration (as I did) so that researcher degrees of freedom are held constant. Even though the effect sizes were not especially large, I believe that I have detected a real pattern.

My use of the term “real pattern” is informed by advances in the philosophy of science. One of the big debates in philosophy of science concerns scientific realism, the view that mature, well-tested scientific theories are true or at least approximately true (Godfrey-Smith, 2009). The most widely cited argument in favor of scientific realism is the no miracles argument. In a nutshell, this argument states that mature scientific theories lead to accurate predictions, and that it would be a miracle if they could do this without being at least approximately true. Since there are no miracles, mature scientific theories must be approximately true (Putnam, 1975, as cited in Alai, 2023). The best argument against scientific realism is the pessimistic induction from the history of science (Psillos, 1996). Defenders of this argument point out that predictively successful theories in the past (like Newton’s theory of gravity) turned out to be technically false, and so we cannot infer on the basis of predictive success that a theory is true.

One version of scientific realism that can be supported by the no miracles argument, but avoid the pessimistic induction, is structural realism. This view allows for a more precise definition of what it means for a theory to be approximately true. A scientific theory is at least approximately true if it captures real mathematical relations or structure. Worrall (1989) argues that well-tested mathematical structures survive theory change. For example, he points out that even though Newton’s theory of gravity was technically falsified by observations and superseded by Einstein’s theory of gravity, Newton’s theory emerges as an approximation of Einstein’s theory within a certain domain of applicability. Ladyman and Ross (2007) developed the idea of

structural realism further using the idea of "real patterns" from Dennett (1991). On their account, a scientific theory is true (or approximately true) if it describes a real pattern[26].

An example from psychology that might help to illustrate is classical conditioning. In behaviorism, classical conditioning is conceptualized as an unconscious automatic mechanism of association. After the cognitive revolution, it has generally been conceptualized as a higher order cognitive process related to expectation (with some evidential basis, see Brewer 1974). However, what survives theory change is the pattern that when a stimulus X that produces response Z is repeatedly paired with stimulus Y, stimulus Y will start to also produce response Z. This "real pattern" is still there, it is just reconceptualized in the new theoretical framework.

Theoretical frameworks are important for helping scientists to generate new predictions, and to integrate the findings. But the previous examples show that the frameworks themselves often do not survive theory change. What survives are the mathematically describable patterns. While the dual-process model may be superseded by a deeper theory that explains a wider range of data, the well replicated patterns that have hitherto been explained by the dual-process model will have to be accounted for in any future theory. If the patterns I have found in this dissertation are replicated in future studies, they will likely survive future theory changes.

[26] Interestingly, this account blurs the line to some degree between scientific realism and scientific anti-realism. Bas Van Fraassen (1980) is perhaps the most sophisticated contemporary defender of scientific anti-realism. He claims that the goal of science is not to discover the truth, but merely to put forward empirically adequate theories that can account for observations. The most difficult argument for any anti-realist position is the no miracles argument. His response to the no miracles argument is the following: "I claim that the success of current scientific theories is no miracle. It is not even surprising to the scientific (Darwinist) mind. Any scientific theory is born into a life of fierce competition, a jungle red in tooth and claw. Only the successful theories survive—the ones which in fact latched on to actual regularities in nature" (Van Fraassen, 1980, p. 40). He is forced to appeal to the existence of actual regularities in nature in order to respond to the no miracles argument, despite the fact that he claims to be a scientific anti-realist. Van Fraassen's account is thus remarkably similar to structural realism (which appeals to "real patterns"), which is perhaps the most sophisticated contemporary version of scientific realism. This constitutes a kind of convergence between scientific realism and scientific anti-realism.

References

Ahlenius, H., & Tännsjö, T. (2012). Chinese and Westerners respond differently to the trolley dilemmas. *Journal of Cognition and Culture, 12*(3-4), 195-201.

Alai, M. (2023). Scientific realism, metaphysical antirealism and the no miracle arguments. *Foundations of Science, 28*(1), 377-400.

Amormino, P., Ploe, M. L., & Marsh, A. A. (2022). Moral foundations, values, and judgments in extraordinary altruists. *Scientific Reports, 12*(1), 22111.

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.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology, 40*(2), 290-302.

Batson, C. D., Batson, J. G., Griffitt, C. A., Barrientos, S., Brandt, J. R., Sprengelmeyer, P., & Bayly, M. J. (1989). Negative-state relief and the empathy—altruism hypothesis. *Journal of Personality and Social Psychology, 56*(6), 922.

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

Bentham, J. (1789). *An introduction to the principles of morals and legislation*.

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making, 12*(2), 148-167.

Bostyn, D. H., De Keersmaecker, J., Van Assche, J., & Roets, A. (2020). Bright mind, moral mind? Intelligence is unrelated to consequentialist moral judgment in sacrificial moral dilemmas. *Psychonomic Bulletin & Review, 27*, 392-397.

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General, 146*(5), e1-e7.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084-1093.

Bostyn, D. H., Roets, A., & Conway, P. (2022). Sensitivity to moral principles predicts both deontological and utilitarian response tendencies in sacrificial dilemmas. *Social Psychological and Personality Science, 13*(2), 436-445.

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences, 106*(18), 7351-7356.

Brosnan, M., & Ashwin, C. (2023). Thinking, fast and slow on the autism spectrum. *Autism, 27*(5), 1245-1255.

Brown, W. M., & Moore, C. (2000). Is prospective altruist-detection an evolved solution to the adaptive problem of subtle cheating in cooperative ventures? Supportive evidence using the Wason selection task. *Evolution and Human Behavior, 21*(1), 25-37.

Bugental, D. B. (2000). Acquisition of the algorithms of social life: a domain-based approach. *Psychological Bulletin, 126*(2), 187-219.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics, 82*, 101455.

Brewer, W. F. (1974). There is no convincing evidence for operant or classical conditioning in adult humans. In *Cognition and the Symbolic Processes* (pp. 1-42). Routledge.

Brosnan, S. F., & De Waal, F. B. (2003). Monkeys reject unequal pay. *Nature, 425*(6955), 297-299.

Brosnan, S. F., & de Waal, F. B. (2014). Evolution of responses to (un) fairness. *Science, 346*(6207), 1251776.

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition, 192*, 103995.

Capraro, V. (2024). The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *Journal of Personality and Social Psychology*.

Capraro, V., Corgnet, B., Espín, A. M., & Hernán-González, R. (2017). Deliberation favours social efficiency by making people disregard their relative shares: evidence from USA and India. *Royal Society Open Science, 4*(2), 160605.

Capraro, V., Everett, J. A., & Earp, B. D. (2019). Priming intuition disfavors instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology, 83*, 142-149.

Chan, Y. L., Gu, X., Ng, J. C. K., & Tse, C. S. (2016). Effects of dilemma type, language, and emotion arousal on utilitarian vs deontological choice to moral dilemmas in Chinese–English bilinguals. *Asian Journal of Social Psychology, 19*(1), 55-65.

Cesarini, D., Johannesson, M., Magnusson, P. K., & Wallace, B. (2012). The behavioral genetics of behavioral anomalies. *Management Science, 58*(1), 21-34.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*(2), 216-235.

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.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Dale, M. T., & Gawronski, B. (2023). Brains, trains, and ethical claims: Reassessing the normative implications of moral dilemma research. *Philosophical Psychology, 36*(1), 109-133.

De Lazari-Radek, K., & Singer, P. (2014). *The point of view of the universe: Sidgwick and contemporary ethics*. Oxford University Press.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy, 88*, 27-51.

Dennett, D. (2007). Philosophy as naive anthropology. In M. Bennett, D. C. Dennett, P. M. S. Hacker & J. R. & Searle (Eds.), *Neuroscience and Philosophy: Brain, Mind, and Language* (pp. 73-95).

Duke, A. A., & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition, 134*(1), 121-127.

Earp, B. D., McLoughlin, K. L., Monrad, J. T., Clark, M. S., & Crockett, M. J. (2021). How social relationships shape moral wrongness judgments. *Nature Communications, 12*(1), 1-13.

West-Eberhard, M. J. (1975). The evolution of social behavior by kin selection. *The Quarterly Review of Biology, 50*(1), 1-33.

Erceg, N., Galić, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection–testing convergent/divergent validity of two measures of cognitive reflection. *Judgment and Decision Making, 15*(5), 741-755.

Everett, J. A., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences, 24*(2), 124-134.

Falconer D. S. (1960). *Introduction to quantitative genetics*. Ronald Press.

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision making competence of older adults. *Psychology and Aging, 25*(2), 271–288

Foot, P. (1967). *The problem of abortion and the doctrine of double effect*. Oxford: Blackwell.

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

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin, 41*(5), 696-713.

Gawronski, B., & Beer, J. S. (2017). What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience, 12*(6), 626-632.

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2018). Effects of incidental emotions on moral dilemma judgments: An analysis using the CNI model. *Emotion, 18*(7), 989-1008.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology, 113*(3), 343-376.

Gleichgerrcht, E., Torralva, T., Rattazzi, A., Marenco, V., Roca, M., & Manes, F. (2013). Selective impairment of cognitive empathy for moral judgment in adults with high functioning autism. *Social Cognitive and Affective Neuroscience, 8*(7), 780-788.

Godfrey-Smith, P. (2009). *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029.

Greene, J. D. (2002). *The terrible, horrible, no good, very bad truth about morality and what to do about it*. Princeton University.

Greene, J. D. (2008) The secret joke of Kant's soul. In T. Nadelhoffer, E. Nahmias & S. Nichols (Eds.), *Moral Psychology* (pp. 359-372).

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. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D. (2015). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *The Law and Ethics of Human Rights, 9*(2), 141-172.

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

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

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

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998-1002.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Pearson College Division.

Hume, D. (1960). *Treatise of human nature* (L. A. Selby-Bigge, Ed.). Oxford University Press. (Original work published 1739).

Jackson, R. E., & Cormack, L. K. (2007). Evolved navigation theory and the descent illusion. *Perception & Psychophysics, 69*, 353-362.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513-541.

Johnson, E. D., Tubau, E., & De Neys, W. (2014). The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*(36), 2441-2446.

Jonas, E., McGregor, I., Klackl, J., Agroskin, D., Fritsche, I., Holbrook, C., ... & Quirin, M. (2014). Threat and defense: From anxiety to approach. In *Advances in Experimental Social Psychology* (Vol. 49, pp. 219-286). Academic Press.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience, 10*(5), 551-560.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131-164.

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

Kant, I. (1785). *Groundwork for the metaphysics of morals*.

Klenk, M. (2022). The influence of situational factors in sacrificial dilemmas on utilitarian moral judgments: A systematic review and meta-analysis. *Review of Philosophy and Psychology, 13*(3), 593-625.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience, 7*(6), 708-714.

Komter, A. (2010). The evolutionary origins of human generosity. *International Sociology, 25*(3), 443-464.

Krebs, D. L. (2008). Morality: An evolutionary account. *Perspectives on Psychological Science, 3*(3), 149-172.

Ladyman, J., & Ross, D. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press.

Lane, D., & Sulikowski, D. (2017). Bleeding-heart conservatives and hard-headed liberals: The dual processes of moral judgements. *Personality and Individual Differences, 115*, 30-34.

Li, S., Ding, D., Wu, Z., Yi, L., Lai, J., & Dang, L. (2020). Do high psychopaths care more about moral consequences than low psychopaths in Chinese culture? An exploration using the CNI model. *Healthcare, 8*(4), 505-518.

Li, Z., Xia, S., Wu, X., & Chen, Z. (2018). Analytical thinking style leads to more utilitarian moral judgments: An exploration with a process-dissociation approach. *Personality and Individual Differences, 13*1, 180-184.

Lieberman, D., Tooby, J., & Cosmides, L. (2003). Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270*(1517), 819-826.

Luke, D. M., Neumann, C. S., & Gawronski, B. (2022). Psychopathy and moral-dilemma judgment: an analysis using the four-factor model of psychopathy and the CNI model of moral decision-making. *Clinical Psychological Science, 10*(3), 553-569.

Maranges, H. M., Chen, S. K., & Conway, P. (2022). Insecure and insensitive: Avoidant and anxious attachment predict less concern for others in sacrificial moral dilemmas. *Personality and Individual Differences, 185*, 111274.

McGregor, I., Haji, R., Nash, K. A., & Teper, R. (2008). Religious zeal and the uncertain self. *Basic and Applied Social Psychology, 30*(2), 183-188.

McGregor, I., & Jordan, C. H. (2007). The mask of zeal: Low implicit self-esteem, threat, and defensive extremism. *Self and Identity, 6*(2-3), 223-237.

McGregor, I., Nail, P. R., Marigold, D. C., & Kang, S. (2005). Defensive pride and consensus: Strength in imaginary numbers. *Journal of Personality and Social Psychology, 89*, 978–996.

McGregor, I., Nash, K., & Prentice, M. (2010). Reactive approach motivation (RAM) for religion. *Journal of Personality and Social Psychology, 99*(1), 148-161.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103-115.

Mill, J. S. (1863). Utilitarianism. Collected Works of John Stuart Mill, Vol. 10. *Essays on Ethics, Religion, and Society*.

Miller, G. F. (2007). Sexual selection for moral virtues. *The Quarterly Review of Biology, 82*(2), 97-125.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology, 120*(2), 443-460.

Patrick, C. J. (2010). *Operationalizing the triarchic conceptualization of psychopathy: Preliminary description of brief scales for assessment of boldness, meanness, and disinhibition*. Unpublished manuscript. University of Minnesota. Minneapolis, MN.

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

Persson, B. N., & Kajonius, P. J. (2016). Empathy and universal values explicated by the empathy-altruism hypothesis. *The Journal of Social Psychology, 156*(6), 610-619.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*(3), 145-171.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16*(2), 331-348.

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science, 30*(9), 1389-1391.

Pölzler, T. (2018). How to measure moral realism. *Review of Philosophy and Psychology, 9*, 647-670.

Psillos, S. (1996). Scientific realism and the 'pessimistic induction'. *Philosophy of Science, 63*(3), 306-314.

Qian, Y., Takimoto, Y., Wang, L., & Yasumura, A. (2023). Exploring cultural and gender differences in moral judgment: A cross-cultural study based on the CNI model. *Current Psychology*, 1-11.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science, 27*(9), 1192-1206.

Rehman, S., & Dzionek-Kozłowska, J. (2020). The Chinese and American students and the trolley problem: A cross-cultural study. *Journal of Intercultural Communication, 20*(2), 31-41.

Rosas, A., Hannikainen, I., Lam, J., & Aguiar, F. (2023). Individual attitudes towards moral costs and benefits drive responses to moral dilemmas. *European Journal of Social Psychology*, 1-13.

Shariff, A. F. (2015). Does religion increase moral behavior? *Current Opinion in Psychology, 6*, 108-113.

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.

Sidgwick, H. (1981). *The methods of ethics* (7th ed.). Hackett Publishing. (Original work published 1907).

Simpson, D. (2021). *The dual-process model and moral dilemmas: Reflection does not drive self-sacrifice* [Master thesis, Georgia State University].

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs, 1*(3)., 229-243.

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making, 34*(3), 322-343.

Smith, K. B., Alford, J. R., Hibbing, J. R., Martin, N. G., & Hatemi, P. K. (2017). Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology. *American Journal of Political Science, 61*(2), 424-437.

Smith, K., & Hatemi, P. K. (2020). Are moral intuitions heritable? *Human Nature, 31*(4), 406-420.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision making, 13*(3), 260-267.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences, 23*(5), 701-717.

Street, S. (2017). Nothing 'really' matters, but that's not what matters. In *Does Anything Really Matter,* (pp. 121-148). Oxford University Press.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119*(3), 454-458.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning, 23*(3), 207-234

Szekely, R. D., Opre, A., & Miu, A. C. (2015). Religiosity enhances emotion and deontological choice in moral dilemmas. *Personality and Individual Differences, 79*, 104-109.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273-1296.

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

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275-1289.

Trémolière, B., De Neys, W., & Bonnefon, J. F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition, 124*(3), 379-384.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35-57.

Trivers, R. (2006). Reciprocal altruism: 30 years later. In P. Kappeler, & C van Schaik, *Cooperation in Primates and Humans: Mechanisms and Evolution* (pp. 67-83).

Elgin, C. (2008). Emotion and understanding. In G. Brun, U. Doguolu, & D. Kuenzlem (Eds.), Epistemology and Emotions (pp. 33-50).

Turner, J. H. (1997). The evolution of morality. *Critical Review, 11*(2), 211-232.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology, 12*(12), 1012-1015.

Wei, M., Russell, D. W., Mallinckrodt, B., & Vogel, D. L. (2007). The experiences in Close Relationship Scale (ECR)-Short Form: Reliability, validity, and factor structure. *Journal of Personality Assessment, 88,* 187-204.

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition, 126*(3), 364–372.

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica, 43*(2), 99-124.

Yang, Y., Liang, P., Lu, S., Li, K., & Zhong, N. (2009). The role of the DLPFC in inductive reasoning of MCI patients and normal aging: An fMRI study. *Science in China Series C: Life Sciences*, 52(8), 789-795.

Zheng, H., Lu, X., & Huang, D. (2018). tDCS over DLPFC leads to less utilitarian response in moral-personal judgment. *Frontiers in Neuroscience, 12*, 193.

APPENDIX

Cognitive Reflection Test

You will read three short problems. Please read them carefully and answer to the best of your ability.

1. If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?
2. Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?
3. Sally is making tea. Every hour, the concentration of tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration?

These questions were taken from Finucane and Gullion (2010).

Self-Sacrificial Dilemmas

Incongruent and Congruent Self-Sacrificial Dilemmas. Percentages indicate how many people endorse self-sacrifice. Studies 1, 2, 3, and 4 combined (N = 1,353). Study 5 is not included because the dilemmas were changed to include a relationship with the person.

Incongruent Dilemma Variant (Altruists would engage in self-sacrifice, Utilitarians would not)	Congruent Dilemma Variant (Altruists and Utilitarians would both engage in self-sacrifice)
Family Incongruent (55.7%) You are the oldest child, and you have one younger sibling. You are a math tutor, and your sibling is average at math. He has homework due tomorrow. If you spent the evening with them, he would probably increase his grade from a B+ to an A-. However, a movie comes to the theatres today, and you really wanted to see this movie on opening night. Is it appropriate for you to	Family Congruent (85.9%) You are the oldest child, and you have three younger siblings. You are a math tutor, and your siblings all struggle with math. They have homework due tomorrow. If you spent the evening with them, it would make a big difference, and they'd likely do much better. However, a movie comes to the theatres today, and you really wanted to see this movie on opening night. Is it appropriate for you to

watch the movie and let your sibling get a B+ instead of an A-?	watch the movie and let your siblings do poorly on their homework?
Piano Incongruent (15.3%) You are in the middle of an economic recession during the summer. Your family and your next door neighbour's family are both poor, but your family can at least entertain themselves by playing songs on the family piano. Your next door neighbours are bored. They started building a fort for their kids, but they can't afford more wood. The only way to give them wood is to give them the family piano to chop up so they can finish the fort. The piano is a family heirloom, but you have no other wood to give. Is it appropriate for you to keep the piano and let the other family not have a fort?	Piano Congruent (79.7%) You are in the middle of an economic depression during the winter. Your family and your next door neighbor's family are both poor, but your family at least has firewood to keep warm. Your next door neighbors have no firewood, and their children are at risk of freezing to death or at least catching a serious cold. The only way to give them wood is to give them the family piano to chop up and turn into firewood. The piano is a family heirloom, but you have no other wood to give. Is it appropriate for you to keep the piano and let the other family not have firewood?
Vacation Incongruent (17.6%) You are a skilled painter who is good at your craft. There is a construction project that is building sets for an elementary school play, and they have indicated that they need your expertise. This play will be a production of "Oliver Twist". However, the project is taking place over the December holidays, and you really wanted to use the money you had saved up to travel to Italy. Without you, the sets will still be built, but they won't be painted very well. Is it appropriate for you to go to Italy and let the elementary school production have poorly painted sets?	Vacation Congruent (78.0%) You are a skilled construction worker who is good at leading projects. There is a construction project that is building housing for the homeless, and they have indicated that they need your expertise. These would be the first low-income housing units in your city's history. However, the project is taking place over the December holidays, and you really wanted to use the vacation time you had saved up to travel to Italy. Without you, the project will likely not finish. Is it appropriate for you to go to Italy and let the housing project not finish?
Cliff Incongruent (29.3%) You are driving your car down a road on the side of a mountain. The road is narrow, has no barrier, and is beside a steep cliff. You are going fast, around 70 kilometers per hour. Your brakes fail, and your car starts to slide towards one ninety year old man. Since the road is so narrow, you're left with two choices. You can either run into the ninety year old man (which will certainly kill him) or you can veer off the cliff beside the road (which will certainly kill you). The man has	Cliff Congruent (74.4%) You are driving your car down a road on the side of a mountain. The road is narrow, has no barrier, and is beside a steep cliff. You are going fast, around 70 kilometers per hour. Your brakes fail, and your car starts to slide towards five pedestrians. Since the road is so narrow, you're left with two choices. You can either run into the pedestrians (which will certainly kill them) or you can veer off the cliff beside the road (which will certainly kill you). The pedestrians are all around your age.

only a few years left, but you have your whole life ahead of you. Is it appropriate for you to hit the man and save yourself?	Is it appropriate for you to hit the five pedestrians and save yourself?
Trolley Incongruent (44.4%) You are a construction worker and you have become stuck on a trolley track. A trolley is rushing towards you. There is, however, a switch that is within arm's reach. If you pull this switch, the trolley will go down a different path. Unfortunately, this path has another person stuck, who will die if you pull the switch. He is just an ordinary worker like you. If you do nothing, the trolley will kill you. Is it appropriate for you to save yourself and cause the trolley to kill one person?	Trolley Congruent (73.1%) You are a construction worker and you have become stuck on a trolley track. A trolley is rushing towards you. There is, however, a switch that is within arm's reach. If you pull this switch, the trolley will go down a different path. Unfortunately, this other track has five people stuck, all of whom will die if you pull the switch. They are all ordinary workers like you. If you do nothing, the trolley will kill you. Is it appropriate for you to save yourself and cause the trolley to kill five people?
War Incongruent (49.4%) It is World War II. German troops have taken over your village, and they have orders to put the remaining civilians in prison for a week as retribution. You and one of your friends have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. They are seconds from discovering the cellar. You are the only one who knows about a secret exit that leads to the woods. The only way to save your friend is to tell him about the exit, then distract the soldiers and thus be captured yourself. This would give enough time for him to escape. You are the only one who knows German, so you are the only one who could distract the soldiers effectively. You are also the closest to the backdoor, so you are the only one who could escape without a distraction. Is it appropriate for you to run away and let the other person go to prison?	War Congruent (71.2%) It is World War II. German troops have taken over your village, and they have orders to kill all remaining civilians. Some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. They are seconds from discovering the cellar. You are the only one who knows about a secret exit that leads to the woods. The only way to save the others is to tell them about the secret exit, then distract the soldiers and thus be captured and killed yourself. This would give enough time for the others to escape. You are the only one who knows German, so you are the only one who could distract the soldiers effectively. You are also the closest to the backdoor, so you are the only one who could escape without a distraction. Is it appropriate for you to run away and let the others die?
Grenade Incongruent (10.6%) You are a soldier who has been trained for special missions. You and 1 other soldier are rescuing a hostage from a terrorist group. A grenade gets thrown in the middle of your	Grenade Congruent (65.0%) You are a soldier who has been trained for special missions. You and 3 other soldiers are rescuing a hostage from a terrorist group. A grenade gets thrown in the middle of your

pair, but you're the only one who notices. You are also the only one standing near an open window. You only have a split second to act. You cannot throw it away in time, your only choices are to jump out the window (and save yourself) or jump on the grenade (and save your fellow soldier). You are a grenade expert, and you know that if you jump out the window, the other soldier will lose his leg. You also know that if you jump on the grenade to save him, you will die. Is it appropriate to jump out the window and let the other soldier lose his leg?	team, but you're the only one who notices. You are also the only one standing near an open window. You only have a split second to act. You cannot throw it away in time, your only choices are to jump out the window (and save yourself) or jump on the grenade (and save your fellow soldiers). You are a grenade expert, and you know that if you jump out the window, all 3 of the soldiers will die. You also know that if you jump on the grenade to save them, you will die. Is it appropriate to jump out the window and let the other 3 soldiers die?
Malaria Incongruent (7.0%) You work as a waiter at a local family restaurant. One day, you get a letter in the mail from the Home Depot telling you that a donation of $100 will help them polish their old floors. You are convinced that $100 will help them do this and consider sending in a donation. However, you have also been saving up for a new cell phone (your old one doesn't work well anymore). In your last paycheck, you finally made the last $100 of spending money you need for the phone. Is it appropriate for you to spend the 100$ on a new cell phone instead of donating to Home Depot? (modified from Kahane et al., 2015)	Malaria Congruent (60.1%) You work as a waiter at a local family restaurant. One day, you get a letter in the mail from the "Against Malaria Foundation," telling you that a donation of $100 will save a child's life in Africa. You are convinced that $100 will save a child's life and consider sending in a donation. However, you have also been saving up for a new cell phone (your old one doesn't work well anymore). In your last paycheck, you finally made the last $100 of spending money you need for the phone. Is it appropriate for you to spend the $100 on a new cell phone instead of saving a child's life? (modified from Kahane et al., 2015)
Tsunami Incongruent (12.8%) You are a college student who has been taking the bus for years. You always wanted to have your own car, but you could never afford it while paying for your education. One semester, your uncle passed away and left you $7000—enough money to purchase a decent used car (spending less is unwise as cheap cars soon require expensive repairs). On your way to the used car dealership, you hear a news broadcast about a Tsunami in South-East Asia that has destroyed a government owned art museum. There is a website where you could make a donation to help rebuild the museum. You contemplate donating the	Tsunami Congruent (42.4%) You are a college student who has been taking the bus for years. You always wanted to have your own car, but you could never afford it while paying for your education. One semester, your uncle passed away and left you $7000—enough money to purchase a decent used car (spending less is unwise as cheap cars soon require expensive repairs). On your way to the used car dealership, you hear a news broadcast about a Tsunami in South-East Asia that has left thousands dead, wounded, and homeless. There is a website where you could make a donation to help these victims. You contemplate donating the

$7000. You know that such a donation would allow people to enjoy art. However, you would have to continue living without a car. Is it appropriate for you to spend your uncle's $7000 on the car instead of promoting art? (modified from Kahane et al., 2015)	$7000. You know that such a donation would save the lives of several of the hospitalized tsunami victims. However, you would have to continue living without a car. Is it appropriate for you to spend your uncle's $7000 on the car instead of saving the tsunami victims? (modified from Kahane et al., 2015)
Organ Incongruent (8.2%) You live on a small island with a limited number of potential organ donors. The local clinic has one elderly person dying from a failing heart. However, you are there and have all healthy organs. The doctor says that if he kills you painlessly, he can take your heart and save the person from dying by giving him your heart. He will not do it without your permission. Is it appropriate for you to refuse, save yourself, and let the person die?	Organ Congruent (24.5%) You live on a small island with a limited number of potential organ donors. The local clinic has five people dying: one from a failing heart, one from a failing liver, one from failing kidneys, one from an intestinal problem, and one from failing lungs. However, you are there and have all healthy organs. The doctor says that if he kills you painlessly, he can take your organs to save the five from dying. He will not do it without your permission. Is it appropriate for you to refuse, save yourself and let the five die?

The tsunami and malaria dilemmas were modified from Kahane et al. (2015), the rest were internally generated.

Other-Sacrificial Dilemmas

Incongruent and Congruent Other-Sacrificial Dilemmas (taken from Conway & Gawronski, 2013). Percentages indicate how many people endorse other-sacrifice. Studies 1, 2, 3, and 4 combined (N = 1,353). Study 5 is not included because there were no other-sacrificial dilemmas included in that study.

Incongruent Dilemma Variant (Utilitarians would engage in other-sacrifice, Deontologists would not)	Congruent Dilemma Variant (Altruists and Utilitarians would both not engage in other-sacrifice)
Abortion Incongruent (95.3%) You are a surgeon. A young woman you know becomes pregnant, but her body reacts in an unusual fashion. She develops a severe case of preeclampsia, a dangerous syndrome that leads to rapid increases in blood pressure. The only treatment is to deliver the baby. Unless the baby is delivered soon, the mother	Abortion Congruent (84.4%) You are a surgeon. A young woman you know becomes pregnant, but she is not yet ready for children. She has not finished high school, has no income, and was abandoned by the father. If she has the baby now, she will be stuck as a single mother on welfare for the rest of her life. This will make things very

will die. However, the baby is too young to survive on its own. If it is delivered, it will die. So, although it is very difficult for her, the mother asks you to abort the baby. Is it appropriate for you to perform an abortion in order to save the mother's life?	hard on her and the baby. She thinks that it would be smarter to wait and have children later. So, although it is very difficult for her, she asks you to abort the baby. Is it appropriate for you to perform an abortion in order to let the mother live a better life?
Vaccine Incongruent (90.1%) You are a doctor in a health clinic overrun by patients with a serious disease. You just received a shipment of drugs that can cure the disease but the drugs have their own severe side-effects. If you administer the drugs to your patients, a small number will die from the side effects but most will live. If you do not, most will die from the disease. Is it appropriate for you to administer the drug to your patients?	Vaccine Congruent (35.3%) You are a doctor in a health clinic overrun by patients with the latest flu virus. You just received a shipment of drugs that can cure the flu but the drugs have their own severe side-effects. If you administer the drugs to your patients, a small number will die from the side effects but most will live. If you do not, most will continue to suffer from the effects of the flu virus for some time. Is it appropriate for you to administer the drug to your patients?
Car Incongruent (82.4%) You are driving through a busy city street when all of a sudden a young mother carrying a child trips and falls into the path of your vehicle. You are going too fast to brake in time; your only hope is to swerve out of the way. Unfortunately, the only place you can swerve is currently occupied by a little old lady. If you swerve to avoid the young mother and baby, you will seriously injure or kill the old lady. Is it appropriate to swerve and hit the old lady in order to avoid the young mother and child?	Car Congruent (11.6%) You are driving through a busy city street when all of a sudden a young mother carrying a child trips and falls into the path of your vehicle. You are going too fast to brake in time; your only hope is to swerve out of the way. Unfortunately, the only place you can swerve is currently occupied by a group of children on their way to elementary school. If you swerve to avoid the young mother and baby, you will seriously injure or kill several of them. Is it appropriate to swerve and hit the schoolchildren in order to avoid the young mother and child?
Animal Incongruent (71.1%) You have been hired by a pharmaceutical company to conduct research on their products. Since products must be fit for human use, they are first tried out on animals. Your job is to find out the effects various chemicals have on rats, pigeons, rabbits, and monkeys. Most chemicals have only minor effects on the animals, but some cause them discomfort or even permanent damage. The chemicals you are researching are slated to form part of a new AIDS drug cocktail that	Animal Congruent (27.1%) You have been hired by a pharmaceutical company to conduct research on their products. Since products must be fit for human use, they are first tried out on animals. Your job is to find out the effects various chemicals have on rats, pigeons, rabbits, and monkeys. Most chemicals have only minor effects on the animals, but some cause them discomfort or even permanent damage. The chemicals you are researching are slated to form part of a new acne facial cleanser that

will give new hope to millions of AIDS victims around the world. You anticipate saving many lives with the chemicals. Is it appropriate to test these chemicals on animals?	will give new hope to people with pimples and greasy skin. You anticipate making many people feel better about their appearance with the chemicals. Is it appropriate to test these chemicals on animals?
Torture Incongruent (71.2%) You are a police officer, and have recently caught a criminal you have been hunting for some time. He is allegedly responsible for rigging a series of explosive devices: some that have already gone off and some that have yet to detonate. He places explosives outside city cafes and sets them to go off at a time when people are drinking coffee on the patios. In this manner, he has injured many people and might injure many more. Now that the criminal is in custody, you want to know where the unexploded bombs are so you can defuse them. He refuses to talk, so you decide to use "aggressive interrogation techniques" like holding his head under water and beating him. Is it appropriate for you to use "aggressive interrogation techniques" in order to find and defuse the unexploded bombs?	Torture Congruent (37.8%) You are a police officer, and have recently caught a criminal you have been hunting for some time. He is allegedly responsible for rigging a series of explosive devices: some that have already gone off and some that have yet to detonate. He places explosives outside city cafes and sets them to go off at a time when no one is around. His explosives are inside paint cans so that they spray nearby objects with paint. In this manner, he has sprayed many cafes with paint and might spray many more. Now that the criminal is in custody, you want to know where the unexploded bombs are so you can defuse them. He refuses to talk, so you decide to use "aggressive interrogation techniques" like holding his head under water and beating him. Is it appropriate for you to use "aggressive interrogation techniques" in order to find and defuse the unexploded bombs?
Time Incongruent (69.9%) You find a time machine and travel back to the year 1920. While checking into a hotel, you meet a young Austrian artist and veteran of the First World War. You realize this is Adolf Hitler before his rise to power in Nazi Germany. He is staying in the hotel room next to yours and the doors are not locked. It would be easy to simply smother him with a pillow in his sleep and disappear, stopping the Second World War and the Nazi party before they even start. However, he has not committed any crimes yet and it seems wrong to hurt an innocent person. Is it appropriate for you to kill an innocent young Hitler in order to prevent the Second World War?	Time Congruent (45.2%) You find a time machine and travel back to the year 1920. While checking into a hotel, you meet a young petty criminal. You realize this is George Brackman, a man who later on abducted a child and held her for a week until her family paid him some ransom money. He is staying in the hotel room next to yours and the doors are not locked. It would be easy to simply smother him with a pillow in his sleep and disappear, stopping the abduction and ransom demands before they even start. However, he has not committed any crimes yet and it seems wrong to hurt an innocent person. Is it appropriate for you to kill George Brackman in order to prevent him from taking a child hostage?

Crying Baby Incongruent (68.7%) It is war time. Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. A baby with no parents begins to cry loudly. You cover her mouth to block the sound. If you remove your hand from the baby's mouth her crying will summon the attention of the soldiers who will kill you and the others hiding out in the cellar. To save yourself and the others you must smother the child to death. Is it appropriate for you to smother the child in order to save yourself and the other townspeople from being killed?	Crying Baby Congruent (42.1%) It is war time. Enemy soldiers have taken over your village. They have orders to capture all remaining civilians to make them work quarrying stone in a mine. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. A baby with no parents begins to cry loudly. You cover her mouth to block the sound. If you remove your hand from her mouth the crying will summon the attention of the soldiers who will capture you and the others hiding out in the cellar. To save yourself and the others from laboring in the mine you must smother the child to death. Is it appropriate for you to smother the child in order to save yourself and the other townspeople from being captured?
Border Incongruent (45.2%) You are a soldier guarding a border checkpoint between your nation and one troubled by insurgent violence. You notice a young man in a cheap car approaching the checkpoint with a determined look on his face. You suspect he means to bomb the checkpoint, killing all the soldiers inside. He is rapidly approaching your station. Is it appropriate for you to shoot and kill the approaching man?	Border Congruent (11.6%) You are a soldier guarding a border checkpoint between your nation and one troubled by insurgent violence. You notice a young man in a cheap car approaching the checkpoint with a determined look on his face. You suspect he means to cross the border in order to work illegally inside your country. He is rapidly approaching your station. Is it appropriate for you to shoot and kill the approaching man?
Relationship Incongruent (22.5%) You are divorced now and your former spouse lives far away. You don't expect to see your former spouse for a very long time. However, you have recently started dating a new person that is positively wonderful. Your new partner is head over heels in love for the first time and you feel equally strong about your partner. You want to do everything in your power to keep the relationship progressing smoothly to what you believe will be a second marriage. Unfortunately, your new partner has told you many times that dating someone who is divorced is totally unacceptable. It is	Relationship Congruent (3.2%) You are divorced now and your former spouse lives far away. You don't expect to see your former spouse for a very long time. However, you have recently started dating a new person that is positively wonderful. Your new partner is head over heels in love for the first time and you feel equally strong about your partner. You want to do everything in your power to keep the relationship progressing smoothly to what you believe will be a second marriage. Unfortunately, your ex-spouse gave you a very serious and incurable Sexually Transmitted Disease (STD). Your partner

expressly forbidden in your partner's religion. Your partner doesn't know about your ex-spouse. You consider simply pretending that you have never been divorced so that your current relationship can continue. Is it appropriate for you to leave your new partner in the dark about your previous relationship in order to keep her/him happy and the relationship alive?	does not know about it. You consider simply pretending that you don't know about the STD so that your current relationship can continue. Is it appropriate for you to leave your new partner in the dark about your STD in order to keep her/him happy and the relationship alive?
Hard Times Incongruent (11.6%) You are the head of a poor household in a developing country. Your crops have failed for the second year in a row, and it appears that you have no way to feed your family. Your sons, ages eight and ten, are too young to go off to the city where there are jobs, but your daughter could fare better. You know a man from your village who lives in the city and who makes sexually explicit films featuring girls such as your daughter. In front of your daughter, he tells you that in one year of working in his studio your daughter could earn enough money to keep your family fed for several growing seasons. Is it appropriate for you to employ your daughter in the pornography industry in order to feed your family?	Hard Times Congruent (2.7%) You are the head of a poor household in a developing country. Your sons, ages eight and ten, work on the family farm every day in order to teach them the value of hard work and providing for others. Your daughter, however, cannot work on the farm due to the heavy lifting required. You worry that she will not learn the value of money if she does not work. You know a man from your village who lives in the city and who makes sexually explicit films featuring girls such as your daughter. In front of your daughter, he tells you that in one year of working in his studio your daughter could earn enough money to keep your family fed for several growing seasons. Is it appropriate for you to employ your daughter in the pornography industry in order to teach her the value of money?

Reason/Emotion Manipulation

Reason condition stimulus before the dilemmas:

Sometimes people make decisions by using logic and relying on their reason. Other times, people make decisions by using feeling and relying on their emotion.

Many people believe that reason leads to good decision-making. When we use logic, rather than feelings, we make rationally satisfying decisions. Please deal with the following dilemmas by relying on reason, rather than emotion.

In light of these instructions, which will you be relying on when dealing with the dilemmas?

Reason

Emotion

Reason condition stimulus halfway through the dilemmas:

REMINDER: Please continue to rely on reason, not emotion.

Emotion condition stimulus before the dilemmas:

Sometimes people make decisions by using feeling and relying on their emotion. Other times, people make decisions by using logic and relying on their reason.

Many people believe that emotion leads to good decision-making. When we use feelings, rather than logic, we make emotionally satisfying decisions. Please deal with the following dilemmas by relying on emotion, rather than reason.

In light of these instructions, which will you be relying on when dealing with the dilemmas?

Reason

Emotion

Emotion condition stimulus halfway through the dilemmas:

REMINDER: Please continue to rely on emotion, not reason.

Other Scales

Oxford Utilitarianism Scale

From strongly disagree to strongly agree (seven-point Likert scale)

1. "If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice." (Impartial beneficence)

2. “It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.” (Instrumental harm)
3. “From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don’t need two kidneys to survive, but really only one to be healthy.” (Impartial beneficence)
4. “If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.” (Instrumental harm)
5. “From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.” (Impartial beneficence)
6. “It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.” (Instrumental harm)
7. “It is just as wrong to fail to help someone as it is to actively harm them yourself.” (Impartial beneficence)
8. “Sometimes it is morally necessary for innocent people to die as collateral damage—if more people are saved overall.” (Instrumental harm)
9. “It is morally wrong to keep money that one doesn’t really need if one can donate it to causes that provide effective help to those who will benefit a great deal.” (Impartial beneficence)

These questions were taken from Kahane et al. (2018).

Experiences in Close Relationships Attachment Scale- Short Form

From strongly disagree to strongly agree (seven-point Likert scale)

1. It helps to turn to my romantic partner in times of need.
2. I need a lot of reassurance that I am loved by my partner.
3. I want to get close to my partner, but I keep pulling back.
4. I find that my partner doesn't l want to get as close as I would like.
5. I turn to my partner for many things, including comfort and reassurance.
6. My desire to be very close sometimes scares people away.
7. I try to avoid getting too close to my partner.
8. I don't worry about being abandoned.
9. I usually discuss my problems and concerns with my partner.
10. I get frustrated if my romantic partner is not available when I need them.
11. I am nervous when my partner gets too close to me.
12. I worry that a romantic partner won't care about me as much as I care about them.

These questions were taken from Wei et al. (2007).

Triarchic Psychopathy Measure

From True to Somewhat true to Somewhat false to False.

The empathy facet consists of items 1, 8, 11, 20, 29, 33, 36, 48, 52, and 55.

1. I'm optimistic more often than not.
2. How other people feel is important to me.

3. I often act on immediate needs.
4. I have no strong desire to parachute out of an airplane.
5. I've often missed things I promised to attend.
6. I would enjoy being in a high-speed chase.
7. I am well-equipped to deal with stress.
8. I don't mind if someone I dislike gets hurt.
9. My impulsive decisions have caused problems with loved ones.
10. I get scared easily.
11. I sympathize with others' problems.
12. I have missed work without bothering to call in.
13. I'm a born leader.
14. I enjoy a good physical fight.
15. I jump into things without thinking.
16. I have a hard time making things turn out the way I want.
17. I return insults.
18. I've gotten in trouble because I missed too much school.
19. I have a knack for influencing people.
20. It doesn't bother me to see someone else in pain.
21. I have good control over myself.
22. I function well in new situations, even when unprepared.
23. I enjoy pushing people around sometimes.
24. I have taken money from someone's purse or wallet without asking.
25. I don't think of myself as talented.

26. I taunt people just to stir things up.

27. People often abuse my trust.

28. I'm afraid of far fewer things than most people.

29. I don't see any point in worrying if what I do hurts someone else.

30. I keep appointments I make.

31. I often get bored quickly and lose interest.

32. I can get over things that would traumatize others.

33. I am sensitive to the feelings of others.

34. I have conned people to get money from them.

35. It worries me to go into an unfamiliar situation without knowing all the details.

36. I don't have much sympathy for people.

37. I get in trouble for not considering the consequences of my actions.

38. I can convince people to do what I want.

39. For me, honesty is the best policy.

40. I've injured people to see them in pain.

41. I don't like to take the lead in groups.

42. I sometimes insult people on purpose to get a reaction from them.

43. I have taken items from a store without paying for them.

44. It's easy to embarrass me.

45. Things are more fun if a little danger is involved.

46. I have a hard time waiting patiently for things I want.

47. I stay away from physical danger as much as I can.

48. I don't care much if what I do hurts others.

49. I have lost a friend because of irresponsible things I've done.
50. I don't stack up well against most others.
51. Others have told me they are concerned about my lack of self-control.
52. It's easy for me to relate to other people's emotions.
53. I have robbed someone.
54. I never worry about making a fool of myself with others.
55. It doesn't bother me when people around me are hurting.
56. I have had problems at work because I was irresponsible.
57. I'm not very good at influencing people.
58. I have stolen something out of a vehicle.

These questions were taken from Patrick (2010).

Religious Zeal Scale

From strongly disagree to strongly agree (five-point Likert scale)

1. I am confident in my religious beliefs.
2. I aspire to live and act according to my religious beliefs.
3. My religious beliefs are grounded in objective truth.
4. Most people would agree with my religious beliefs if they took the time to understand it rather than just relying on stereotypes about it.
5. If my religious beliefs were being publicly criticized I would argue to defend them.
6. I would support a war that defended my religious beliefs.
7. If I really had to, I would give my life for my religious beliefs.

8. In my heart I believe that my religious beliefs are more correct than others.
9. It is wise to keep a wary distance from people who distract me from living according to my religious beliefs.
10. In the end, those who oppress my religious beliefs will suffer for their ignorance.
11. If everyone followed my religious beliefs, the world would be a much better place.
12. Harmful misinformation is too often spread about my religious beliefs by ignorant people.
13. If necessary, I would endure much pain and suffering to stay true to my religious beliefs.
14. I will do whatever is necessary to help my religious beliefs prosper in society.
15. If I was sincerely convinced that God wanted me to do something extreme, I would do it.
16. Today society is in desperate need of the wisdom of my religious beliefs.
17. I believe that a powerful God or Godlike force shapes human destiny.
18. Most important events in our world are guided by, and to some extent controlled by, the will of God or a Godlike force.
19. My strongest relationships are with those who have the same religious beliefs as I do.

These questions were taken from McGregor et al. (2008).

Supplemental Analyses- Study 2

In addition to my hypothesis that 1) reason drives utilitarian judgments, both other-sacrificial and self-sacrificial, I also made the following two hypotheses: 2) emotion drives deontological judgments, 3) emotion drives purely altruistic judgments. I derived four additional predictions from hypotheses 2 and 3, and also made four miscellaneous predictions. Five out of these eight additional predictions were confirmed, but they were not relevant to the argument I was making in the manuscript, so they are included here.

Pre-registered prediction 5 (derived from hypothesis 2). I predicted that the CRT would not be positively correlated with the D parameter. The CRT did not correlate with the D parameter, $r = .024$, $p = .691$. The prediction was confirmed.

Pre-registered prediction 6 (derived from hypothesis 3). I predicted that the CRT would not be positively correlated with the A parameter. It correlated negatively with the A parameter, $r = -.120$, $p = .050$. The prediction was not confirmed since there was a significant correlation. But since the hypothesis was that emotion drives Altruistic responses, a negative correlation with cognitive reflection is still consistent with the hypothesis.

Pre-registered prediction 7-8 (derived from hypothesis 2 and 3). I predicted that the emotion prime would increase the deontology and altruistic parameters relative to the reason prime. The reason/emotion manipulation did not have a significant effect on the A parameter or the D parameter (both ps $> .625$). These predictions were not confirmed.

Pre-registered prediction 9. I predicted that the other-sacrificial utilitarian parameter would be positively correlated with the self-sacrificial utilitarian parameter. The SU parameter

and OU parameter were positively correlated with each other, $r = .360$, $p < .001$. This prediction was confirmed.

Pre-registered prediction 10. I predicted that the deontology parameter would be positively correlated with the altruism parameter. The D parameter and A parameter were positively correlated with each other, $r = .229$, $p < .001$. This prediction was confirmed.

Pre-registered prediction 11. I Predicted that the other-sacrificial utilitarian parameter would not be correlated with the deontology parameter. They were not correlated, $r = .027$, $p = .659$. This prediction was confirmed.

Pre-registered prediction 12. I predicted that the self-sacrificial utilitarian parameter would not be correlated with the altruism parameter. They were not correlated, $r = .099$, $p = .105$. This prediction was confirmed.