8-10-2021

The Dual-Process Model and Moral Dilemmas: Reflection Does Not Drive Self-Sacrifice

David Simpson

The Dual-Process Model and Moral Dilemmas: Reflection Does Not Drive Self-Sacrifice

by

David Simpson

Under the Direction of Eddy Nahmias, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2021

ABSTRACT

Greene uses evidence from psychology and neuroscience to argue that manual mode (slow, deliberative thinking) is conducive to utilitarian judgments. He further argues that these data, in conjunction with philosophical premises, lend normative support to utilitarianism. After defending Greene's philosophical premises against critics, I contend that the current state of the evidence suggests that manual mode does *not* drive utilitarian responses to moral dilemmas involving self-sacrifice. I performed an experiment which replicated the positive association between cognitive reflection test (CRT) scores (which measure reliance on manual mode) and utilitarian responses to dilemmas that involved sacrificing the interests of others. However, I did not find a positive association between CRT scores and self-sacrificial utilitarian responses. The lack of a connection between manual mode and self-sacrifice presents a problem for Greene's argument that manual mode drives utilitarianism in general. Prima facie, my results indicate that reflection only drives other-sacrificial utilitarian judgments, not self-sacrificial ones. Greene is left without a basis to say (as he does) that cognitive science lends support to the normative conclusion that we ought to engage in utilitarian self-sacrifice by, for example, giving more to charity. I conclude by discussing other implications of my data for Greene's argument, and outlining directions for future research.

INDEX WORDS: Dual-process model, Moral psychology, Utilitarianism, Deontology, Reason, Emotion

The Dual-Process Model and Moral Dilemmas: Reflection Does Not Drive Self-Sacrifice

by

David Simpson

Committee Chair: Eddy Nahmias

Committee: Eyal Aharoni

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2021

DEDICATION

This thesis is dedicated to my Mom. She is the best listener, best reviewer, and best counselor I have ever met. I know how lucky I am to have her. I also dedicate this thesis to my Dad: thank you for being a good role model and for always making me laugh. As Edgar Albert Guest said, “Only a dad, but he gives his all to smooth the way for his children small. This is the line for him I pen: only a dad, *but the best of men*”.

ACKNOWLEDGEMENTS

I would like to thank Eddy Nahmias and Eyal Aharoni for their helpful comments on previous drafts, as well as their input regarding the design of my study. It has been an honor to do experimental work under their supervision. Connor Kianpour and Steven Falco also gave helpful comments on previous drafts. In conversations inside and outside of class, they have challenged many of my philosophical views, which I view as an enormous gift. I will add that there are many writers, dead and living, without whom I would not be the thinker I am today. Karl Popper, Daniel Dennett, Steven Pinker, Ludwig Wittgenstein, and Joshua Greene are all giants on whose shoulders I am proud to stand. Finally, I want to thank the many teachers who have encouraged/challenged me over the years. I hope that I will one day inspire students as much as you all inspired me.

TABLE OF CONTENTS

1 BACKGROUND

In this section I explain Joshua Greene's dual-process model of moral cognition. On the basis of evidence from psychology and neuroscience, he argues that automatic mode (quick, intuitive, and emotional thinking) causes deontological judgments, and that manual mode (slow, deliberative, and reflective thinking) causes utilitarian judgments. He further argues that this model has philosophical implications because it casts doubt on the reliability of deontological judgments, especially when they are given in response to unfamiliar moral problems. He suggests that for such problems, utilitarianism should be used as a basis for resolving disagreement.

1.1 Greene's Dual-Process Model: Manual Mode Causes Utilitarian Judgments

The trolley dilemma is perhaps the most famous thought experiment in all of philosophy. It was invented by Foot (1978) and modified by Thomson (1985). There are two versions which have come to be known as the switch dilemma and the footbridge dilemma. In the switch dilemma, the question is whether it is permissible to pull a switch that will change a trolley from being on a track which will lead to five deaths to being on a track which will lead to one death. In the footbridge dilemma, the question is whether it is permissible to push a large man off a footbridge in front of a trolley to save five others. Intuitively, it seems that killing one to save five is permissible in the first case, but not in the second. Most ordinary people have this intuition (Petrinovich & O'Neill, 1996), as do the aforementioned philosophers. Foot (1978) argues that this difference in intuition is justified. She cites the doctrine of double effect, the idea that harm in service of a good end is justified so long as the harm is a side effect of the intended action. In the switch case, the killing is justified because it is a side effect, albeit foreseen. But in the footbridge case, the killing is not a side effect; it is the intended means to bring about the outcome of saving five people. Hence, there is a morally relevant difference between the two cases. The implication of the doctrine of double effect is in line with our intuition that the cases are morally different.

Greene (2008) proposed a different approach to this pair of dilemmas. Rather than try to determine what philosophical principles *justify* the difference in responses (or not), he tries to discover empirically what psychological mechanisms *cause* the difference[1]. He and other researchers have used the trolley dilemmas and a host of other dilemmas involving the option to commit up close and personal harm for the sake of benefiting a larger number of people (e.g. whether to smother a baby to save lives, Suter & Hertwig, 2011). Henceforth, such dilemmas will be called "other-sacrificial dilemmas" because they involve a decision to harm someone else. Greene (2015) has used research involving other-sacrificial dilemmas to measure people's utilitarian tendencies. Utilitarianism, according to the mainstream interpretation, is the view that the morality of an action is determined by the extent to which it maximizes happiness impartially (De Lazari- Radek & Singer, 2014). Deontology, by contrast, is the view that the morality of an action is determined by the intrinsic nature of the action, especially whether it is consistent with moral rules (Gawronski & Beer, 2017). This definition means that a deontologist would condemn an action that violates a moral rule, even if it produces better consequences. Conversely, a utilitarian would view an action conducive to beneficial consequences as morally preferable, even if it violates a moral rule.

[1] He proposes this approach as a first step. As we will see later, he does go on to draw normative implications from his empirical model. Crucially, he does this in conjunction with normative premises, and thus does not violate the is-ought gap.

Utilitarianism and deontology come in different versions and each has a complex history, but Greene (2008) gives narrow operational definitions. His operational definition of utilitarianism is approval of using up close and personal harm towards a small number to save a greater number, while his operational definition of deontology is disapproval of up close and personal notwithstanding the good consequences. Greene does not require that the participants endorse (implicitly or explicitly) these moral theories, he merely means that these judgments are “utilitarian” and “deontological” in the sense that the judgments are more easily justified by those theories. These definitions constitute attempts by Greene (2008) to single out distinct types of ordinary moral reasoning or “psychological natural kind(s)” with “functional roles” (p. 360). It has been pointed out that these dilemmas do not predict real-world behavior (Bostyn et al., 2018), but Plunkett and Greene (2019) respond that this is not their role, and ask us to consider visual illusions as an analogy. Visual illusions (such as the Muller-Lyer illusion) do not predict individual variation in perceptual capacities; they experimentally separate different internal perceptual processes. Greene claims that other-sacrificial dilemmas serve a similar role in that they experimentally separate processes that drive what Greene describes as utilitarian and deontological judgments.

Greene (2015) further postulates that humans have a dual-process system for dealing with moral problems: automatic mode and manual mode.[2] Automatic mode is quick and intuitive, while manual mode is slow and deliberative. According to Greene’s taxonomy, manual mode is rooted in slow, controlled cognition, whereas automatic mode is rooted in emotion. The crucial next step in Greene’s argument is to show that automatic mode tends to produce deontological judgments (as a first response), and that manual mode tends to produce (upon reflection)

[2] This is akin to the popular distinction between system 1 and system 2 (Stanovich and West, 2000).

utilitarian responses. There are multiple convergent lines of evidence for the connection between manual mode and utilitarian judgments in other-sacrificial dilemmas (for summaries of the literature, see Greene, 2015; Patil et al., 2018). I will review some of the most compelling evidence here. One widely used measure of reliance on manual mode (or system 2) is the cognitive reflection test or CRT (Frederick, 2005). One item on one version of the test is the famous ball and bat question: "a baseball bat and a ball cost $1.10 together, and the bat costs $1.00 more than the ball, how much does the ball cost". People have a quick, intuitive response (from automatic mode), which is "10 cents". If they slow down and do the arithmetic (manual mode), they get the correct answer, which is "5 cents". Performance on this test is strongly predictive of reliance on algorithmic thinking in other contexts (Toplak et al., 2011), and performance on the test is also strongly correlated with giving utilitarian answers to other-sacrificial dilemmas (Paxton et al., 2012, Conway & Byrd, 2019).

One brain region that is widely thought to be associated with controlled cognition (or manual mode) is the dorsolateral prefrontal cortex or DLPFC (Glenn et al., 2009, Yang et al., 2009). Using functional magnetic resonance imaging (or fMRI), Greene et al. (2004) found that there was a correlation between DLPFC activity and utilitarian responses to other-sacrificial dilemmas. Furthermore, when the DLPFC is interfered with using transcranial direct current stimulation (tDCS), participants are *less* likely to give utilitarian answers (Zheng et al., 2018). One can also reduce dependence on manual mode by giving participants a mentally difficult task that they must perform while reading the moral dilemmas. This is called increasing cognitive load, and it has been shown to reduce people's dependence on controlled reasoning (Gilbert, Tafarodi & Malone, 1993; Johnson, Tubau, & De Neys, 2014). Several studies find that participants are less likely to give a utilitarian response when cognitive load is increased

(Trémolière, B., De Neys, W., & Bonnefon, 2012; Conway & Gawronski, 2013; Białek & De Neys, 2017). Similarly, when participants are given fewer time constraints and more encouragement to be deliberative (which presumably increases the influence of manual mode), they are more likely to give utilitarian answers (Suter & Hertwig, 2011).

Greene (2015) uses similar lines of evidence to argue that deontological (or non-utilitarian) judgments are caused by automatic mode. Deontological responses to other-sacrificial dilemmas positively correlate with activity in the amygdala, an emotion center of the brain (Shenhav & Greene, 2014), and are more likely to occur when empathy is induced (Conway & Gawronski, 2013). With regard to all of the arguments given by philosophers in defense of deontological judgments (like the aforementioned doctrine of double effect), Greene (2008) contends that they are merely post-hoc rationalizations.

The claim that arguments offered in favor of deontological responses are post-hoc rationalizations is similar to Haidt's (2001) rejection of rationalist conceptions of moral cognition. Haidt argued that moral cognition in general is driven by automatic mode, and that arguments offered in favor of moral judgments are post-hoc rationalizations. In a famous study, he presented subjects with viscerally repulsive but harmless actions, such as one-time sex with a sibling using birth control which is stipulated not to result in relational damage (Haidt, Bjorklund, & Murphy, 2000). Most people judged these harmless taboo violations to be wrong, and when pressed were able to give reasons in favor of their judgment. However, when these reasons were refuted, people maintained their initial judgment. This finding constitutes evidence that the arguments given were rationalizations, and that the initial judgment was not rooted in reasoning. Haidt (2001) thought that his findings applied to all moral reasoning, and that moral cognition involves an initial intuitive reaction which is then justified in a post-hoc manner by reasoning. By contrast, Greene's contention is that this description only applies to deontological judgments and not utilitarian ones. Contra Haidt, utilitarian judgments are an exception: they are arrived at through the use of slow reflective reason.

1.2 The Philosophical Significance of Greene's Dual-Process Model

In addition to arguing that manual mode tends to facilitate utilitarian judgments, Greene (2015) has argued that the outputs of manual mode are more reliable. Greene (2013) is agnostic about moral realism, and so no strong meta-ethical commitments should be read into the term "reliable". To a first approximation, Greene appears to mean something like "ability to achieve consensus" when he says "reliable". This can be seen in his appeal to the common currency argument. The common currency argument is that while human groups disagree about many moral starting points, we all agree that maximizing happiness is good (all else being equal). He concludes that we should rely on the principle of happiness maximization as a way to adjudicate disagreement. Given that (according to Greene) the worlds' cultures do not generally agree on other moral axioms, we should use this axiom as our collective basis (i.e. our common currency) for adjudicating moral disputes (Greene, 2013). This view presupposes that consensus-building capacity is the main metric by which we should judge the reliability of a moral judgment.

Greene defends the reliability of manual mode through an appeal to what he calls "the no 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 automatic mode were reliable in dealing with an unfamiliar moral problem. Since the footbridge dilemma centers 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 (2015) 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 manual mode 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.

If it is true that a) manual mode is more reliable than automatic mode, and b) manual mode causes utilitarian judgments, then it seems to follow that we should put more trust in utilitarian judgments than deontological judgments, at least for unfamiliar moral dilemmas. Lott (2016) objects, saying that the psychological origin of a moral judgment does not determine whether it is justified. For example, Lott goes over the example of incest. People's moral condemnation of incest appears to be caused by emotionally rooted psychological mechanisms, whose evolutionary function is to prevent deleterious genetic effects. Whether or not to commit incest when these deleterious effects can be prevented with birth control is an unfamiliar moral problem which (according to Greene) should be solved with manual mode. This is an example of a harmless taboo violation. Furthermore, the fact that our initial aversion to this harmless taboo violation is caused by automatic mode casts doubt on that reaction, according to Greene (2015).

But Lott points out that we might still upon reflection accept the emotionally rooted anti-incest judgment for various other reasons. In other words, the mere fact that the initial judgment was caused by automatic mode is not sufficient grounds for rejecting the judgment.

A different argument that Greene has appealed to in favor of the reliability of manual mode-driven utilitarian judgments is what has come to be known as the argument from morally irrelevant factors (e.g. by Paulo, 2019). The argument is that automatic mode (but not manual mode) is driven by factors that both utilitarians and non-utilitarians deem irrelevant. As mentioned, philosophers have contended that the reason that it is impermissible to kill one to save five in the footbridge case (but not the switch case) is that a person is being used as a means (e.g. Foot, 1978). However, if the footbridge case is adjusted so that you have to pull a switch that makes the large person fall through a trapdoor to stop the trolley, people are as likely to approve as they are in the switch case (Greene, 2015). The person is still being used as a means, but there is no up close and personal harm. Clearly, the fact that the person is being used as a means is not what is driving people's condemnation. What psychologically drives their condemnation is the fact that the action involves up close personal contact (i.e. low physical distance), which is not something that moral philosophers would appeal to because it seems morally irrelevant. On the basis of this evidence, Conway et al. (2018) contend that the sophisticated principles (like the doctrine of double effect) given as a defense of deontological intuitions stand a good chance of being post-hoc rationalizations. Lest the reader think that philosophers are immune from engaging in post-hoc rationalizations of this sort, Schwitzgebel and Cushman (2015) used moral philosophers as subjects, and found that they are also susceptible to the influence of irrational factors (like the order in which dilemmas are presented) when evaluating other-sacrificial dilemmas.

Even if evidence from cognitive science can tell us that deontological judgments tend to be caused by automatic mode, Lott (2016) is still right that this evidence still must be conjoined with good philosophical reasons to reject the deontological judgments in question. When the argument from irrelevant factors is used against any particular deontological judgment, philosophical premises are required. Most notably, the factors (such as physical distance) discovered to cause the judgment must in fact be morally irrelevant in order for the judgment to be rejected. The claim that physical distance is morally irrelevant is a philosophical claim that cannot be empirically demonstrated. The value of cognitive science is that it can empirically show us that an argument in favor of a judgment is a post-hoc rationalization. This is compatible with the argument being valid (post-hoc rationalizations can be correct), but discovering that the argument psychologically emerged as a rationalization should make us to be vigilant in looking for flaws in the argument.

I will spell this point out by applying it to a specific example. Suppose that people generally consider a movie to be aesthetically good, and that the justification they offer is that the movie had sophisticated character development. Further suppose that a psychologist experimentally demonstrated that the fact that the movie has a red title is the factor that makes people like the movie. When people are shown the movie, but the title is made to be blue, they no longer say that they like it. This is a strong indication that when people say, "the movie had sophisticated character development", they are actually engaging in post-hoc rationalization. This justification is thrown into doubt, and we should feel prompted to look closely at it. However, it still could turn out to be true upon reflection that the movie does have good character development. The same point applies in the ethical case. The fact that physical distance (or lack thereof) is what in fact causally drives people to make deontological judgments in the footbridge case does cast doubt on the justifications offered by deontological philosophers. But we might still decide upon reflection that the doctrine of double effect (a principle used to defend the wrongfulness of sacrificing someone in the footbridge case) is right, just as we might decide upon reflection that the movie really does have good character development.

This analogy helps defend Greene's argument (or at least a conservative version of it) against philosophical objections. Empirical evidence can show that an argument offered in favor of a moral judgment (say, a deontological one) is a post-hoc rationalization. The fact that an argument can be empirically demonstrated to be a post-hoc rationalization does cast some doubt upon its validity, as is seen in the movie analogy. This should lead philosophers to be vigilant in looking for flaws in the argument. However, this is conservative version of Greene's argument because the data cannot by themselves show that an argument offered in defense of a moral judgment is a bad argument. A post-hoc rationalization still can be discovered to be valid upon reflection. In the end, the specific flaws in the post-hoc rationalization must be pointed out. In the next section, I will leave aside these philosophical concerns and focus on Greene's empirical premise, namely that manual mode causes utilitarian judgments.

2 MANUAL MODE DOES NOT CAUSE SELF-SACRIFICIAL JUDGMENTS

The empirical objections to Greene’s model have centered around his use of other-sacrificial dilemmas. Many studies have demonstrated that utilitarian responses (or “prosacrificial responses”, as Greene’s critics call them) to these dilemmas do not correlate with utilitarian responses to other sorts of dilemmas. This finding casts doubt on the link between manual mode and utilitarianism in general, since utilitarianism applies to a much wider scope of actions than sacrificing some people to benefit more people. In particular, under certain conditions utilitarianism leads to the conclusion that one ought to sacrifice one’s own interests (I call these judgments self-sacrificial utilitarian responses). I outline different strands of evidence which suggest that automatic mode, not manual mode, drives self-sacrificial utilitarian responses. I go on to explain how my account of moral cognition converges with the views of Henry Sidgwick, the great utilitarian philosopher. Sidgwick argued that through the use of reason alone (no reliance on emotion) one could arrive at hedonism, which he subdivided into egoistic hedonism and utilitarian hedonism. In situations where these two systems conflict (i.e. situations in which utilitarianism requires self-sacrifice), Sidgwick concluded that reason alone could not determine what the normatively justified answer is, and that non-rational psychological factors will be what makes the difference. He does not specify what non-rational factors he has in mind, but presumably he meant affective or emotional factors.

2.1 Empirical Objections to Greene’s Thesis

I begin by pointing out that other-sacrificial dilemmas are used as the measure of utilitarianism in all of the relevant studies that Greene cites. The claim that utilitarian judgments in other-sacrificial contexts are associated with utilitarian judgments in other contexts is an empirical one, and it is one that has been challenged extensively. Because of this, Greene's critics often call "utilitarian" responses to such dilemmas "prosacrificial" responses (Everett & Kahane, 2020). Both terms refer to the same response (i.e. approval of sacrificing one to save many), but the term "prosacrificial" avoids conflation with utilitarian responses in other contexts.

There is considerable evidence correlating prosacrificial responses to other-sacrificial dilemmas with unsavory traits. For example, Bartels and Pizarro (2011) found that they were correlated with measures of Machiavellianism and psychopathy. Kahane et al. (2015) carried out several correlational studies finding that prosacrificial responses were associated with rational egoism, the belief that an action is rational only if it maximizes self-interest, and were also associated with a decreased sense of connection to all people. Furthermore, they found that people who gave prosacrificial responses were no more likely to do any of the following: donate money, report believing in an obligation to help distant people in poverty, or report believing in an obligation to help people in the future. Each of these are plausible measures of utilitarian beliefs. Kahane et al. (2015) conclude that prosacrificial responses to other-sacrificial dilemmas are not measures of genuine utilitarianism, but rather measures of low harm-aversion. In other words, people do not condone sacrificing one to save five because they care about the five, but because they do not care about harming the one.

This hypothesis is consistent with evidence that higher blood alcohol content *increases* utilitarian responses to trolley-like dilemmas (Duke & Bègue, 2015). Alcohol impairs deliberative cognition and social cognition. Greene's hypothesis should predict that alcohol should *decrease* utilitarian responses, or at least make no difference if the impairment of deliberation and empathizing cancel each other out. But the fact that alcohol increases utilitarian responses seems to indicate that low empathizing is a more dominant cause of such responses.

Greene's critics have also marshalled neuroscientific evidence: Wiech et al. (2013) found that decreased activity in the subgenual cingulate cortex (a region implicated in empathy) was associated with utilitarian judgments.[3] Strictly speaking, this is compatible with the dual-process model thesis, since Greene (2015) explicitly says that low empathy (or low reliance on automatic mode) should decrease utilitarian judgments. However, Kahane et al. (2012) also failed to replicate the correlation between the DLPFC (implicated in manual mode) and utilitarian responses to other-sacrificial responses. This evidence provides cumulative support for the contention that reduced sensitivity to harm (more so than increased deliberative reasoning) underlies prosacrificial judgments.

In response to the challenges that a) other-sacrificial responses do not predict utilitarian responses in general and b) other-sacrificial responses are primarily caused by a lack of harm-aversion (rather than increased deliberation), Conway and Gawronski (2013) used a method called process dissociation in order to test whether a genuinely utilitarian factor (as opposed to the mere lack of harm-aversion) could be extracted from prosacrificial responses. They used 10 traditional other-sacrificial dilemmas in which harm maximizes good consequences, but also included 10 parallel dilemmas in which similar harm did not maximize good consequences. For example, one traditional dilemma was whether to torture someone in order to locate lethal bombs. The parallel dilemma was whether to torture someone in order to locate paint bombs. The extent to which a person approves of harm in the latter kinds of dilemmas is subtracted from their tendency to approve of harm in the former kinds of dilemmas. When this is done,

[3] One confounding variable in traditional other-sacrificial dilemmas is that the utilitarian response is counterintuitive, whereas the deontological answer is not. Kahane et al. (2012) attempted to control for this confound by giving participants a battery of dilemmas including some in which the deontological answer was counterintuitive.

something called the U (utilitarian) parameter[4] is extracted. The U parameter captures a person's willingness to harm one to save many, subtracted by their general willingness to harm others.

Conway et al. (2018) replicated Kahane's findings that prosacrificial responses to other-sacrificial dilemmas correlated with psychopathy and egoism. However, they also found that participants' U parameter scores, which control for the mere absence of harm-aversion, correlated negatively with psychopathy. Furthermore, the U parameter scores correlated positively with various moral tendencies, including group focus (a tendency to report caring for the welfare of everyone involved in the dilemma), moral conviction about harm (the extent to which they felt that harm was wrong), and moral identity internalization (how much they identified with moral adjectives like "honest" and "honorable"). These results were interpreted as showing that utilitarian responses to other-sacrificial dilemmas do reflect some genuinely moral, even genuinely utilitarian, tendencies. Since U parameter scores are also reduced when participants are under cognitive load but do not change when participants are caused to be more empathic (Conway and Gawronski, 2013), there is also reason to believe they are uniquely influenced by manual mode.

Everett and Kahane (2020) responded to Conway et al. (2018) by pointing out that, crucially, the U parameter did not correlate (positively or negatively) with amounts participants would donate to charity (hypothetically) or to their responses to greater good dilemmas, which are about whether it is obligatory to sacrifice one's own interests (or the interests of one's family or ingroup) for the greater good. Thus, process dissociation still fails to show that utilitarian

[4] More precisely, the U parameter is equivalent to the likelihood of a participant saying harm is unacceptable in a congruent dilemma, minus their likelihood of saying harm is unacceptable in an incongruent dilemma: U= p(unacceptable|congruent) − p(unacceptable|incongruent). Suppose a participant says that harm is unacceptable in 9/10 congruent dilemmas, and says that harm is unacceptable in 3/10 incongruent dilemmas. He would have a U parameter score of .6.

responses to other-sacrificial dilemmas measure genuine utilitarianism when it comes to self-sacrifice. The finding that utilitarian responses to other-sacrificial dilemmas do not correlate with self-sacrificial responses is especially problematic for Greene (2013) because he appeals to the association between manual mode and "utilitarian" judgments to support the conclusion that we should (for example) give more to life-saving charities. But if the above criticism is correct, then this leap is unwarranted: there is no association (in everyday folk psychology) between approving of sacrificing one to save many and believing in donating to charity. In other words, other-sacrificial utilitarian tendencies do not predict self-sacrificial utilitarian tendencies. This is even true when process dissociation is used.

Conway et al. (2018) actually provide some evidence that automatic mode underlies self-sacrificial utilitarianism. People's responses to greater good dilemmas correlated with the D parameter (a measure of a deontological tendency extracted from process dissociation[5]) but not the U parameter. That is, people with more deontological responses were more likely to say you are obligated to sacrifice one's own (or one's group's) interests for the greater good. Since there is evidence (summarized earlier) that deontological responses are driven by automatic mode, self-sacrifice might also be driven by automatic mode. There is additional psychometric evidence for the connection between automatic mode and self-sacrificial utilitarianism. Kahane et al. (2018) designed a measure of utilitarian judgment called the Oxford Utilitarianism Scale. It has two scales which were derived from factor analysis: instrumental harm and impartial benevolence. The former scale is an other-sacrificial scale: it is about the permissibility of harming people for the sake of the greater good. The latter scale seems to be misnamed: four out

[5] The formula for the D parameter is D= p(unacceptable|incongruent) / (1−U), where U is the U parameter (explained in the previous footnote). A participant who said that harm is unacceptable in 9/10 congruent dilemmas, and says that harm is unacceptable in 3/10 incongruent dilemmas would have a deontological parameter score of .3/(1-.6)=.75

of five items measure willingness to sacrifice one's own interests (such as "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."), while only one is actually about impartial benevolence per se: "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." In terms of face validity, the scale seems to be primarily tracking self-sacrifice. This scale correlated positively with empathy and religiosity (Kahane et al., 2018). Religiosity, in turn, is associated with automatic mode: it correlates both with having emotional reactions in other-sacrificial dilemmas, and forming deontological judgments in those dilemmas (Szekely et al., 2015).

Furthermore, Rand (2016) performed a meta-analysis of experiments that included manipulations aimed at promoting intuition with one of the following manipulations: cognitive load, time constraints, ego depletion (i.e. a task that taxes one's self-control), and induction. Induction involves simply telling people to think a certain way, in this case intuitively or deliberatively. The studies provided very strong evidence that intuition tends to increase pure cooperation, whereas deliberation tends to decrease pure cooperation. Pure cooperation was defined as cooperation that was against one's strategic self-interest. Bear and Rand (2016) have also worked out an evolutionary game theoretic model which shows that agents who cooperate as an intuitive first response but whose deliberation favors defecting will evolutionarily outcompete agents whose deliberation favors cooperation. It is natural to use it as a further basis for predicting that automatic mode will produce self-sacrificial responses in the context of moral dilemmas. Taken together, there is compelling evidence that self-sacrificial tendencies are rooted in automatic mode, in people's emotional, intuitive responses to moral issues.

2.2 Convergence with Sidgwick

There is convergence between the above evidence and arguments from Henry Sidgwick (1901). Sidgwick explores three different ethical theories: common sense morality (CSM), egoism, and utilitarianism. He argues that CSM is filled with contradictions, which indicates the need for a system that can resolve those contradictions. He further argues that CSM often appeals to utilitarianism (implicitly) when resolving tensions between different duties. For example, when faced with the tension between the common-sense duty to save life and the common-sense duty to not lie, it is commonplace for ordinary people to appeal to the consequences when deciding which duty to follow, and we decide to lie to save lives. This is why Sidgwick says that CSM is "unconsciously utilitarian" (p. 424). Sidgwick also gives a conceptual argument for hedonism, the theory that only conscious states have intrinsic value. He asks us to imagine a world identical to this one, but merely lacking in conscious states. There are numerous other candidates of things that possess intrinsic value: freedom, friendship, preserving life, and so on. If any of these candidates actually are intrinsically valuable, they should be able to endow the world with value even if there were no conscious states. In today's terms, Sidgwick's hypothetical world amounts to a world filled with philosophical zombies.[6] Sidgwick expects this thought experiment to trigger the intuition that any world which is devoid of conscious states is also devoid of intrinsic value (p. 396). He concludes that in such a world nothing would be of value—including the freedom, friendships, even lives of such beings.

[6] Of course, the notion that a philosophical zombie is conceivable rests on premises that have been extensively criticized (Dennett, 1995), but this is not the place to explore that question.

This argument for hedonism is compatible both with hedonistic egoism (the view that only one's own pleasurable conscious states should be maximized) and hedonistic utilitarianism (the view that everyone's pleasurable conscious states should be maximized). Sidgwick (1901) recognizes this tension, and devotes considerable space to arguing that in most cases, egoism and utilitarianism will converge on the same answers. Given that most people in most situations derive positive emotion from prosocial behavior, it is in the egoist's best interest to behave the same way that a utilitarian would. Sidgwick rightly says that the extent to which this is true is an empirical question. However, he concludes by saying that:

> In the rarer case of a recognized conflict between self-interest and duty, practical reason, being divided against itself, would cease to be a motive on either side; the conflict would have to be decided by the comparative preponderance of one or other of two groups of non-rational impulses. (p. 508)

This means that in cases where egoism and utilitarianism conflict, reason cannot decide the issue: other "non-rational" psychological factors must do so. Insofar as the happiness that comes from giving to charity does not outweigh the happiness that one loses, the self-sacrificial dilemmas are cases in which egoism and utilitarianism conflict. In such cases, Sidgwick says that rational factors will not be the difference that makes a difference.

Sidgwick's argument provides conceptual justification for a distinction between self-sacrificial utilitarian judgments and other-sacrificial utilitarian judgments. Someone could be persuaded by Sidgwick's argument for the intrinsic value of conscious states, and make utilitarian judgments for all dilemmas that do not affect their own self-interest. However, he contends that in cases where one must choose between maximizing the general welfare and one's own welfare, psychological forces other than reason will be what makes the difference. He leaves open exactly what these will be, but presumable he has some affective factors in mind. Sidgwick's philosophical argument matches the evidence cited above: cold reasoning is not sufficient to lead to the conclusion that one must sacrifice one's own interest for utilitarian ends. When faced with a self-sacrificial dilemma, non-rational or emotional psychological factors are the difference that make a difference. In the next section, I outline an experimental test that I performed to test my hypothesis.

3 EXPERIMENTAL TEST

In this section I explain the study that I performed to determine whether reliance on manual mode or reliance on automatic mode is more likely to undergird self-sacrificial judgments. I lay out two hypotheses: 1) the ability to control the influence of intuition and emotion should be positively associated with utilitarian responses to other-sacrificial dilemmas, and 2) the ability to control the influence of intuition and emotion should be negatively associated with utilitarian responses to self-sacrificial responses. I use two measures of the ability to control the influence of intuition and emotion: the cognitive reflection test (Frederick, 2005) and the reappraisal scale of the emotion regulation questionnaire (Gross & John, 2003). I find some support for hypothesis one (which replicates previous work), but no support for hypothesis two. I instead find no association (positive or negative) between reliance on manual mode and self-sacrificial judgments. This still presents problems for Greene's thesis, which I spell out in the discussion.

3.1 Introduction

Despite the fact that Sidgwick gives philosophical arguments in favor of the claim that non-rational factors drive self-sacrificial judgments, that claim is strictly empirical. As already mentioned, there is evidence that performance on the cognitive reflection test (CRT) is a valid measure of reliance on rational algorithmic thinking (Toplak et al., 2011). Using the CRT, one can test whether people who are more likely to engage manual mode are more or less likely to make self-sacrificial utilitarian judgments. If self-sacrificial utilitarian judgments are caused by automatic mode (as I am conjecturing), then CRT scores should *negatively* correlate with self-sacrificial judgments.

There is also a validated scale called the emotion regulation questionnaire (ERQ), which has two components: a reappraisal scale and a suppression scale (Gross & John, 2003). The reappraisal scale measures how much someone tends to cognitively change their assessment of a situation in order to control their emotions, and there is evidence that it predicts certain kinds of moral responses. Feinberg et al. (2012) gave participants a set of moral dilemmas from Haidt (2001) involving harmless taboo violations, such as having sex with a sibling whilst using birth control. Higher scores on the reappraisal scale of the ERQ were predictive of decreased condemnation of harmless taboo violations. Failing to condemn harmless taboo violations looks like a paradigm case of a judgment driven by manual mode: most people have some initial revulsion towards the taboo violation, and manual mode is what reduces said revulsion. The result from Feinberg et al. (2012) thus indicates that the reappraisal scale tracks people's ability to reduce the influence of emotion on moral judgment. If my hypothesis is correct, then scores on the ERQ reappraisal scale should also *negatively* correlate with self-sacrificial judgments.

In Paxton et al. (2012), the CRT was found to be correlated positively with other-sacrificial responses, but only in the condition in which the CRT was presented first. I conducted a pilot study that tested the association between CRT scores and self-sacrificial responses, and I similarly found that the correlation (in this case, negative) between CRT scores and self-sacrificial responses only obtained when the CRT was presented first. Thus, I had two conditions, one in which the CRT is presented before the dilemmas (and the ERQ is presented after), and one in which the ERQ is presented before the dilemmas (and the CRT is presented after).

3.2 Method

243 Georgia State University students were recruited via instructor invitations and were given course credit for participation. The exclusion criteria were set in advance. 61 participants were excluded for doing one or more of the following: failing the attention check, failing any of the comprehension checks, taking less than five minutes, or taking more than 60 minutes. Applying these criteria left 182 participants. 67.6% of the participants were female, 31.3% were male, and 1.1% responded "other/prefer not to answer", with an average age of *M* = 19.9 years (*SD* = 4.0). Self-reported racial affiliation was 43.8% Black or African American, 31.0% White or Caucasian, 16.4% Asian, 14.6% Hispanic or Latino, 2.6% American Indian or Alaska Native, 1% Native Hawaiian or other Pacific Islander, 2.2% prefer not to answer (these categories were not exclusive). The average completion time in minutes was *M* = 22.5 (*SD* = 10.8).

All participants were presented with the cognitive reflection test (CRT), which consisted of three word problems involving arithmetic. Each question had an intuitive (but incorrect) answer, and a correct answer that required reflection (Byrd & Conway, 2019). For example, one question was "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?", where the intuitive answer is 200 minutes, but the correct answer is two minutes. As mentioned, the emotion regulation questionnaire (ERQ) was also included. Participants were randomly assigned to either the CRT-first condition (n = 89) or the ERQ-first condition (n = 93). The Appendix contains all CRT and ERQ questions.

Each participant was given two self-sacrificial dilemmas, both of which were modified from Kahane et al. (2015), and two other-sacrificial dilemmas, modified from Greene et al. (2008). All dilemmas were modified to be in second person ("you"), and the order of the dilemmas was randomized. Here is an example of a self-sacrificial dilemma, and an other-sacrificial dilemma, respectively:

> [SELF-SACRIFICIAL] 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.
>
> How acceptable would it be for you to spend the money on a new cell phone instead of the malaria charity, even though a child you could have helped will die?
>
> [OTHER-SACRIFICIAL] It is war time. Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and 10 other 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.
>
> How acceptable would it be for you to smother the child[7], even though that would mean you, the child, and the 10 other townspeople would be killed?

For exploratory analyses, participants were also given two dilemmas about the acceptability of failing to prioritize the greater good for each of the following categories: family dilemmas (in which the choice concerns helping many or helping a member of one's family), ingroup dilemmas (in which the choice concerns helping a few in one's ingroup or many in one's outgroup), and animal welfare dilemmas. Most of these dilemmas were from Conway et al. (2018), but two were internally developed. Participants also completed the Oxford Utilitarianism Scale, which included two subscales: instrumental harm and impartial benevolence (Kahane et

[7] One problem with the traditional other-sacrificial dilemmas was that higher permissibility ratings were more utilitarian, whereas in the remaining four types of dilemmas, higher permissibility ratings corresponded to a less utilitarian response. To solve this, we randomized the valence of the other- sacrificial dilemmas: half of participants were asked how acceptable it is to kill one to save others, and the other half were asked how acceptable it was to NOT kill one to save others. The latter question was reverse coded so that the questions could be combined into a single measure.

al., 2018). The former scale measures willingness to harm people for the sake of the greater good, whereas the latter purports to measure one's belief in the obligation to care for all people equally. These scales were included to provide a psychometrically validated measure of two empirically distinct utilitarian tendencies. As mentioned, four out of the five items in the impartial benevolence scale are self-sacrificial items, so in terms of face validity it looks like a self-sacrificial measure. All scales and dilemmas are available in the Appendix. Next, I will outline my hypotheses. Note that all hypothesis tests involving the CRT will only make use of subjects in the CRT-first condition, and all hypothesis tests involving the ERQ will only make use of subjects in the ERQ-first condition.

Hypothesis 1: The ability to control the influence of intuition and emotion should be positively associated with other-sacrificial responses. This hypothesis is, in part, a replication of previous work (Paxton et al., 2012). There are two measures of the ability to control the influence of intuition and emotion: the CRT, and the ERQ reappraisal scale. From this hypothesis I derived four predictions: 1) CRT scores will positively correlate with other-sacrificial responses (i.e. approval of killing one to save many), 2) CRT scores will positively correlate with the instrumental harm subscale, 3) reappraisal scores will positively correlate with other-sacrificial responses, and 4) reappraisal scores will positively correlate with the instrumental harm subscale.

Hypothesis 2: The ability to control the influence of intuition and emotion should be negatively associated with self-sacrificial responses. This is based on the evidence I summarized which suggests a link between automatic mode and self-sacrifice. Since the CRT and the ERQ reappraisal scale both measure one's tendency to control the influence of automatic mode, they should be negatively associated with self-sacrifice. From this hypothesis I derived four further predictions: 5) CRT scores will negatively correlate with self-sacrificial responses (i.e. disapproval of keeping one's money when it could save lives), 6) CRT scores will negatively correlate with the impartial benevolence subscale, 7) reappraisal scores will negatively correlate with self-sacrificial responses, and 8) reappraisal scores will negatively correlate with the impartial benevolence subscale.

3.3 Results

Descriptive statistics. The CRT consisted of three questions. 24% of participants got none of the questions correct, 34.8% got one correct, 29.8% got two correct, and 11.1% got all questions correct. For the reappraisal scale of the ERQ, the highest possible score was 42. In this sample, the mean was *M* = 30.6 (*SD* = 6.0). For the self-sacrificial dilemmas, the highest possible score was 14 (strongly agreeing with the utilitarian option in both cases) and the average response was *M* = 7.5 (*SD* = 2.9). For the other-sacrificial dilemmas, the highest possible score was 14 (strongly agreeing with the utilitarian option for both cases) and the average response was *M* = 7.9 (*SD* = 2.8). The highest score on the impartial benevolence scale was 35, and the average score was *M* = 20.4 (*SD* = 5.3). The highest score on the instrumental harm scale was 28, and the average score was *M* = 15.1 (*SD* = 4.6).

Hypothesis one. This received some support.[8] CRT scores were significantly positively correlated with other-sacrificial responses, *r* = .263, *p* = .013, but the CRT did not significantly correlate with the instrumental harm scale: *r* = .068, *p* = .525. With regard to the ERQ, reappraisal scores were not significantly correlated with other-sacrificial responses: *r* = -.052, *p* =

[8] As indicated in the method section, all hypothesis tests involving the CRT made use of subjects in the CRT-first condition and all hypothesis tests involving the ERQ made use of subjects in the ERQ-first condition. I found no significant order effects for the CRT, ERQ, or the dependent measures for the hypothesis tests.

.618 or the instrumental harm subscale: *r* = .072, *p* = .491. As already explained, the CRT and ERQ reappraisal scale have been treated in other contexts as being theoretically similar: both are measures of the ability to control initial responses through deliberate reflection. It is thus noteworthy that CRT scores and ERQ reappraisal scores had correlations with other-sacrificial responses in opposite directions (with only the CRT reaching significance). An exploratory analysis revealed that, contrary to expectations, CRT and reappraisal scores were negatively (though not quite significantly) correlated: *r* = -.129, *p* = .083.

Hypothesis two. This did not receive any support. CRT scores were not negatively correlated with self-sacrificial responses, *r* = -.012, *p* = .909, nor were they negatively correlated with the impartial benevolence subscale, *r* = -.017, *p* = .875. Similarly, reappraisal scores were not negatively correlated with self-sacrificial responses: *r* = .038, *p* = .716, or with the impartial benevolence subscale: *r* = .102, *p* = .331. These null results were inconsistent with my pilot study, which found that CRT scores were significantly negatively correlated with self-sacrificial responses: *r* = -.399, *p* = .036. However, the pilot study also involved a cognitive load manipulation: participants read the moral dilemmas after memorizing a spatial pattern (either simple or complex). It is possible that CRT scores only predict self-sacrificial responses when people are put under cognitive load.

Exploratory analyses. Consistent with the idea that they are driven by different psychological faculties, self-sacrificial and other-sacrificial utilitarian responses were negatively correlated: *r* = -.202, *p* = .006. Despite the fact that the first scale from Kahane et al. (2018) is called the "impartial benevolence" scale, it only positively correlated with self-sacrificial responses: *r* = .378, *p* < .001. The scale did not significantly correlate with utilitarian responses to dilemmas involving one's family (*r* = .099, *p* = .185), one's ingroup (*r* = .062, *p* = .404), animals (*r* = .096, *p* = .198). Since there was a significant correlation between CRT scores and other-sacrificial utilitarian responses only in the CRT-first condition (the correlation in the CRT-second condition was *r* = .056, *p* = .594), we tested for an order effect being presented with the CRT. Those in the CRT-first condition gave less utilitarian responses to other-sacrificial dilemmas (*M* = 7.573, *SD* = 2.888) than those in the CRT-second condition (*M* = 8.215, *SD* = 2.670), but this difference was not significant: *t*(180) = -1.558, *p* = .121.

3.4 Discussion

Hypothesis one was partially confirmed. That is, the evidence supported the idea that the ability to control the influence of intuition and emotion was positively associated with other-sacrificial utilitarian responses. This finding lends support to Greene's contention that manual mode undergirds utilitarian judgments in other-sacrificial contexts. With regard to hypothesis two, this study failed to replicate the negative association between CRT scores and self-sacrificial responses that I found in the pilot study. Despite not confirming my hypothesis, this finding counts strongly against the argument that manual mode produces utilitarian judgments in such cases. More importantly, it counts against Greene's argument that we should give more to charity because a) manual mode is associated with utilitarian judgments, and b) manual mode should be relied upon for unfamiliar moral problems (Greene, 2013). The fact that CRT scores predict utilitarian responses to other, but not self, sacrificial dilemmas is consistent with the contention that manual mode does not consistently drive utilitarian judgments.

This study also did not find the predicted negative association between the ability to control the influence of emotion (as measured by the ERQ reappraisal scale) and self-sacrificial utilitarian responses. Perhaps this measure is too coarse-grained to detect effects of particular emotions. One could test the extent to which specific emotional factors undergird self-sacrificial utilitarian responses. One candidate factor is empathy. Empathy positively correlates with the impartial benevolence scale, which is predominantly a self-sacrificial measure (Kahane et al., 2018) Contra Bloom's (2016) claim that empathy tends to make people favor their ingroup, Sierksma et al. (2015) found evidence that in fact, inducing empathy tends to increase concern for outgroup members (see also Batson et al., 1981). Future experiments could try to detect effects of inducing empathy on self-sacrificial responses.

Another emotion that could potentially drive self-sacrifice is guilt. One obvious way to test this would be to use psychometrically validated measures of a tendency to feel guilt, such as the guilt and shame proneness scale (Cohen et al., 2011), and look for a correlation with self-sacrificial responses. In order to make causal inferences, one could experimentally induce a sense of guilt before giving self-sacrificial dilemmas. To the extent that guilt is driving self-sacrifice in a given person, one might expect that person to endorse self-sacrifice even if it does not promote the greater good. Process dissociation is a tool that could provide insight here. When process dissociation was used for the other-sacrificial dilemmas, the goal was to dissociate genuine utilitarian motivations from a lack of harm-aversion (Conway & Gawronski, 2013). For self-sacrificial dilemmas, the goal would be to dissociate utilitarian motivations from a general willingness to sacrifice one's own interests even if it does not promote the greater good. This non-utilitarian self-sacrificial tendency, which would be controlled for by process dissociation, could plausibly be associated with excessive guilt.

It would also be useful to take neuroimaging paradigms like those summarized by Greene (2015) and apply them to self-sacrificial dilemmas. One could have participants undergo fMRI scans and test whether there is a correlation between DLPFC activity and self-sacrificial responses. Based on my second hypothesis, I would predict that activity in the emotional centers of the brain (such as the amygdala or VMPFC) would correlate with self-sacrificial responses, but DLPFC activity would not. Using transcranial magnetic stimulation (TMS) or transcranial direct current stimulation (tDCS) to interfere with these regions would allow us to make causal inferences. Zheng et al. (2018) interfered with activity in the DLPFC (using tDCS), and this reduced the proportion of utilitarian responses, but only other-sacrificial dilemmas were used. Again, my prediction, based on the predicted connection between automatic mode and self-sacrifice, would be that interfering with the DLPFC would increase self-sacrificial responses.

With regard to my exploratory analyses, they have implications for the two-dimensional model of Kahane et al. (2018). The so-called "impartial benevolence" scale mostly contains self-sacrificial items (four out of five items). As was pointed out earlier, in terms of face validity, it seems more appropriate to call it a self-sacrificial scale. Furthermore, I found that this scale does not correlate with dilemmas about whether to prioritize the greater good over one's own family, ingroup, or species: it only correlates with self-sacrificial items. In their critique of Conway et al. (2018), Everett and Kahane (2020) point out that there is no evidence that the U parameter is associated with greater impartiality. However, this study shows that their subscale is also not associated with greater impartiality in general. I did not use the full battery of 20 dilemmas that Conway used to extract a U parameter, but future studies should do so. Rather than attempt to correlate the U parameter with greater-good dilemmas in general, future studies should break these dilemmas down into subcategories. This study is a reminder that the psychometric properties of the various types of greater-good dilemmas (which have previously been grouped together) might be different.

There are potential problems with the measures that I have used. For example, one objection that has been raised against the use of the CRT is that it measures both reflectiveness and mathematical ability (Jaquet & Cova, 2021). There are other measures of cognitive reflection that rely more on verbal reflection that could be used, but these measures have a less straightforward relationship with utilitarian judgments (Byrd & Conway, 2019). A more problematic confound in a different context is the fact that there are differences between the other-sacrificial dilemmas and self-sacrificial dilemmas other than the extent to which one's own interests are sacrificed. To name a few, the other-sacrificial dilemmas involve violence, the deaths involved would be up close and personal, and they have more lives on the line. All of these variables could be relevant, and future experiments could test whether they are. [9] There was a tradeoff between maximizing consistency with past studies and maximizing experimental control, and I generally opted for maximizing the former.

The finding that is the most difficult to interpret is the following: there was a *positive* correlation between the CRT and other-sacrificial utilitarian responses in the CRT-first condition (no correlation in the CRT-second condition), but participants were *less* utilitarian in the CRT-first condition (albeit not significantly so). One interpretation is that, paradoxically, doing the CRT induces intuition, perhaps because a majority of participants got one out of three correct or less. It is also possible that doing the CRT is a kind of cognitive load, which decreases most participants' utilitarian responses to other-sacrificial dilemmas. This explanation would suggest that the correlation exists in the CRT-first condition because the very reflective people continue to be willing to sacrifice someone for the greater good even in the case of cognitive load. In the

[9] One additional problem with the other-sacrificial dilemmas that I used is that they are such that if the person does nothing, they will be harmed as well. However, experiments which have included dilemmas where the person will *not* be harmed if they do nothing still find that utilitarian responses to such dilemmas are associated with manual mode (Conway & Gawronski, 2013; Byrd & Conway, 2019).

space remaining, I will focus on the implications of the positive correlation per se, since it was a) statistically significant, b) predicted in advance, and c) easy to interpret.

4 RESPONSE TO OBJECTIONS

There are some arguments that Greene has made that he could use to respond to the above evidence. For example, Greene has pointed to a deep symmetry between other-sacrificial and self-sacrificial dilemmas. He does this through appeal to Singer's (1972) drowning girl example, which I will explain. Singer famously made the analogy between allowing a drowning girl to drown because saving them would ruin one's costly clothing, and not donating comparable amounts of money to life-saving charities. Saving a life at the cost of one's property seems obligatory in the drowning case, but not the charity case. We could call the tension between these two cases the "drowning girl problem". Similarly, the "trolley problem" is that it seems permissible to kill one to save five with a switch, but not with pushing. Greene (2013) argues that there is a deep symmetry between the trolley problem and the drowning girl problem in that emotional salience is what causes the difference in judgment. Letting a girl die right in front of you and directly pushing the fact man both (plausibly) lead to a visceral reaction. The psychological reason why (according to Greene) donating to life-saving charities does not intuitively feel obligatory and why redirecting a trolley with a switch does not intuitively feel wrong is because both cases fail to produce a visceral reaction.

Greene does offer some empirical support for the symmetry between these types of dilemmas: in work that became part of an unpublished manuscript, Musen & Greene (MS, cited in Greene, 2013) performed experiments on people's responses to Singer-like descriptions of a drowning girl. They manipulated different variables, and found that physical distance was the variable that most influenced people's evaluation.[10] In other words, the reason that most people say it is permissible to *not* save lives via giving to charity (but not via jumping into a pond) is

[10] I will leave aside the fact that some researchers have found that distance does not influence moral judgment in these cases when other factors are statistically controlled for (Nagel & Waldmann, 2012).

simply that they are far away. A similar finding obtains in trolley cases: the reason that most people say it is permissible to kill someone via pulling a switch (but not via pushing someone) is simply that they are far away (Greene, 2015). Remember, Greene contends that manual mode uniquely facilitates self-sacrificial utilitarian judgments, whereas I contend that it does not. Let us grant that morally irrelevant factors drive people to make the non-utilitarian judgment about giving to charity. That does not demonstrate that manual mode causes the utilitarian judgment for such dilemmas. It could be true that morally irrelevant factors drive people to make the non-utilitarian judgment, and that other (different) emotive factors lead to the utilitarian judgment.

However, there are also a few recent studies that cast doubt on my argument that reflection does not drive self-sacrifice. Lindauer et al. (2020) presented subjects with a rational argument for charitable giving, which increased charitable giving as much as an emotional appeal did. The rational argument was Singer's argument that letting the drowning girl die is morally equivalent to letting a girl in the third world die, conjoined with an evolutionary debunking argument against our stronger emotional reaction to nearby suffering. I would respond by saying that this result does not demonstrate that (ordinarily) self-sacrificial judgments are rooted in manual mode, only that self-sacrificial judgments can be induced by making powerful arguments in their favor. It is conceivable that one could cause people to make more deontological judgments by showing them a powerful argument from a deontological philosopher. That would be consistent with the evidence that deontological judgments are (ordinarily) rooted in automatic mode. Similarly, the fact that one can induce a self-sacrificial utilitarian judgment through the use of an argument is consistent with the claim that such judgments are normally caused by automatic mode. The finding is nonetheless significant: one can rationally persuade people to give self-sacrificial responses, even if it turns out that they are ordinarily rooted in automatic mode.

Jaquet and Cova (2021) performed a series of experiments to test Greene's hypothesis that manual mode undergirds utilitarianism in general. Unlike the traditional experiments, they used several different types of dilemmas. The dilemma type that is of most relevance is the "demandingness ethics" (DE) type. The DE dilemmas are very similar to the dilemmas that I used as a measure of self-sacrificial utilitarianism. The only substantive difference is that the DE dilemmas do not emphasize the extent to which giving money to charity is a sacrifice (e.g. the dilemmas I used highlight that the person can no longer buy a phone or a car). This will be discussed later. Unlike my pilot study and larger study, Jaquet and Cova (2021) found evidence consistent with the hypothesis that manual mode drives utilitarian responses to DE dilemmas. Cognitive load and time restraints both reduced utilitarian responses to the DE dilemmas, and instructing people to rely on reason increased utilitarian responses. However, they found that CRT scores negatively correlated with utilitarian responses to those dilemmas, which is inconsistent with the view that utilitarian reasoning is more reflective but consistent with my pilot study and larger study. My studies found a negative correlation and a nonsignificant correlation, respectively.

There is thus a conflict: if we use CRT scores as our main measure of reliance on manual mode, Greene's hypothesis is put in jeopardy by Jaquet and Cova (2021) and my studies. If we use cognitive load, time pressure, and instructions to use reason as our main manipulations aimed at influencing people's reliance on manual mode, Greene's hypothesis is corroborated. Future studies should try to estimate the validity of these manipulations. One problem with instructing people to rely on reason is that this is confounded by what judgments people *think* are the kinds of judgments that cold reason tends to make. The idea that utilitarian judgments are cold and calculating might be one that ordinary people make. The fact that instruction to use reason causes people to make utilitarian judgments might merely show that people *believe* that cold, calculating types tend to make utilitarian judgments. This is distinct from the idea that reason *actually* produces utilitarian judgments. The CRT is arguably not subject to this same objection: it measures actual reliance on manual mode, and it does have the same risk of inducing people to make judgments that they think they should be making. Jaquet and Cova (2021) point out (as I did in the discussion) that the CRT is an imperfect measure of rational reflection because it also measures mathematical ability. They conclude that it should not be relied upon as much as their other operational definitions of reliance on manual mode. I will point out that Toplak et al. (2011) used several measures of rationality and of heuristic-based reasoning as the criterion variables, and showed that the CRT has predictive validity independent of IQ. CRT scores and IQ scores are correlated, but they independently predict performance on a wide range of tests.

A more salient issue is that the DE dilemmas do not (as my dilemmas do) make salient the personal sacrifice involved in giving money to charity. It could be that the contribution in the hypothetical dilemma is thought to be sufficiently trivial so as to not make much different to the agent. My hypothesis would be that if one were to make self-sacrifice salient, the findings in Jaquet and Cova (2021) would not replicate. According to Popper (1974, cited in Bamford, 1994), it is acceptable to defend a theory from apparent falsification by adding a new hypothesis to the theory, so long as that hypothesis itself leads to testable predictions. By contrast, if one continually adds hypotheses that themselves add no further predictions, the theory is being illegitimately protected from falsification. My defense of the claim that manual mode does not cause self-sacrificial judgments against this recalcitrant evidence does lead to new testable predictions. If I am right that the low salience of self-sacrifice is the reason that Jaquet and Cova (2021) found evidence that manual mode drives utilitarian judgments, then a study which used the same manipulations but used dilemmas that made self-sacrifice salient should not find evidence that corroborates Greene's thesis.

CONCLUSION

Greene makes a convincing empirical argument that, at least in cases that involve sacrificing another person, manual mode (slow, reflective thinking) tends to cause utilitarian judgments. He goes on to argue that automatic mode (which produces deontological or non-utilitarian judgments in such cases) could not be reliable for unfamiliar moral problems (such as whether to give to life-saving charities) in the absence of cognitive miracles. He concludes that automatic mode should not be relied upon for said problems. Furthermore, automatic mode is causally driven by factors that utilitarians and non-utilitarians would not upon reflection endorse, which makes it seem likely that the arguments put forward in deontological judgments are post-hoc rationalizations. These normative arguments from Greene on the basis of his findings have been challenged by Lott (2016), but there is a defensible way to make philosophical use of these findings. I put forward the following analogy: if it were experimentally shown that the color of a movie title is what caused people to like or dislike a movie, this would cast doubt on whatever arguments they put forward for their aesthetic judgment. It might still be the case that these arguments turn out to be valid upon reflection, but the knowledge that they are likely post-hoc rationalizations should lead us to be more vigilant in looking for errors in reasoning. One could apply this argument to the moral domain as well. If a moral judgment is caused by factors that we deem irrelevant, that does not by itself invalidate the moral judgment or the reasons offered for it, but it does cast some doubt on it.

I outlined an empirical objection to Greene's dual-process model, namely the fact that the evidence Greene cites only uses dilemmas that involve other-sacrifice. Even when process dissociation is used, utilitarian responses to other-sacrificial response are not associated with utilitarian responses to self-sacrificial responses (Conway et al., 2018). I summarized various pieces of psychological evidence indicating that automatic mode tends to drive self-sacrifice, and explained how this converged with the views of the utilitarian philosopher Henry Sidgwick. Sidgwick (1907) argued that reason cannot decide egoistic and utilitarian hedonism. When a dilemma requires one to choose between one's own welfare and the greater welfare, non-rational factors in a person's psychology will be the deciding factor. This is consistent with the hypothesis that automatic mode (a non-rational psychological factor) drives self-sacrifice. I performed an experiment to test this hypothesis. My experiment replicated the finding that CRT scores positively correlate with utilitarian responses to other-sacrificial dilemmas, but found no correlation with utilitarian responses to self-sacrificial dilemmas.

The finding that manual mode does not drive utilitarian responses to self-sacrificial dilemmas casts doubt on Greene's thesis, since it depends on the premise that manual mode drives utilitarian judgments in general. However, utilitarianism could still be justified by his other arguments. As was already discussed, Greene (2013) has an argument for utilitarianism which does not depend on empirical findings from psychology, namely the common currency argument. The argument goes as follows: given the substantial disagreement between cultures about which moral axioms we should accept, there is a need for some collectively agreed upon metric for resolving disagreement. Greene likens this metric to a common currency. He then argues that utilitarianism is the ideal moral theory for being used as a common currency, since it relies on an axiom that almost uniformly accepted. Since almost everyone agrees that, ceteris paribus, maximizing happiness is good, utilitarianism is a promising system for resolving otherwise intractable moral disagreements.

As discussed, there is preliminary evidence suggesting that automatic mode causes self-sacrifice. For example, there is evidence of a positive correlation between empathy self-sacrificial utilitarian judgments (Kahane et al., 2018), and evidence that intuition drives pure cooperation (i.e. cooperation that is against one's strategic self-interest) in economic games (Rand, 2016). If automatic mode does drive utilitarian responses to self-sacrificial dilemmas, then defenders of Greene could lean more heavily on the common currency argument as a defense for using utilitarianism to resolve moral disagreement. As a result, they would have to conclude that he does not give automatic mode enough credit. If automatic mode sometimes causes utilitarian judgments (which are the judgments we should be relying on, according to the common currency argument) in unfamiliar moral problems, then automatic mode should be taken more seriously in these contexts, by Greene's own lights. This would mean that solving contemporary moral disagreements requires more than cold, robotic cogitation. Emotions and intuitions likely have a role to play as well. This counts against recent popular books arguing that we should be distrustful of empathy (Bloom, 2016), and that reliance on reason and science alone can solve moral problems (Harris, 2011).

Remember, Greene argued that automatic mode cannot be reliable for unfamiliar moral problems in the absence of a cognitive miracle. The defining feature of manual mode is its flexibility, which is what allows it to work in unfamiliar situations. If automatic mode is reliable for at least some unfamiliar problems, this would open up the question of how it is possible for automatic mode to be reliable for problems it was not designed to solve. Perhaps it is the case that although our evolutionary past did not specifically shape automatic mode for these unfamiliar problems, these problems are sufficiently similar to problems we faced in our ancestral environment. If the problems are sufficiently similar, we can trust allow automatic mode to be reliable for these problems. The question of what constitutes sufficient similarity is a question that future empirical and philosophical work could address.

In conclusion, it would be convenient if ordinary responses to moral dilemmas were neatly divided into utilitarian and deontological ones. It would be even more convenient if the utilitarian responses were consistently and exclusively undergirded by manual mode. I have argued that there are empirical and theoretical reasons to think that manual mode is not associated with utilitarian responses to self-sacrificial dilemmas. While my study replicated the finding that reliance on manual mode is associated with utilitarian judgments in other-sacrificial dilemmas, it found no support for a similar association with self-sacrifice. I have responded to objections, and outlined directions for future research.

REFERENCES

Bamford, G. (1993). Popper's explications of ad hocness: Circularity, empirical content, and scientific practice. *The British Journal for the Philosophy of Science, 44*(2), 335-355.

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.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences, 113*(4), 936-941.

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.

Bloom, P. (2016). *Against empathy: The case for rational compassion.* Random House.

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.

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.

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: a new measure of guilt and shame proneness. Journal of Personality and Social Psychology, 100(5), 947-966.

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.

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

Dennett, D. C. (1995). The unimagined preposterousness of zombies. *Journal of Consciousness Studies, 2*(4), 323-326.

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

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

Feinberg, M., Willer, R., Antonenko, O., & John, O. P. (2012). Liberating reason from the passions: Overriding intuitionist moral judgments through emotion reappraisal. *Psychological Science, 23*(7), 788-795.

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

Foot, P. (1978). Abortion and the doctrine of double effect. *Virtues and Vices; and other essays in moral philosophy, Clarendon.*

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.

Glenn, A. L., Raine, A., Schug, R. A., Young, L., & Hauser, M. (2009). Increased DLPFC activity during moral decision-making in psychopathy. *Molecular Psychiatry, 14*(10), 909-911.

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. (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.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. Journal of Personality and Social *Psychology, 85*(2), 348-362.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology, 65*(2), 221-233.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814-834.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). *Moral dumbfounding: When intuition finds no reason.* Unpublished manuscript, University of Virginia, 191-221.

Harris, S. (2011). *The moral landscape: How science can determine human values.* Simon and Schuster.

Jaquet, F., & Cova, F. (2021). Beyond moral dilemmas: The role of reasoning in five categories of utilitarian judgment. *Cognition, 209*(1), 104572.

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.

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.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience, 7*(4), 393-402.

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

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.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., Cushman, F. (2020). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. Available at: https://psyarxiv.com/q86vx

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

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.

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.

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.

Sierksma, J., Thijs, J., & Verkuyten, M. (2015). In-group bias in children's intention to help can be overpowered by inducing empathy. *British Journal of Developmental Psychology, 33*(1), 45-56.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*(5), 645-665.

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

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.

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.a

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.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476-477.

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.

Yang, Y., Liang, P., Lu, S., Li, K., & Zhong, N. (2009). The role of the DLPFC in inductive reasoning of MCI patients and normal agings: 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?

Emotion Regulation Questionnaire

1. When I want to feel more positive emotion (such as joy or amusement), I change what I'm thinking about. (Reappraisal scale)
2. I keep my emotions to myself. (Suppression scale)
3. When I want to feel less negative emotion (such as sadness or anger), I change what I'm thinking about. (Reappraisal scale)
4. When I am feeling positive emotions, I am careful not to express them. (Suppression scale)
5. When I'm faced with a stressful situation, I make myself think about it in a way that helps me stay calm. (Reappraisal scale)

6. I control my emotions by not expressing them. (Suppression scale)
7. When I want to feel more positive emotion, I change the way I'm thinking about the situation. (Reappraisal scale)
8. I control my emotions by changing the way I think about the situation I'm in. (Reappraisal scale)
9. When I am feeling negative emotions, I make sure not to express them. (Suppression scale)
10. When I want to feel less negative emotion, I change the way I'm thinking about the situation. (Reappraisal scale)

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 benevolence)
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 benevolence)
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 benevolence)
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 benevolence)
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 benevolence)

Dilemmas

Once you get to the scenarios, please read them carefully. Even though some scenarios may seem similar, each scenario is different in important ways. Please note that some scenarios refer to things that may seem unpleasant to think about. This is because we are interested in people's thoughts about difficult, real-life issues.

From totally unacceptable to totally acceptable (seven-point Likert scale)

First self-sacrificial dilemma. 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.

How acceptable would it be for you to spend the money on a new cell phone instead of the malaria charity, even though a child you could have helped will die?

Second self-sacrificial dilemma. 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 save the lives of several of the hospitalized tsunami victims. However, you would have to continue living without a car.

How acceptable would it be for you to spend your uncle's $7000 on the car, even though this means that some tsunami victims will die who could have been saved?

First other-sacrificial dilemma. You are the captain of a military submarine traveling underneath a large iceberg. An onboard explosion has caused the submarine to lose most of the oxygen supply and has injured one of your crew who is quickly losing blood. The injured crew member may die from his injuries. The computer indicates that the remaining oxygen is not sufficient for the entire crew of 8 to make it to the surface. The only way to save the other crew members is to shoot dead the injured crew member so that there will be just enough oxygen for the rest of the crew to survive.

How acceptable would it be for you to (NOT[11]) shoot the injured crew member, even though that would mean all of the crew members would suffocate?

Second other-sacrificial dilemma. It is war time. Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and 10 other 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.

How acceptable would it be for you to [NOT] to smother the child, even though that would mean you, the child, and the 10 other townspeople would be killed?

[11] Whether or not this was included was randomized. This is also true of the next dilemma.

First family dilemma. You are a firefighter who is rescuing people from a burning building. The building is about to collapse, so the firefighters are rushing out. You are the last firefighter in the building, and will only have time to rescue one more person on your way out. In the last room, you find two people trapped. You immediately recognize one as a famous peace negotiator. This negotiator won the Nobel Peace Prize for reducing resolving armed conflicts around the world, and is next heading to Syria to strengthen the fragile peace there. Unfortunately, you realize that the second person is your own mother—a poor, uneducated housekeeper. Your mother is not important for world peace, but she was always a good mother to you. Now you must choose whether you should save the famous peace negotiator or your own mother in the few seconds before the building collapses.

How acceptable would it be for you to save your mother from the burning building, even though this means the famous peace negotiator will die instead of heading to Syria?

Second family dilemma. Your mother is feeling really lonely and would like you to come and spend the day with her. You would like to cheer up your mother, but you have already agreed to spend the day volunteering for Habitat for Humanity, a charity organization that builds houses for poor families. As Head Engineer, if you do not show up to the build site, the other volunteers will be legally unable to work, so they will have to go home. So, by volunteering, you would be helping a family in need for a long time, rather than cheering up your mother for one day.

How acceptable would it be for you to visit your mother to cheer her up, even though this means that no volunteer work will occur on the habitat for humanity house today?

First ingroup-outgroup dilemma. You are an American businessman who plans to donate $1000 to help sick children. You are currently deciding which of two charities you should donate to- you will give all the money to one charity but not both. One charity focuses on treating pediatric diseases in the United States, such as leukemia. If you donated $1000 to the American charity, your donation would purchase a drug treatment regimen that would save the life of one underprivileged American child with leukemia. The other charity focuses on preventing widespread diseases in Sub-Saharan Africa, such as measles. If you donated $1000 to the Sub-Saharan charity, your donation would purchase drug treatment regimens that would save the lives of 10 underprivileged African children with measles.

How acceptable would it be for you to donate to the American charity to save the life of one child with leukemia, even though this means you would not donate to the Sub-Saharan charity, so 10 children with measles will die?

Second ingroup-outgroup dilemma. You are a philanthropist who plans to donate $1000 to a homeless shelter. There are two homeless shelters near you: one in your hometown and one in the nearest big city (Cincinnati). The homeless shelter in Cincinnati is bigger and more efficient: a donation of $1000 would provide food and shelter for 50 individuals for a week. Your hometown has a smaller homeless shelter: a donation of $1000 would provide food and shelter for 10 individuals for a week. However, you feel a special connection to your hometown. It is where you grew up, and you want to see it taken care of.

How acceptable would it be for you to donate the $1000 to the homeless shelter in your own hometown, even though donating to the homeless shelter in Cincinnati would help many more homeless people?

First animal dilemma. You recently read a book describing the methods of modern factory farming in gruesome detail. The book persuaded you that by eating meat, people are supporting the factory farming industry, thereby causing many animals to undergo great suffering. Each time a person becomes vegetarian, it means one less customer for factory farming, which reduces animal suffering, so you consider becoming vegetarian. On the other hand, you think that humans are more important than animals, it is natural for humans to eat meat, and you really like eating meat.

How acceptable would it be for you to continue eating meat, even though you know this means animals will continue to suffer for your dining pleasure?

Second animal dilemma. You recently became the CEO of a company that develops and markets new forms of plastic surgery. You find out that the company uses animal experimentation with pigs and rabbits to test these new methods. These experiments cause great suffering for the animals, and often result in permanent disfigurement or death of the animals experimented on. On the other hand, these experiments do help your company improve the procedures, which in turn helps them make a bigger profit and provide better plastic surgery for people.

How acceptable would it be for you to allow the company to continue to engage in animal experimentation, even though you know this means animals will continue to suffer?