Cognition 179 (2018) 241–265

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers

Paul Conway[a,b,*], Jacob Goldstein-Greenwood[a], David Polacek[a], Joshua D. Greene[c]

[a] Florida State University, Department of Psychology, 1107 W. Call St., Tallahassee, FL 32306-4301, USA
[b] Social Cognition Center Cologne, Department of Psychology, University of Cologne, Richard-Strauss-Str. 2, Cologne 50931, Germany
[c] Harvard University, Department of Psychology, Center for Brain Science, 33 Kirkland St., Cambridge, MA 02138, USA

ARTICLE INFO

Keywords:
Moral dilemmas
Dual-processes
Process dissociation
Moral conviction
Trolley dilemmas

ABSTRACT

Researchers have used "sacrificial" trolley-type dilemmas (where harmful actions promote the greater good) to model competing influences on moral judgment: affective reactions to causing harm that motivate characteristically deontological judgments ("the ends don't justify the means") and deliberate cost-benefit reasoning that motivates characteristically utilitarian judgments ("better to save more lives"). Recently, Kahane, Everett, Earp, Farias, and Savulescu (2015) argued that sacrificial judgments reflect antisociality rather than "genuine utilitarianism," but this work employs a different definition of "utilitarian judgment." We introduce a five-level taxonomy of "utilitarian judgment" and clarify our longstanding usage, according to which judgments are "utilitarian" simply because they favor the greater good, regardless of judges' motivations or philosophical commitments. Moreover, we present seven studies revisiting Kahane and colleagues' empirical claims. Studies 1a–1b demonstrate that dilemma judgments indeed relate to utilitarian philosophy, as philosophers identifying as utilitarian/consequentialist were especially likely to endorse utilitarian sacrifices. Studies 2–6 replicate, clarify, and extend Kahane and colleagues' findings using process dissociation to independently assess deontological and utilitarian response tendencies in lay people. Using conventional analyses that treat deontological and utilitarian responses as diametric opposites, we replicate many of Kahane and colleagues' key findings. However, process dissociation reveals that antisociality predicts reduced deontological inclinations, not increased utilitarian inclinations. Critically, we provide evidence that lay people's sacrificial utilitarian judgments also reflect moral concerns about minimizing harm. This work clarifies the conceptual and empirical links between moral philosophy and moral psychology and indicates that sacrificial utilitarian judgments reflect genuine moral concern, in both philosophers and ordinary people.

1. Introduction

A substantial body of research in psychology, neuroscience, and experimental philosophy examines responses to sacrificial moral dilemmas where harmful actions promote the greater good. Such dilemmas have proven useful in studies of children, lesion patients, psychopaths, and research employing functional neuroimaging, psychophysiological measures, endocrinological measures, non-invasive brain stimulation, virtual reality, pharmacological interventions, genotyping, and behavioral measures (e.g., Amit & Greene, 2012; Bartels & Pizarro, 2011; Bernhard et al., 2016; Crockett, Clark, Hauser, & Robbins, 2010; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Koenigs et al., 2007; Mendez, Anderson, & Shapira, 2005; Montoya et al., 2013; Moore, Clark, & Kane, 2008; Nichols & Mallon, 2006; Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014; Pellizzoni, Siegal, & Surian, 2010; Wiech et al., 2013; see Greene, 2013, 2014 for reviews). Dilemma research has flourished because sacrificial dilemmas, including classic trolley dilemmas (Foot, 1967; Thomson, 1986), straddle a major fault line in human cognitive architecture, engaging competing processes (Cushman & Greene, 2012).

Philosophers originally crafted moral dilemmas to draw out the implications of competing philosophical positions. Rejecting harmful actions (even those that promote the greater good) is broadly consistent with a deontological moral philosophy (Kant, 1785/1959). Endorsing

* Corresponding author at: Florida State University, Department of Psychology, 1107 W. Call St., Tallahassee, FL 32306-4301, USA.
E-mail addresses: conway@psy.fsu.edu (P. Conway), jrg15b@my.fsu.edu (J. Goldstein-Greenwood), dap13c@my.fsu.edu (D. Polacek), jgreene@wjh.harvard.edu (J.D. Greene).

https://doi.org/10.1016/j.cognition.2018.04.018
Received 21 June 2017; Received in revised form 19 April 2018; Accepted 20 April 2018
Available online 02 July 2018
0010-0277/

such harmful actions is broadly consistent with a utilitarian/consequentialist moral philosophy (Mill, 1861/1998).[1] Hence, researchers have described such judgments as 'characteristically deontological/utilitarian' (Greene, 2007, 2014). According to Greene and colleagues' dual-process theory (Greene, 2007; Greene et al., 2001, 2013), when people consider causing harm to save lives, automatic emotional reactions to causing harm motivate characteristically deontological responses, whereas deliberate cost-benefit reasoning motivates characteristically utilitarian responses. For example, in the *footbridge* dilemma, where pushing a man in front of a trolley will save five lives (Thomson, 1986), negative emotional reactions lead people to disapprove of pushing him, whereas cost-benefit reasoning motivates people to approve of pushing him to save lives.

Recently, Kahane and colleagues (Kahane, 2015; Kahane et al., 2015) argued (a) against the use of sacrificial dilemmas, (b) against calling sacrificial responses "utilitarian," and (c) against the dual-process theory. They present these objections as following from an empirical discovery: That sacrificial judgments favoring the greater good are driven primarily—if not exclusively—by antisocial tendencies and therefore bear no meaningful relationship to utilitarian thought. For support, they cite a growing body of findings associating sacrificial judgments with antisocial tendencies, such as psychopathy, Machiavellianism, and narcissism (Bartels & Pizarro, 2011; Djeriouat & Trémolière, 2014), low empathic concern (Gleichgerrcht & Young, 2013), high testosterone (Carney & Mason, 2010), decreased aversion to causing harm (Miller, Hannikainen, & Cushman, 2014; Wiech et al., 2013), and intoxication (Duke & Bègue, 2015). Moreover, Kahane et al. (2015) found that sacrificial dilemma judgments fail to track other metrics of utilitarian values, such as donating to the poor.

However, Kahane and colleagues' critique rests upon two conceptual assumptions that we do not accept. The first concerns how researchers employ the term *utilitarian judgment*. According to Kahane and colleagues, a judgment qualifies as "utilitarian" only if it arises from motivations that are generally consistent with utilitarian philosophical principles. In other words, they assume that what makes a judgment "utilitarian" is the mindset of the judge. Kahane and colleagues assume that others share this understanding (p. 194), but we do not. This definition diverges from how the term was originally employed in the scientific literature: researchers explicitly applied the term "utilitarian" to *judgments* rather than *judges* (e.g., Amit & Greene, 2012; Greene, 2007; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, 2013, 2014). Understood this way, judgments that maximize good consequences are utilitarian *by definition*, regardless of the psychological processes that drive such judgments. One can make a utilitarian judgment without being a utilitarian or even having any generally utilitarian traits, just as one can make an Italian meal without being Italian or having generally Italian traits. Naturally, researchers may, if they prefer, reserve the term "utilitarian judgment" for judgments reflective of a general commitment to utilitarian values. However, if they do this, they should be clear that they are using this term in a more restrictive way than it was originally used in the scientific literature.

Second, we suggest that Kahane and colleagues mischaracterize the commitments of dual-process theory. Greene and colleagues (e.g., 2004) originally claimed that characteristically utilitarian decisions arise from ordinary cost-benefit reasoning—not from a general commitment to utilitarian values. What's more, because sacrificial dilemmas are thought to pit cost-benefit reasoning against negative emotional responses to causing harm, the dual-process theory actually *predicts* that antisocial tendencies will be associated with characteristically utilitarian judgments, assessed via conventional techniques (Koenigs et al., 2007; Mendez et al., 2005; Ciaramelli, Muccioli, Làdavas, & di Pellegrino, 2007).

Conceptual issues aside, Kahane and colleagues did raise a substantive empirical question: Do ordinary people's sacrificial utilitarian dilemma judgments reflect genuine concern for the greater good? They presented four studies suggesting the answer is "no:" utilitarian sacrificial dilemma judgments correlated positively with various measures of antisociality (e.g., psychopathy), but not with measures of prosociality (e.g., donating to the poor). Such findings raise questions about whether sacrificial utilitarian judgments ever reflect moral or prosocial considerations. However, Kahane and colleagues employed only conventional dilemma analyses that treat deontological and utilitarian responses as diametric opposites. Such analyses cannot distinguish genuine moral concern about aggregate outcomes from reduced concern about causing harm and may therefore distort the true relations between people's response tendencies and their psychological traits (Conway & Gawronski, 2013).

Replicated and, in some cases, extended of Kahane and colleagues' studies, employing process dissociation to assess utilitarian and deontological tendencies independently (Conway & Gawronski, 2013) and to clarify their relations with antisociality and prosociality. We also examined the judgments of philosophers to determine whether philosophers who endorse utilitarian/consequentialist principles typically endorse outcome-maximizing (utilitarian) harm in response to sacrificial dilemmas. Our findings suggest that sacrificial utilitarian judgments often do reflect moral concern for the greater good, both among philosophers and ordinary people.

1.1. What qualifies as a utilitarian judgment? A five-level taxonomy

The critique presented by Kahane (2015) and Kahane et al. (2015) has two layers. In the foreground are empirical results indicating that the ordinary people who approve of utilitarian sacrifices do not have generally utilitarian values and seem to be relatively antisocial. In their critique's background are assumptions about what it takes for a judgment to qualify as "utilitarian." These conceptual assumptions set the height of the bar for calling a judgment "utilitarian," while the empirical findings determine which judgments can clear the bar, given its height. These assumptions play an essential role Kahane and colleagues' critique, underwriting their conclusion that judgments routinely described as utilitarian are only "so called" utilitarian judgments. Before presenting our empirical results, it is important to clarify the meaning of the term "utilitarian judgment." Here we describe five levels at which one can set the bar for calling a judgment "utilitarian." We use this framework to clarify how our critical background assumptions differ from those of Kahane and colleagues and why we disagree with their critique.

1.1.1. Level 1: Utilitarian by judgment content

A judgment can be described as utilitarian simply because it favors the greater good, regardless of the mindset, intentions, or philosophical commitments of the judge. Whether or not a judgment qualifies as Level-1 utilitarian is a conceptual, not a psychological, matter. It reflects what utilitarian philosophy says about the judgment, not what the judgment says about the judge. Antisocial individuals can make Level-1 utilitarian judgments, even if they care not at all about the greater good (e.g., Koenigs et al., 2011). Such judgments are utilitarian because they coincide with what utilitarianism requires.

When we label a judgment as "utilitarian," we mean that it qualifies as Level-1 utilitarian by definition. Whether it qualifies as higher than Level-1 is a further empirical question. From the outset, Greene and colleagues described utilitarian judgments as "judgments that maximize aggregate welfare" (2004) and "approving of harmful actions that maximize good consequences" (2008). They never described such

[1] Consistent with Kahane et al. (2015), we confine our discussion of utilitarianism to *act utilitarianism*, according to which each action ought to promote the greater good, rather than other forms of utilitarianism, such as *rule utilitarianism*, according to which actions ought to conform to the rules that most reliably promote the greater good (see Kahane & Shackel, 2010). The sacrificial utilitarian judgments discussed here are required by act utilitarianism, but not necessarily by rule utilitarianism.

judgments as reflecting general moral commitments or traits. Greene has explicitly addressed this terminological issue (Amit & Greene, 2012, p. 867; Greene, 2007, p. 39), including in response to Kahane and colleagues (Greene, 2014, p. 699). Thus, following our longstanding usage, the sacrificial judgments in question are utilitarian judgments by definition, and not "so called" utilitarian judgments (Kahane et al., 2015), as they are the judgments required by utilitarian philosophy, regardless of decision-makers' motivations.

1.1.2. Level 2: Utilitarian through aggregate cost-benefit reasoning

Although the label "utilitarian judgment" implies no assumptions about the judge's reasons or motivations, the dual-process theory does make claims about psychological processes. It makes the empirical claim that ordinary people's utilitarian judgments are typically the products of deliberate, aggregate cost-benefit reasoning (Greene et al., 2001, 2004, 2013). Consistent with this claim, people's judgments are sensitive to aggregate consequences: People who endorse killing one to save five typically do not endorse killing five to save one, or killing one to save zero.[2] Kahane and colleagues agree that people's sacrificial judgments often qualify as Level-2 utilitarian, reflecting the, "modest, unremarkable, and ordinary thought that it is, ceteris paribus, morally better to save a greater number" (Kahane et al., 2015, pp. 206–7). Although they present this claim as an alternative account, it remains perfectly consistent with the dual-process account.

1.1.3. Level 3: Utilitarian through concern for the greater good

One can make level-1 or level-2 utilitarian judgments without actually caring about the greater good, as psychopaths appear to do (Koenigs, Kruepke, Zeier, & Newman, 2011). But when ordinary people approve of utilitarian sacrifices, or hesitate before disapproving, they may do so out of genuine concern for the greater good *within the context of the decision* (Greene, 2007, 2013), if not out of a general commitment to utilitarian values. Level-3 utilitarian judgment marks the point where our empirical conclusions diverge from those of Kahane and colleagues. Examining various measures of antisociality and prosociality, they find no evidence for genuine moral concern influencing ordinary people's sacrificial utilitarian judgments, and therefore suggest that such judgments "merely express a calculating yet selfish mindset" (p. 197). Below, we provide evidence against this strong claim using process dissociation. Our findings, however, stop short of implying that sacrificial judgments reflect a general commitment to utilitarian values.

1.1.4. Level 4: Utilitarian through a general commitment to utilitarian values

Some people's concern for the greater good may go beyond merely caring about aggregate outcomes in specific contexts. They may instead adopt, to some degree, the broader moral commitments championed by utilitarian philosophers, emphasizing the well-being of socially distant humans and animals. The instruments that Kahane et al. (2015) used to measure "genuine utilitarian impartial concern for the greater good" (p. 193) are essentially tests for Level-4 utilitarian judgment. They find no evidence that ordinary people's sacrificial utilitarian judgments are motivated by general commitments to utilitarian values, and neither do we. However, we know of no one who claims that ordinary people's sacrificial judgments are motivated by such commitments.

1.1.5. Level 5: Explicit commitment to utilitarian values

Finally, sacrificial utilitarian judgments may reflect explicit commitments to utilitarian principles. Neither we nor Kahane and colleagues (p. 207) expect ordinary participants to make level-5 utilitarian judgments. Level-5 utilitarian judgment is relevant, however, because of Studies 1a–1b examining the judgments of philosophers, who can be expected to make level-5 utilitarian judgments. Their judgments speak to the relationship between utilitarian philosophy and sacrificial dilemmas.

1.1.6. Taxonomy implications

This taxonomy offers five key takeaways, one for each level. First, ordinary people's sacrificial utilitarian judgments should not be described as "so called" utilitarian. They are, by definition, Level-1 utilitarian judgments—judgments favoring the greater good and therefore required by utilitarianism. This claim is definitional, not empirical. Second, the dual-process theory makes the further empirical claim that ordinary people's Level-1 utilitarian judgments also qualify as Level-2 utilitarian judgments, reflecting cost-benefit reasoning. This is a point of empirical agreement (Kahane et al., 2015, p. 207). Third, the most controversial empirical question, and our present focus, concerns Level-3 utilitarian judgment—sacrificial judgments motivated by genuine concern about aggregate outcomes. Kahane and colleagues' most provocative suggestion is that ordinary people's utilitarian judgments do not rise to Level 3—that such judgments "merely express a calculating yet selfish mindset" (p. 197). We provide evidence against this claim below.

Fourth, Kahane et al. set the bar for calling a judgment "utilitarian" at Level 4, requiring that sacrificial judgments reflect a general commitment to utilitarian values. They provide evidence that ordinary people's judgments do not rise to this level, and our findings corroborate this claim. This is not surprising, however, as neither we nor anyone else has ever claimed that ordinary people's approval of utilitarian sacrifices reflects a general commitment to utilitarian values. Nevertheless, some researchers in addition to Kahane et al. (Rosas & Koenigs, 2014; Royzman, Landy, & Leeman, 2015; Sheskin & Baumard, 2016) seem to be operating on the assumption that a judgment is not "utilitarian" unless it is Level-4 utilitarian, reflecting a consistent commitment to utilitarian values. Consensus returns for the fifth key takeaway: whereas philosophers may make Level-5 utilitarian judgments (reflecting an explicit commitment to utilitarianism or related theories), no one expects this from ordinary respondents.

As noted above, there is a large discrepancy between our definition of utilitarian judgment (Level-1 or higher) and that of Kahane and colleagues (Level-4 or 5 only). In the General Discussion, we consider reasons for, and implications of, this divergence. For now, we turn to our primary empirical question concerning Level-3 utilitarian judgment: Do sacrificial judgments reflect concern for the greater good?

1.2. Process dissociation and the detection of competing moral influences

1.2.1. Conventional analyses are ambiguous

Kahane et al. (2015) claim that ordinary people's sacrificial utilitarian judgments "do not reflect impartial concern for the greater good" (p. 193) and suggest that such judgments "merely express a calculating yet selfish mindset" (p. 197). The first claim, made in the paper's title, can be read as merely denying that people's sacrificial judgments rise to Level 4 (reflecting a general commitment to utilitarian values). However, the second claim appears to rule out even Level-3 utilitarian judgment, implying that such judgments involve no genuine moral concern at all (i.e., merely qualify as Level-2).

However, both of these claims rest upon research employing conventional sacrificial moral dilemmas that pit concerns about causing harm against concerns for the greater good. Such analyses remain ambiguous with respect to people's motivations and traits. Level-1 utilitarian responses on conventional dilemmas may reflect either prosocial tendencies, a *relatively* strong desire to promote the greater good, or antisocial tendencies, a *relatively* weak desire to avoid harming people. Conventional analyses cannot distinguish between these possibilities. Thus, although evidence abounds that utilitarian sacrificial judgments are associated with antisocial traits (e.g., Bartels & Pizarro,

[2] For example, a majority of people in the current studies endorsed causing harm in the incongruent version of the car accident dilemma (2.2) where causing harm would save lives overall, but rejected causing harm in the congruent version (2.1) where causing harm would not save lives overall.

2011; Patil, & Silani, 2014; Miller et al., 2014), it remains unclear whether such findings truly reflects the psychology involved in maximizing good outcomes, or simply the absence of concerns about causing harm.

Moreover, further evidence is required to support the stronger claim that sacrificial judgments do not reflect prosocial tendencies—an argument based on null findings. Although null findings can indicate the absence of the effect, they can also result from *suppression*—the case where two same-direction effects cancel out when pitted against one another. There are reasons to believe that just such a suppression effect may occur for the relationship between measures of prosociality and conventional dilemma judgments. For example, Conway and Gawronski (2013) found that moral identity internalization positively predicted both utilitarian and deontological response tendencies, which cancelled out for relative judgments. Reynolds and Conway (2018) found a similar pattern for aversion to witnessing others' suffering, and many other papers have documented simultaneous influences on dilemma responding that remain invisible to conventional dilemma analyses (e.g., Conway, Weiss, Burgmer, & Mussweiler, 2018; Muda, Niszczota, Bialek, & Conway, 2017). Therefore, conventional analyses may underestimate the extent to which sacrificial utilitarian judgments reflect prosocial motivations.[3]

1.2.2. Process dissociation clarifies ambiguity

To overcome the limitations of conventional analyses, we employed process dissociation (PD) to independently assess harm rejection and outcome-maximization response tendencies that contribute to dilemma responses. Thus, PD allows one to assess, for a given participant, whether her sacrificial (Level-1) utilitarian judgments (if present) reflect increased concern for aggregate outcomes or decreased concern for causing harm. Jacoby (1991) developed PD to examine memory performance, but PD has provided insight into a growing range of topics (Payne & Bishara, 2009), including moral dilemmas (Conway & Gawronski, 2013). The insight behind PD is that when two response tendencies jointly contribute to an outcome, one must assess both decisions where tendencies converge as well as decisions where they compete to clarify how each tendency operates.

Thus, PD employs *incongruent* and *congruent* dilemmas. Incongruent dilemmas, like most familiar trolley cases, pit concerns about maximizing overall outcomes against concerns about causing harm. In the matched *congruent* dilemmas, the same harmful action cannot be justified (or is harder to justify) on utilitarian grounds, but is equally unappealing from a deontological perspective. Yet, in congruent dilemmas there remain nonmoral/antisocial reasons to accept harm, such as self-interest, vengeance, or sadism. In one incongruent dilemma, a driver may swerve into and kill one elderly person to avoid killing a mother and her child. In the congruent version of this dilemma, the driver may swerve into and kill several children to avoid killing the mother and child.[4] Then, PD involves applying participant responses to both sets of dilemmas to a processing tree (see Fig. 1), and calculating two parameters for each participant (see Appendix A). These parameters reflect the tendency to reject causing harm regardless of outcomes (i.e., deontological inclinations), and the tendency to maximize good outcomes regardless of whether doing so entails causing harm (i.e., utilitarian inclinations).[5]

PD findings suggest that the utilitarian (U) and deontological (D) parameters reflect independent contributions to a joint outcome, because they both correlate robustly with conventional sacrificial judgments, yet they themselves typically remain uncorrelated—a finding confirmed meta-analytically (Friesdorf, Conway, & Gawronski, 2015). In addition, the utilitarian and deontological PD parameters appear to largely align with the processes posited by the dual-process theory (Greene et al., 2001; Greene et al., 2004; Greene, 2013). For example, the utilitarian parameter uniquely predicted need for cognition (Cacioppo & Petty, 1982) and was uniquely impaired by cognitive load, while the deontological parameter uniquely predicted empathic concern and was uniquely increased by making harm salient (Conway & Gawronski, 2013). Evidence from other laboratories largely corroborates this distinction (e.g., Christov-Moore, Conway, & Iacoboni, 2017; Lee & Gino, 2015; Park, Kappes, Rho, & Van Bavel, 2015).

That said, the picture is undoubtedly more complex—not all labs have replicated such effects (e.g., Gawronski, Conway, Friesdorf, Armstrong, & Hütter, 2017), some find links between deliberative processing and deontological tendencies (e.g., Gamez-Djokic & Molden, 2016), and others demonstrate links between affective concerns and utilitarian tendencies (e.g., Reynolds & Conway, 2018), and such models ignore higher-order processes like strategic self-presentation (Rom & Conway, 2018). Yet, the preponderance of evidence suggests that the dual-process model is not so much incorrect as merely incomplete. Although other processes not modeled may play a role, ultimately deontological responses seem reflect relatively more affective processing centered on harmful actions, whereas utilitarian responses appear reflect relatively more deliberative reasoning centered on outcomes.[6]

Most important for present purposes, PD has demonstrated greater sensitivity than conventional judgments. For example, Conway and Gawronski (2013) found that both parameters positively predicted moral identity internalization, or the centrality of morality to the self-concept (Aquino & Reed, 2002)—and these dual positive relationships cancelled out in conventional judgments that pit deontological response inclinations directly against utilitarian ones (i.e., a suppression effect). Likewise, Reynolds and Conway (2018) demonstrated that concerns about witnessing harm (outcome aversion) positively predicted both deontological and utilitarian tendencies (which cancelled out for conventional judgments), whereas concerns about causing harm (action aversion) predicted increased deontological but reduced utilitarian tendencies (see Miller et al., 2014). Moreover, many other studies demonstrate suppression effects where a manipulation simultaneously impacts both PD parameters and largely cancels out for conventional analyses (e.g., Conway et al., 2018; Hayakawa, Tannenbaum, Costa, Corey, & Keysar, 2017; Muda et al., 2017). Consistent with these findings, we anticipated that PD analyses in the current work would prove more sensitive than conventional dilemma analyses, revealing relationships between sacrificial responses and prosocial concerns that are difficult to detect using conventional correlational methods, thereby clarifying the role of antisocial and prosocial motivations in utilitarian judgments.

[3] Note that Kahane et al. (2015) sometimes controlled statistically for antisocial tendencies (e.g., p. 196) in their analyses, but unfortunately this technique is insufficient because measure of dilemma responding itself remains ambiguous with regard to suppression effects. When suppression occurs, there will be no correlation between a given measure and conventional dilemma judgments regardless of whether one controls for other measures or not. Hence, controlling for antisociality does nothing to address this concern, and therefore cannot reveal whether or not there is a correlation between utilitarian responding and prosociality.

[4] Technically, congruent dilemmas are more accurately described as "decision scenarios" than dilemmas, as only one option qualifies as moral, similar to "low-conflict" (Greene et al., 2008; Koenigs et al., 2007) personal dilemmas introduced by Greene et al. (2001).

[5] The utilitarian parameter may be described as the tendency to minimize negative outcomes whether or not doing so entails causing harm—i.e., making a series of Level-1 utilitarian judgments, some of which entail causing harm, and others which entail refraining from directly causing harm. This pattern suggests a sensitivity to outcomes, suggesting that such a pattern reflects, at minimum, Level-2 utilitarian judgment. It remains an empirical question whether such responses also qualify as Level-3 or higher.

[6] Further evidence for this pattern emerged in Study 6, where the D parameter selectively correlated with concern for the individual to be harmed, whereas the U parameter selectively correlated with concern for the overall group (see Robinson, Joel, & Plaks, 2015).

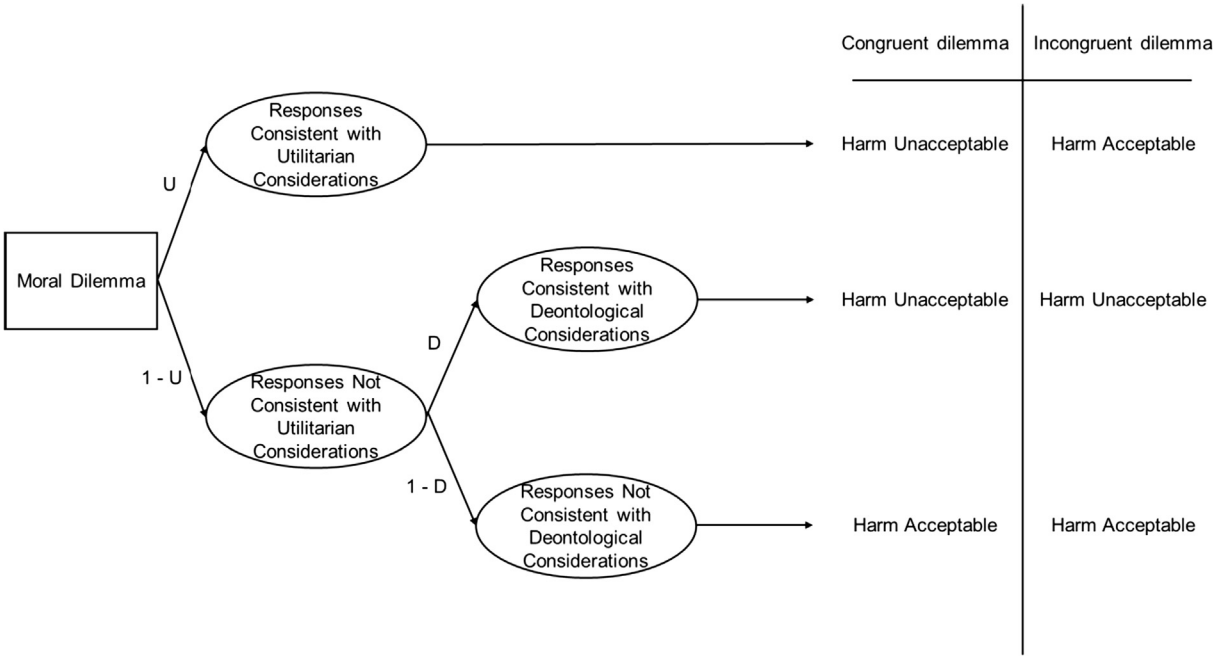

Fig. 1. Processing tree illustrating the components underlying responses to congruent and incongruent moral dilemmas employed in Studies 2–6.

1.3. Overview of the current work

We hypothesized that utilitarian judgments reflect genuine concern for the greater good, but that this relationship is often masked by a confounding relationship between antisocial tendencies and utilitarian judgment, driven by an "un-deontological" indifference to causing harm. We tested this hypothesis using two methods.

First, we analyzed data from two large surveys of professional philosophers to determine whether philosophers who endorse utilitarian/consequentialist ethics are more likely to make sacrificial utilitarian judgments in response to moral dilemmas (Studies 1a and 1b). We predicted that the judgments of philosophers will reveal a robust relationship between the endorsement of utilitarian/consequentialist philosophy and sacrificial utilitarian judgments. If this prediction is confirmed, it will establish a link between sacrificial utilitarian judgments and genuine concern for the greater good, on the assumption that philosophers who endorse utilitarian sacrifices do so out of genuine moral concern and are not merely exhibiting a calculating, selfish mindset.

Second, to clarify the relationship between utilitarian judgment and antisociality in ordinary people, we repeated all four studies conducted by Kahane et al. (2015), but replaced their conventional dilemma battery with the Conway and Gawronski (2013) PD dilemma battery (Studies 2–5). Study 6 repeats Kahane et al.'s Study 4 a second time, using modified materials that allow us to more precisely describe the nature of ordinary people's utilitarian thinking. The PD dilemma battery used in Studies 2–6 allows researchers to calculate conventional dilemma judgments that pit acceptance of outcome-maximizing harm (characteristic of utilitarianism) against rejection of harm (characteristic of deontology). We expected that such analyses would replicate many of Kahane and colleagues' findings, demonstrating links between antisociality and accepting sacrificial harm.

However, the PD dilemma battery also provides independent estimates of the utilitarian and deontological parameters that jointly contribute to conventional dilemma judgments. We expected this analysis to paint a more nuanced picture: Consistent with meta-analytic findings, we expected to find evidence for distinct utilitarian (U) and deontological (D) parameters that jointly contribute to conventional dilemma judgments (Friesdorf et al., 2015). We expected most of the variance in antisociality to load negatively on the deontological parameter, indicating a lack of concern about causing harm among people high in psychopathy and other antisocial traits. However, we anticipated that antisociality would fail to predict, or negatively predict, utilitarian inclinations assessed independently of low deontological inclinations. Moreover, we expected that general measures of moral concern, such as moral identity internalization (Aquino & Reed, 2002) and moral conviction about harm (see Skitka, Bauman, & Sargis, 2005), would demonstrate suppression effects: they would positively predict both the deontological and utilitarian parameters, suggesting that both response tendencies reflect genuine moral motivations. Likewise, consistent with the dual-process model, we expected the utilitarian parameter to correlate positively with a focus on group harm, but not individual harm, and the reverse for the deontological parameter. Such findings would suggest that utilitarian responses to sacrificial moral dilemmas reflect genuine concern for the greater good (Level-3 utilitarian judgment), and such findings would emerge only if one uses process dissociation to account for the influence of suppression effects.

In Studies 2–6, we present bivariate correlations between conventional dilemma judgments, the U and D process dissociation parameters, and various trait and decision measures. We also conducted analyses regressing each measure on both the U and D parameters simultaneously, while controlling for gender and age. As these regression analyses yielded findings quite similar to the correlational results, we present results of regression analyses in the Supplementary Materials. All data and analyses are available at osf.io/8vdaj.

2. Studies 1a and 1b

Research in moral psychology has focused on trolley problems and other sacrificial dilemmas because they are thought to nicely capture the central tension between the utilitarian/consequentialist and deontological schools of philosophical thought (e.g., Greene, 2007, 2013, 2014). Kahane and colleagues claim that sacrificial trolley judgments have "little (or nothing)" to do with utilitarianism (Kahane, 2015, p. 551), and that there is "very little relation between sacrificial judgments in the hypothetical dilemmas that dominate current research, and a genuine utilitarian approach to ethics" (Kahane et al., p. 193). There are two interpretations of this objection. First, one might object

to studying the judgments *of ordinary people*, whose thought processes may bear little resemblance to those of philosophers. Second, one might object to studying sacrificial dilemmas more generally, claiming that they are of limited philosophical relevance for both ordinary people and philosophers. Kahane et al. (2015) raise the first objection, but Kahane (2015) goes further and makes the second objection, claiming that sacrificial dilemmas have little to do with the "grand questions" (p. 552) of moral philosophy.

Kahane notes that trolley dilemmas originated as part of a debate within the deontological school (Foot, 1967; Thomson, 1986) and says, "To the extent that the aim of this recent empirical research on moral dilemmas is to use the hypothetical cases that most sharply divide utilitarians and their opponents, then this research may be focusing on the wrong examples." (p. 552). We disagree. Like the authors of many introductory philosophy textbooks (MacKinnon & Fiala, 2014, pp. 99–109; Sandel, 2010, pp. 21–24; Vaughn, 2012, pp. 84–95), we suggest that sacrificial dilemmas do, in fact, relate to the "grand questions" that divide consequentialists and deontologists. If we are correct, then philosophers' sacrificial dilemma decisions should systematically relate to whether they identify as consequentialists, deontologists, or something else. Kahane's argument would suggest no such systematic relationships. Studies 1a–1b test these predictions. Moreover, Studies 1a–1b are relevant for Kahane and colleagues' more plausible claim that ordinary people's dilemma judgments are philosophically irrelevant—a point we return to in the General Discussion.

2.1. Method

2.1.1. Participants and procedure, Study 1a

We drew the first sample from *The Philosophical Survey* conducted in 2009 by Bourget and Chalmers (2014). They examined the views of 931 professional philosophers, drawn from the faculties of 99 leading philosophy departments (77.2% male, 17.4% female, 5.3% unspecified). Of these respondents, 753 responded to two questions that are essential for present purposes, one concerning their general views on normative ethics and one posing a compact version of the trolley *switch* dilemma. Participants also responded to demographic questions and a range of questions regarding other philosophical topics, not considered here. For details, see Bourget and Chalmers (2014).

The general normative ethics question asked respondents whether they *accept* or *lean toward* each of the three main theories of Western normative ethics: consequentialism, deontology, and virtue ethics. To indicate their views, respondents selected as many of the following options as applied to them: *Accept: consequentialism*, *Lean toward: consequentialism*, *Accept: deontology*, *Lean toward: deontology*, *Accept: virtue ethics*, *Lean toward: virtue ethics*. Following Bourget and Chalmers, we gave each participant a score indicating their degree of endorsement of (and fidelity to) consequentialism versus other theories. Participants received +2 points for accepting consequentialism, +1 for leaning toward it, −1 for leaning toward one of the other two theories, and −2 for accepting one of the other two theories. These point values were then summed. For example, a participant who accepted consequentialism (+2), but also leaned toward virtue ethics (−1), obtained a final consequentialism score of +1. As this analysis tracks the combined endorsement of multiple theories, we also conducted an analysis examining categorical endorsement of utilitarianism by sorting participants into three exclusive categories: (1) accepting or leaning toward consequentialism exclusively, (2) accepting or leaning toward another theory exclusively, or (3) accepting or leaning toward more than one theory.

The compact trolley dilemma read as follows: Trolley problem (five straight ahead, one on side track, turn requires switching): straight or turn? Participants responded by selecting one or more of the following: accept: straight (−2), lean toward: straight (−1), lean toward: turn (+1), and accept: turn (+2). As above, participants could select multiple answers, and the original researchers summed these values, such that higher scores reflect greater endorsement of killing one person to save five lives.

2.1.2. Participants and procedure, Study 1b

We obtained a second (unpublished) data set collected by Fiery Cushman & Eric Schwitzgebel. They recruited 2466 participants through emails to philosophy and non-philosophy departments at 25 major research universities with strong Ph.D. programs in philosophy (and a few additional participants through academic blogs). Here we focus on the 273 participants who identified themselves as holding an advanced degree (M.A. or Ph.D.) in philosophy. Participants characterized their normative ethical views by selecting *Consequentialism* ($n = 66$), *Deontology* ($n = 74$), or *Virtue Ethics* ($n = 92$). As 41 participants selected *None of the Above* instead, we examined data from the remaining 232 participants (169 males, 63 females, $M_{age} = 32.83$, $SD = 8.71$). Participants completed the study via the Moral Sense Test website (www.moralsensetest.com). In addition to the normative ethics question, the study examined responses to 17 moral scenarios, including three sacrificial moral dilemmas (*submarine*, *vaccine*, and *ecologists/safari*) that were adapted from the original set of *personal* dilemmas (Greene et al., 2001) and subsequently categorized as *high conflict* (Koenigs et al., 2007). We restrict our analyses to these three dilemmas. Each dilemma describes an action that causes harm but maximizes overall good outcomes. Participants evaluated the moral quality of the action using a scale ranging from 1 (*extremely morally good*) to 7 (*extremely morally bad*). We reverse coded responses so that higher scores indicate that participants evaluate causing harm that maximizes good outcomes more positively.

2.2. Results

2.2.1. Study 1a

As predicted, the degree to which philosophers accepted turning the trolley (thereby killing one to save five) correlated robustly with endorsement of consequentialist normative ethics, $r(753) = .29$, $p < .001$. To further explore this relationship, we examined whether the proportion of philosophers endorsing consequentialism would vary across dilemma judgments (see Fig. 2). If dilemma judgments reflect ethical views, then as dilemma judgments increase from accepting keeping the trolley straight (more deontological) to accepting turning the trolley (more utilitarian), the proportion of consequentialists should increase, the proportion of non-consequentialists should decrease, and the proportion of philosophers who endorse a mixture of normative ethics should peak in the middle.

To conduct this analysis, we created three dummy codes representing endorsing versus not endorsing each of our three philosophical categories (consequentialist, non-consequentialist, and mixed), and subjected each to one-way ANOVAs across the five levels of dilemma judgment. As predicted, the proportion of philosophers who accepted or leaned toward consequentialism significantly increased across the five levels of dilemma judgment, $F(4, 752) = 17.97$, $p < .001$, $\eta_p^2 = .08$. Post-hoc comparisons indicated that the proportion of consequentialists was higher among those who fully accepted turning the trolley ($M = .41$, $SD = .49$), than among those only leaning toward turning ($M = .24$, $SD = .43$), $M_{diff} = 0.17$, $SE = .04$, $p < .001$. In turn, the proportion of consequentialists was higher among those leaning toward turning the trolley than among those who arrived at mixed judgments ($M = .08$, $SD = .27$), $M_{diff} = 0.17$, $SE = .05$, $p = .002$, who, in turn, did not differ from the proportion of consequentialists who leaned toward ($M = .10$, $SD = .31$), $M_{diff} = -0.03$, $SE = .08$, $p = .737$, or fully accepted, keeping the trolley straight ($M = .06$, $SD = .24$), $M_{diff} = 0.01$, $SE = .07$, $p = .848$.

Conversely, the proportion of philosophers who accepted or leaned toward other views (non-consequentialists) significantly decreased across the five levels of dilemma judgment, $F(4, 752) = 13.32$, $p < .001$, $\eta_p^2 = .07$. Post-hoc comparisons indicated that the

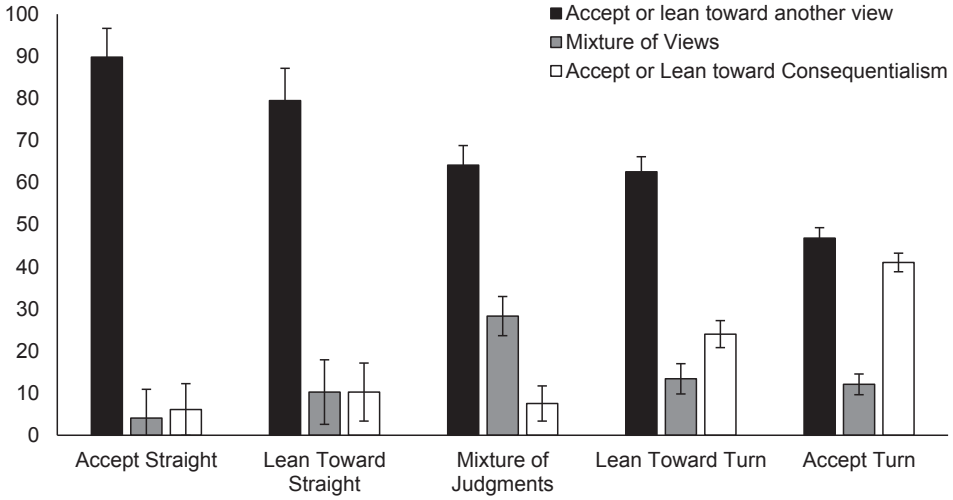

Fig. 2. Proportion of philosophers endorsing consequentialism, another view, or a mixture of views as a function of responses to the trolley *switch* dilemma in Study 1a. Accepting turning the trolley maximizes the number of lives saved, consistent with utilitarianism/consequentialism. Error bars reflect standard errors.

proportion of non-consequentialists was similar among those who fully accepted (M = .89, SD = .30), or leaned toward keeping the trolley straight (M = .79, SD = .40), M_{diff} = 0.10, SE = .10, p = .316, but the proportion of non-consequentialists was higher in the former group than among those who arrived at mixed judgments (M = .64, SD = .48), M_{diff} = 0.26, SE = .08, p = .002, whereas the latter two groups did not differ, M_{diff} = 0.15, SE = .09, p = .088. The proportion of non-consequentialists among those who arrived at mixed judgments was not different from those who leaned toward turning the trolley (M = .47, SD = .50), M_{diff} = 0.02, SE = .06, p = .788, but was higher than among those who fully accepted turning the trolley (M = .63, SD = .49), M_{diff} = 0.17, SE = .05, p = .001. Finally, the proportion of non-consequentialists among those who leaned toward turning the trolley was higher than among those who fully accepted turning the trolley, M_{diff} = 0.16, SE = .04, p < .001.

Moreover, the proportion of those who endorsed a mixture of normative ethical positions also varied systematically across condition, $F(4, 752) = 6.03$, p < .001, η_p^2 = .03. However, unlike the directional effects exhibited by consequentialists and non-consequentialists, post hoc tests indicated that the proportion of philosophers who endorsed a mixture of normative ethics positions was higher among those who made a mixture of dilemma judgments (M = .28, SD = .45) than among any other group (next highest M = .13, SD = .34, all p's < .006), none of which significantly differed from one another (all p's > .093).

The relationship between endorsement of consequentialism and dilemma decisions was most pronounced among those whose dilemma responses were unequivocal: Among the 380 philosophers who fully accepted turning the trolley, 41% (156 people) identified themselves as leaning toward or fully accepting consequentialism. Among the 49 philosophers who fully rejected turning the trolley, only 6% (3 people) identified themselves as leaning toward or fully accepting consequentialism. Thus, philosophers who fully endorsed the utilitarian response on the trolley dilemma were nearly seven times more likely to endorse consequentialism than philosophers who fully rejected that response.

2.2.2. Study 1b

To examine endorsement of outcome-maximizing harm (upholding utilitarianism), we conducted a 3 (philosophical endorsement: deontology, virtue ethics, or consequentialism) × 3 (dilemma: submarine, virus, or safari) repeated measures ANOVA, where the first factor was between-subjects and the second was within-subjects (see Fig. 3). We obtained the predicted main effect of endorsement, $F(2, 229) = 25.61$, p < .001, η_p^2 = .18, as well as a theoretically uninteresting main effect

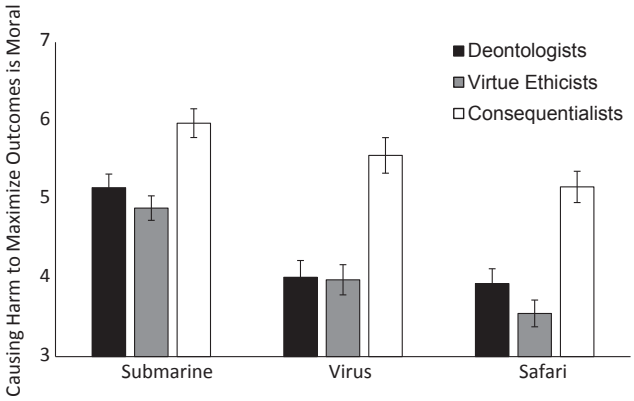

Fig. 3. Philosophers who identified as consequentialists (as compared to deontologists and virtue ethicists) rated outcome-maximizing harm (upholding utilitarianism) as more morally acceptable across all three dilemmas in Study 1b. Error bars reflect standard errors.

of dilemma, $F(2, 228) = 61.09$, p < .001, η_p^2 = .35. The interaction was not significant, $F(4, 458) = 2.01$, p = .093, η_p^2 = .02.

Post-hoc tests indicated that for each dilemma, philosophers who identified as consequentialists were more willing to accept outcome-maximizing harm than those who identified as either deontologists or virtue ethicists. Specifically, on the submarine dilemma, consequentialists reported greater endorsement of the action (M = 5.97, SD = 1.37) than either deontologists (M = 5.15, SD = 1.46), M_{diff} = 0.82, SE = .25, p = .001, or virtue ethicists (M = 4.87, SD = 1.56), M_{diff} = 1.10, SE = .24, p < .001, whereas the latter two groups did not significantly differ, M_{diff} = 0.28, SE = .23, p = .230. On the virus dilemma, consequentialists reported greater endorsement of the action (M = 5.56, SD = 1.44) than either deontologists (M = 4.01, SD = 2.00), M_{diff} = 1.55, SE = .31, p < .001, or virtue ethicists (M = 3.95, SD = 1.96), M_{diff} = 1.61, SE = .30, p < .001, whereas the latter two groups did not significantly differ, M_{diff} = 0.07, SE = .29, p = .814. Finally, on the safari dilemma, consequentialists reported greater endorsement of the action (M = 5.17, SD = 1.61) than either deontologists (M = 3.93, SD = 1.64), M_{diff} = 1.23, SE = .28, p < .001, or virtue ethicists (M = 3.52, SD = 1.62), M_{diff} = 1.65, SE = .26, p < .001, whereas the latter two groups did not significantly differ, M_{diff} = 0.41, SE = .25, p = .107.

Among the 50 participants who most consistently accepted sacrificial judgments favoring the greater good (mean ratings 6–7), 62%

identified as consequentialists. Among the 11 participants who most consistently rejected such sacrifices (mean ratings 1–2), only 9% (1 participant) identified as a consequentialist. Among the 172 participants who generally tended to approve of such sacrifices (mean rating ≥ 4.0), 37% identified as consequentialists, whereas among the 59 participants who tended to reject such sacrifices (mean rating < 4.0), only 3% (2 participants) identified as consequentialists.

2.3. Discussion

Across two separate data sets examining the sacrificial dilemma judgments of hundreds of philosophical experts, a clear pattern emerged: Those who endorsed consequentialist ethics were far more likely to accept causing harmful actions that promote the greater good than their peers who did not endorse consequentialism. This pattern held across multiple dilemmas,[7] and emerged most clearly for those whose dilemma decisions were unequivocal. Indeed, philosophers who most consistently made utilitarian judgments on sacrificial dilemmas were nearly seven times more likely to endorse consequentialist normative ethics than philosophers who reliably rejected outcome-maximizing harm.

These findings stand in tension with Kahane's (2015) suggestion that sacrificial dilemmas, including trolley dilemmas, are the "wrong examples" to study if one is interested in understanding the psychology behind the "grand questions" of moral philosophy, including "the division between utilitarianism and deontology" (p. 552).[8] Instead, we observe a robust relationship between how philosophers respond to sacrificial dilemmas, including the trolley *switch* case, and the extent to which they identify as consequentialist, deontologists, or virtue ethicists. These findings indicate a clear psychological link, in the minds of philosophers, between their thinking about sacrificial dilemmas and their general philosophical views.

These findings do not tell us whether ordinary people, in wrestling with these dilemmas, experience something related to the philosophical tension between consequentialism and deontology. We expect few ordinary people to endorse utilitarian sacrifices out of a general commitment to utilitarian values (Level 4), and even fewer out of an explicit commitment to utilitarianism (Level 5). However, we hypothesize that ordinary people often endorse utilitarian sacrifices out of a genuine concern for the greater good within the context of the dilemma (Level 3) and that such judgments do not merely reflect antisocial, calculating selfishness. The remaining experiments address this hypothesis.

3. Study 2

Studies 2–6 revisit Kahane and colleagues' claim that utilitarian judgments in sacrificial moral dilemmas do not reflect genuine concern for the greater good, and instead merely reflect antisocial tendencies. They presented four studies comparing participants' sacrificial dilemma responses to measures of antisociality (e.g., psychopathy) and prosociality (e.g., charity donations). Their findings suggest that utilitarian dilemma judgments correlate well with antisociality and poorly with prosociality, therefore suggesting that such judgments reflect lack of concern for causing harm rather than concern for maximizing overall outcomes. In this regard, their findings accord with a growing body of work linking utilitarian dilemma judgments to antisociality (e.g., Bartels & Pizarro, 2011; Carney & Mason, 2010; Djeriouat & Trémolière, 2014; Duke & Bègue, 2015; Gleichgerrcht & Young, 2013; Miller et al., 2014; Wiech et al., 2013).[9] We expected to replicate most of these findings when using similar dilemma measurement techniques.

However, all of these findings employ conventional dilemma analytic techniques that may obscure the true relationships between utilitarian dilemma responding, antisociality, and prosociality. Conventional analyses entail treating deontological responses as the pure inverse of utilitarian responses, and thus cannot distinguish high levels of concern for maximizing outcomes (high U) from low levels of concern about causing harm (low D). Therefore, it remains possible that the overall positive correlation between utilitarian judgment and antisocial traits actually reflects the absence of concerns about harm (i.e., low deontological response tendencies) rather than concerns about achieving the best overall outcomes (i.e., high utilitarian tendencies). Moreover, the presence of prosocial motivations (high levels of concern for maximizing good outcomes) may be masked by the presence of confounding antisocial motivations (low levels of concern about causing harm).

To examine these possibilities, we replicated each of Kahane et al.'s (2015) four studies (Studies 2–5), and conducted an additional study (Study 6), using process dissociation to independently assess utilitarian and deontological response tendencies underlying conventional dilemma judgments. Using conventional analyses, we anticipated replicating most of the links between antisociality and utilitarian dilemma judgments documented by Kahane and others. We anticipated that a PD analysis would reveal that these effects are due to negative relations between antisociality and deontological inclinations, which manifest in conventional analyses as positive relationships between antisociality and utilitarian judgments. Critically, we predicted that, using PD, utilitarian inclinations would correlate *negatively* with some measures of antisociality, suggesting that conventional analyses distort the true underlying relationship between antisociality and concern for maximizing outcomes. Finally, (in later studies) we predicted that some measures of prosociality would correlate positively with both utilitarian and deontological response tendencies, and cancel out for conventional analyses (a suppression effect). In other words, we predicted that conventional analyses would (a) overestimate the relationship between antisociality and utilitarian response tendencies, and (b) underestimate the relationship between prosociality and utilitarian response tendencies.

3.1. Study 2 methods

3.1.1. Participants

We recruited 182 participants via Amazon's Mechanical Turk (Amazon, 2015), who received $1.50 for participating. We excluded two participants who failed to complete all dilemmas (following Conway & Gawronski, 2013), and eight participants who failed an attention check item requiring a specific response regarding ambient air

[7] We note that Study 1a, unlike Study 1b, focused exclusively on the "impersonal" (Greene et al., 2001, 2009) *switch* dilemma and did not include any "personal" dilemmas. In addition, Study 1b did not include the best-known personal dilemma, i.e. the *footbridge* case. We would like to have tested the *footbridge* case specifically, but were unable to choose the dilemmas included in these data sets. With that said, the dilemmas included provide a more than adequate test of our hypothesis. The lessons drawn from research using these dilemmas are not intended to depend on any particular case, and, consistent with this, Kahane et al.'s objections are meant to apply to the use of sacrificial dilemmas in general, and not specifically to the *footbridge* case or even to personal dilemmas exclusively. The *switch* case tends to generate less variance than the *footbridge* case, but it is nevertheless philosophically controversial (Thomson, 2008), as it pits the utilitarian requirement to save more lives against a strict deontological prohibition against actively (but not passively) causing harm. Consistent with this, we find here that the *switch* case does a surprisingly good job of predicting philosophers' positions, although the judgments are overall tend toward the utilitarian response.

[8] Note that we are not claiming that utilitarian responses to sacrificial dilemmas reflect a general commitment to utilitarian values. More specifically, these dilemmas do not *and were never intended to* identify individuals who are committed to impartially maximizing good outcomes for all humanity (e.g., Kahane et al., 2017). Our claim, consistent with (Greene, 2007, 2013, 2014) is that sacrificial dilemmas reflect one of the key points of *disagreement* between consequentialist and deontological philosophies. This does not imply that these dilemmas reflect all that is important to either philosophical school of thought.

[9] As noted above, these findings are consistent with dual-process theory.

Table 1
Correlations between conventional utilitarian vs. deontological judgments, the utilitarian and deontological process dissociation parameters, acceptance of business ethics violations, psychopathy, empathic concern, gender, and age, Study 2.

	Conventional utilitarian vs. deontological judgments	Utilitarian PD parameter	Deontology PD parameter	Accept business ethics violations	Psychopathy	Empathic concern	Gender
Utilitarian PD parameter	**.60****						
Deontology PD parameter	**−.80*****	−.06					
Accept business ethics violations	.14	−.01	**−.22****				
Psychopathy	**.27****	**−.19***	**−.46****	**.44****			
Empathic concern	**−.20****	.05	**.29****	**−.35****	**−.64*****		
Gender ($m = 1$, $f = 2$)	**−.35*****	**−.19***	**.30*****	−.06	−.14	**.28*****	
Age	−.14	.08	**.21***	−.18	**−.29*****	.13	**.21****

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

temperature (see Oppenheimer, Meyvis, & Davidenko, 2009), leaving a final sample of 172 (105 males, 67 females, $M_{age} = 34.37$, $SD = 9.92$).

3.1.2. Procedure

Participants completed all personality and decision-making questions from Kahane et al.'s (2015) Study 1, the Conway & Gawronski process dissociation dilemma battery, and provided demographic information.

3.1.3. Personality and decision-making

Following Kahane et al. (2015), we assessed willingness to commit business ethics violations (Cooper & Pullig, 2013), psychopathy (Levenson, Kiehl, & Fitzpatrick, 1995), and empathic concern (Davis, 1983).

3.1.4. Moral dilemma task

Participants completed a set of 10 moral dilemmas, each with one incongruent and one congruent version, presented in a fixed random order (Conway & Gawronski, 2013). Each dilemma entailed deciding whether to directly harm some individuals in order to produce a given effect for other individuals. Participants indicated whether performing each harmful action was *appropriate* or *not appropriate* (following Greene et al., 2001). Incongruent dilemmas entail causing harm that maximizes overall outcomes, thereby involving a conflict between utilitarian and deontological answers, similar to classic high-conflict moral dilemmas (Koenigs et al., 2007). For example, the set of incongruent dilemmas includes a case in which one can torture one person to prevent an explosion from killing many people, and a case in which one can kill a crying baby to prevent a massacre.

Congruent versions of each dilemma are worded identically to incongruent versions, except that causing the same harm no longer maximizes overall outcomes. For example, congruent dilemmas involve deciding whether to torture one person to prevent messy but nonlethal property damage and whether to kill a crying baby to prevent imprisonment. Hence, in such cases rejecting the harmful action is consistent with both standard versions of deontological and utilitarian philosophies. While there may be no compelling moral reasons for accepting harm in such cases, there may be amoral or immoral reasons driven, for example, by self-interest or sadism. Thus, by employing both congruent and incongruent dilemmas, we can distinguish a pattern of responding consistent with utilitarian philosophy (only accepting harm that maximizes outcomes) from a general willingness to accept or even favor harm.

By applying participants' responses to both congruent and incongruent cases to a processing tree (see Fig. 1), we can algebraically estimate two independent parameters using the equations described by Conway and Gawronski (2013). The utilitarian (U) parameter reflects the degree to which participants systematically selected answers that maximize good outcomes, regardless of whether doing so requires harming another person. The deontology (D) parameter reflects the degree to which participants systematically reject causing harm under all conditions, regardless of whether doing so maximizes overall outcomes. Whereas conventional analyses treat these tendencies as perfectly inversely correlated ($r = -1.0$), process dissociation treats their relationship as an empirical matter. Meta-analytic findings across 40 datasets indicate that the deontology and utilitarian parameters correlate about $r = .01$, even though both typically correlate with conventional dilemma judgments around $r = .60$–$.70$ in the expected directions (Friesdorf et al., 2015). Moreover, Conway and Gawronski (2013) found evidence suggesting that the deontological parameter taps affective reactions to harm, whereas the utilitarian parameter taps cognitive evaluations of outcomes. Moreover, both parameters correlated with moral identity internalization (Aquino & Reed, 2002)—and these dual positive effects cancelled out for conventional dilemma judgments. Finally, it is worth briefly noting that past meta-analyses have found robust gender differences primarily on the deontological, but not utilitarian, parameter (Armstrong, Friesdorf, & Conway, 2018; Friesdorf et al., 2015). We anticipated replicating the meta-analytic correlational patterns and gender differences.

3.2. Results and discussion

3.2.1. Conventional analyses

We began by examining the correlations between each personality measure and willingness to cause outcome-maximizing harm on the 10 standard incongruent dilemmas. Higher scores on this conventional dilemma measure reflect relatively more harm-acceptance (utilitarian) judgments, whereas low scores reflect relatively more harm-rejection (deontological) judgments, in line with Kahane and colleagues and standard practice. As predicted, this analysis replicated many of the findings in Kahane and colleagues' Study 1 (see Table 1): Conventional utilitarian responses were associated with increased psychopathy, lower empathic concern, and increased acceptance of business ethics violations. Additionally, men made more utilitarian judgments than women, consistent with past findings (e.g., Fumagalli et al., 2010).

3.2.2. Process dissociation analysis

Before we examined how each personality measure correlated with the utilitarian and deontological parameters, we conducted preliminary analyses to determine whether the U and D parameters correlated in the expected manner with the conventional measure of utilitarian dilemma judgments (see Table 1). As expected, the U parameter correlated positively with conventional utilitarian judgment, whereas the D parameter correlated negatively. Likewise, as expected, the U and D parameters were not significantly correlated with each other. This pattern held across all studies and was consistent with meta-analytic findings

(Friesdorf et al., 2015), suggesting that each parameter reflects an independent process and that the two processes jointly influence conventional dilemma judgments. In addition, process dissociation found that the gender differences evident in conventional judgments loaded on both parameters, with a stronger effect on the deontological parameter, consistent with past meta-analyses (Friesdorf et al., 2015). Again, this pattern held across all studies, and we will therefore not comment on it further.

Next, we examined the correlations between each parameter and each personality measure. Although psychopathy correlated positively with overall levels of conventional utilitarian judgment (see above), psychopathy correlated *negatively* with both the U and D parameters. We also replicated Kahane and colleagues' finding that high levels of conventional utilitarian judgment correlated with reduced empathic concern. However, as predicted, we found no correlation between the U parameter and reduced empathy—instead, empathic concern correlated positively with the D parameter, consistent with past work (Conway & Gawronski, 2013). In addition, we found no correlation between the U parameter and acceptance of business ethics violations—instead, this measure loaded negatively on the D parameter. Finally, we conducted regression analyses predicting personality variables using both the D and U parameters as simultaneous predictors, controlling for gender and age. As these regressions produced similar results, we relegated them to the Supplementary Materials (see Table S1).

3.3. Discussion

We replicated Kahane et al.'s (2015) Study 1 findings when assessing conventional dilemma judgments that treat utilitarian and deontological tendencies as opposites. Conventional utilitarian judgments correlated positively with psychopathy and unethical business practices, and negatively with empathic concern. Hence, inferences predicated only on such conventional analyses might suggest that there is no genuine concern for the greater good behind these judgments. However, a process dissociation analysis reveals that judgments favoring utilitarian sacrifices have two distinct components. There is a low deontology (D) component that explains the positive association between sacrificial judgments and antisocial tendencies such as psychopathy, low empathic concern, and comfort with unethical business practices. But, critically, there is also a utilitarian component (U) that is negatively associated with these antisocial tendencies (i.e., a partial suppression effect, but one where the relatively greater strength of association between psychopathy and D than U results in a directional effect for conventional judgments).

In other words, of the participants endorsing outcome-maximizing harm on conventional high-conflict dilemmas, some appeared relatively comfortable with causing harm *regardless of whether causing harm maximized good outcomes or not*. These people tended to score higher in psychopathy. Conversely, other people who accepted harm on conventional high-conflict dilemmas only reported acceptance of harm when harm maximizes good outcomes (i.e., high U parameter)—and these people scored lower in psychopathy. Note that these people need not explicitly endorse utilitarian principles, but their judgments appear to be consistent with those principles and inconsistent with antisocial tendencies.

Thus, the results of Study 2 indicate that individual differences in the tendency to make sacrificial utilitarian judgments are not driven entirely by antisocial tendencies (low D), as Kahane and colleagues suggest. Instead, sacrificial utilitarian judgments appear to reflect two independent motivational components which vary across people: (an antisocial reduced concern for causing harm (i.e., low D) and an increased prosocial concern for maximizing good outcomes (i.e., high U). These findings are consistent with the dual-process theory (Greene et al., 2001; Greene et al., 2004; Greene, 2013), according to which individual differences in the frequency of utilitarian judgment may reflect either increased reliance on controlled evaluations of aggregate outcomes, or decreased reliance on affective responses to causing harm.

4. Study 3

Here we examine additional personality variables related to prosocial and antisocial tendencies, following Kahane and colleagues' Study 2. As before, we use process dissociation to clarify how these measures relate to utilitarian and deontological inclinations underpinning conventional dilemma judgments.

4.1. Method

4.1.1. Participants

We recruited 201 American participants via Amazon's Mechanical Turk, who received $1.50 for participating. We excluded one participant for failing to complete all 20 dilemmas, and 18 participants for failing the attention check. This left a final sample of 182 (96 males, 86 females, M_{age} = 33.39, SD = 10.48).

4.1.2. Procedure

Following Kahane et al.'s (2015) Study 2, participants completed the Identification with All Humanity scale (McFarland, Webb, & Brown, 2012), a three-facet measure of egoism, and a question about a hypothetical charitable donation. As in Study 1 of the present work, we again measured empathic concern (α = .89), whereas Kahane and colleagues again measured psychopathy. The Identification with All Humani + 3ty scale consists of nine items assessing how closely participants identify with three groups: their community (IWC; α = .93), Americans (IWA; α = .90), and all humanity (IWAH α = .89). The egoism measure assessed agreement with statements in favor of psychological egoism (*People may sometimes appear to do things for the sake of others, but deep down, the only thing that really motivates people is their own self-interest*), rational egoism (*An action isn't rational if it doesn't aim to promote one's own self-interest*), and ethical egoism (*An action isn't morally right if it doesn't aim to promote one's own self-interest*) on 7-point scales (1 = *strongly disagree*, 7 = *strongly agree*). The hypothetical donation question asked participants to imagine that they have received a $100 bonus from their employer, with the option to donate any amount from $0 to $100 to charity, with their employer doubling and donating whatever amount they select. Participants also completed the same process dissociation dilemma battery as in Study 2.

4.2. Results

4.2.1. Correlational analyses

All personality variables correlated sensibly with one another (see Table 2). For example, all three measures of identification correlated positively with one another, correlated negatively with psychological egoism, and correlated positively with empathic concern, religiosity, and age. All three measures of egoism correlated positively with one another. Two of them correlated negatively with donations, and one correlated negatively with empathic concern.

4.2.2. Conventional analyses

Here, our results were generally similar to those of Kahane and colleagues (see Table 2). Although they found no correlation between IWAH and levels of conventional utilitarian judgment, we found a negative correlation. They found a positive correlation between utilitarian judgments and rational egoism, whereas we found a positive correlation with ethical egoism instead. They found a negative correlation between utilitarian judgments and hypothetical donations, but we found a null effect instead. We replicated our finding from Study 2 that men and people low in empathic concern made more utilitarian judgments. Finally, we found a negative correlation between conventional utilitarian judgment and religiosity, which is well-documented elsewhere (e.g., Conway & Gawronski, 2013).

Table 2
Correlations between conventional utilitarian vs. deontological judgments, the utilitarian and deontological process dissociation parameters, identification with all humanity, community, and Americans, three kinds of egoism, hypothetical charity donations, empathic concern, gender, religiosity, and age, Study 3.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
1. Conventional utilitarian vs. deontological judgments													
2. Utilitarian PD parameter	**.56*****												
3. Deontology PD parameter	**−.74*****	.09											
4. Identification with all humanity	**−.23****	−.12	**.17***										
5. Identification with community	−.12	.01	.14	**.55*****									
6. Identification with Americans	−.08	.03	.10	**.59*****	**.76*****								
7. Psychological egoism	−.02	**−.17***	−.08	**−.24****	**−.21****	**−.22****							
8. Rational egoism	.08	**−.18***	**−.21****	−.04	.01	−.06	**.40*****						
9. Ethical egoism	**.16***	**−.22****	**−.32*****	−.03	−.04	−.09	**.30*****	**.64*****					
10. Hypothetical charity donation	−.08	−.01	.10	**.25****	.08	.01	**−.20****	**−.17***	−.10				
11. Empathic concern	**−.18***	−.06	**.17***	**.50*****	**.42*****	**.48*****	**−.26*****	−.07	−.05	**.26*****			
12. Gender ($m = 1$, $f = 2$)	**−.27*****	−.14	**.21****	**.20****	.02	.08	−.10	**−.23****	**−.15***	**.20****	**.27****		
13. Religiosity	**−.21****	−.02	**.26****	**.19***	**.32*****	**.29*****	−.02	−.08	−.11	**.17***	**.24****	**.16***	
14. Age	−.14	**.19****	**.32*****	**.15***	**.35*****	**.29*****	−.08	−.14	**−.23****	.02	.11	.06	**.18***

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

4.2.3. Process dissociation analysis

Table 2 displays the correlations between the U and D parameters and each personality variable. As usual, the U and D parameters correlated as expected with conventional utilitarian dilemma judgments, but did not significantly correlate with one another. As in Study 2, a PD analysis suggests a much different interpretation from interpretations based on conventional dilemma judgments.

4.2.3.1. IWAH. Although Identification with All Humanity correlated negatively with conventional utilitarian judgments, process dissociation clarified why: IWAH did not correlate significantly with the U parameter, but instead correlated positively with the D parameter.[10] Thus, in this sample, the people who identified most strongly with humanity as a whole did not appear especially concerned with maximizing good outcomes, but instead appeared especially concerned with avoiding causing harm.

4.2.3.2. Egoism. As predicted, we found negative correlations between the U parameter and each of the three egoism measures. We also found negative correlations between the D parameter and rational egoism and ethical egoism. Thus, the U and D parameters appear to reflect distinct, prosocial tendencies, and people high in egoism appear to score low in both such tendencies.

4.2.3.3. Donations. We failed to replicate Kahane et al.'s (2015) finding concerning hypothetical charitable donations. Likewise, we found no correlation between the U and D parameters and hypothetical donation levels.

4.2.3.4. Empathic concern and gender. The PD patterns for empathic concern and gender replicated the results of Study 1. Once again, empathic concern was correlated with gender and with the D parameter, but not with the U parameter. Here, the link between gender and the U parameter did not reach significance, consistent with previous weak and inconsistent effects (Friesdorf et al., 2015). The D parameter selectively correlated with religiosity, replicating past findings (Conway & Gawronski, 2013). In addition, older people scored higher on both the U and D parameters.

Finally, we conducted a set of separate regression analyses for each personality variable, entering the U and D parameters simultaneously as predictors while controlling for age and gender (see Table S2 in the Supplementary Materials). The results were generally consistent with the pairwise correlational analyses presented in Table 2. Although scores for IWAH fluctuated somewhat, all egoism relations remained significant. In addition, the correlation between the D parameter and empathic concern appears driven by gender differences.

4.2.4. Discussion

Overall, findings using conventional analyses largely replicate Kahane et al. (2015), but a process dissociation analysis presents a very different picture. Like them, we found that, people who were willing to accept outcome-maximizing harm on high-conflict dilemmas (i.e., made more utilitarian judgments) tended to score higher in egoism and lower in empathy and Identification with All Humanity. Yet, a process dissociation analysis indicated that these relations reflect reduced deontological tendencies (i.e., negative correlations with the D parameter) and not increased utilitarian tendencies (i.e., positive correlations with the U parameter). Critically, we found that participants scoring high on the U parameter scored relatively low on each measure of egoism, and showed no sign of being especially low in empathy, generosity, or identification with others. Thus, egoism demonstrated partial suppression—it correlated negatively with both parameters, and these simultaneous relationships largely cancelled out in conventional analyses.

Thus, these data indicate that many ordinary people have prosocial reasons for making utilitarian judgments and that antisocial people appear to be more "un-deontological" than genuinely (Level-3) utilitarian. In typical samples, these un-deontological tendencies account for more of the variance in individual differences than prosocial utilitarian tendencies, but the prosocial utilitarian tendencies clearly exist and are readily observed once one adequately controls for the presence of confounding antisocial, un-deontological tendencies. These findings highlight the limitations of using conventional analyses that treat utilitarian and deontological tendencies as opposites when examining individual differences. When the tendencies reflected in the U and D parameters are pitted against one another, as in sacrificial dilemmas, the negative relations between these parameters and measures of antisociality tend to cancel out, resulting in weak or null effects. See, for example, stronger effects for egoism in columns 2 and 3, as compared to

[10] When we computed partial correlations between IWAH and each parameter, controlling for Identification with Community and Identification with America, the utilitarian parameter negatively predicted IWAH, $r = -.17$, $p = .020$, whereas the deontology parameter marginally positively predicted IWAH, $r = .13$, $p = .091$. Likewise, in the regressions, there was a marginal trend toward a negative relationship between the utilitarian parameter and IWAH, controlling for IWC and IWA (see Supplementary material). However, IWC and IWA are highly correlated, introducing a problem of multicollinearity that requires researchers to interpret the results of these analyses with extreme caution (Aiken, West, & Reno, 1991).

Table 3
Correlations between conventional utilitarian vs. deontological judgments, the utilitarian and deontological process dissociation parameters, "real-world" utilitarianism items, "real-world" harm items, hypothetical bonus donations, gender, and age, Study 4.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Conventional utilitarian vs. deontological judgments														
2. Utilitarian PD parameter	**.62*****													
3. Deontology PD parameter	**−.69*****	.12												
4. Pond dilemma: help drowning child	−.10	.08	**.21***											
5. Not donating to poor children is wrong	.05	**.20**†	.16	**.19***										
6. Obligation of west to help world's poor	−.16	.06	**.26****	.12	.06									
7. Helping foreign rather than own country	−.15	−.08	.14	**.20***	−.05	.16								
8. Obligation to prevent climate change	.07	.16	.02	.01	**.19**†	**.19***	.11							
9. Animal experiments wrong	−.10	.03	.16	.06	**.27****	−.14	.01	.10						
10. Abortion wrong	**−.24***	.03	**.37*****	.10	.18	.14	−.04	**−.20***	−.07					
11. Eating meat wrong	−.11	−.04	.07	−.03	.16	.06	−.01	**.35*****	.18	−.08				
12. Torture to save lives is acceptable	**.18**†	**.18**†	−.09	.07	.13	−.08	**−.24***	−.06	−.03	.03	−.12			
13. Hypothetical bonus donation	**−.25****	−.08	**.26****	−.15	.15	−.07	.06	.07	**.21***	.05	.10	.05		
14. Gender ($m = 1, f = 2$)	**−.36*****	−.18	**.30****	.11	.01	.08	.06	.09	**.20***	.03	**.21***	.07	**.21***	
15. Age	.03	−.07	−.15	.02	−.01	−.07	.04	.04	−.10	.07	−.11	.12	.16	.12

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

column 1, in Table 2. Consistent with this, Conway and Gawronski (2013, Study 1) found that both the U and D parameters correlated positively with moral identity, but these effects cancelled out in people's responses to sacrificial dilemmas, which pit deontological concerns against utilitarian concerns. In sum, these findings provide additional evidence for our claim that the U and D parameters reflect distinct, prosocial influences on ordinary people's moral judgments.

5. Study 4

Next, we examined responses to "real-world" moral problems, following Kahane and colleagues' Study 3. As before, we used process dissociation to better distinguish between prosocial and antisocial motivations for making utilitarian judgments in response to sacrificial dilemmas.

5.1. Method

5.1.1. Participants

We recruited 118 participants at a major American public university, who received partial course credit for their participation. We excluded 11 participants who failed to respond to all 20 dilemmas, resulting in a final sample of 107 individuals (81 males, 26 females, $M_{age} = 19.60$, $SD = 1.54$).

5.1.2. Procedure

Participants completed the Conway & Gawronski dilemma battery used in Studies 2–3, responded to Kahane et al.'s (2015) questions regarding "real-world" utilitarian beliefs and real-world harm permissibility, answered a hypothetical donation measure, and provided demographic information.

5.1.2.1. Real-world utilitarianism items. Participants responded to a series of vignettes intended to assess their concern for maximizing overall well-being in the real world. Participants read Singer's (1972) Pond dilemma and indicated how wrong it would be to not save a child drowning in a pond (1 = *not at all wrong*, 7 = *very wrong*). They also indicated on either 5- or 7-point scales how wrong it would be for people to fail to engage in prosocial behavior toward (a) children in poor countries, (b) nations less wealthy than Western countries, (c) foreign poor, and (d) future generations (see Kahane et al., 2015, Supplementary Materials for exact wording). For example, participants indicated their belief about helping foreign poor people on a scale from 1 (*It would be wrong for well-off people in the West to help poor people in developing countries*) to 5 (*Well off people in the West must help poor people in developing countries*). Although Kahane et al. (2015) combined these items into a single measure, they failed to hang together reliably in the current data ($\alpha = .27$). We therefore assessed them separately.

5.1.2.2. Real-world harm items. These items assessed participants' attitudes toward causing harm in real-world situations. Participants indicated the moral acceptability of (a) animal experimentation, (b) abortion, (c) eating meat, and (d) torture in service of saving lives. The first three questions were scored on 7-point scales (with higher scores reflecting rejection of the actions), and the last on a 3-point scale (with higher scores reflecting acceptance of torture). For example, participants responded to the item: *How morally wrong or right is eating meat?* As with the real-world utilitarianism items, the Cronbach's Alpha for these items was extremely low ($\alpha = -.05$). We therefore assessed each item separately.

5.1.2.3. Hypothetical donation measure. Participants completed the same hypothetical donation item as in Study 3.

5.2. Results

5.2.1. Conventional analysis

As before, we examined the relationship between each of the above measures and levels of utilitarian versus deontological responding to sacrificial moral dilemmas (see Table 3). Consistent with Kahane et al. (2015), we did not find any significant correlations between conventional utilitarian judgments and measures of "real-world utilitarianism" or "real-world harm." However, there were marginal correlations between utilitarian judgments and rating abortion as less wrong, rating torture in service of saving people as more acceptable, and lower hypothetical donations. Again, men tended to make more utilitarian judgments.

5.2.2. Process dissociation analysis

As in Studies 2–3, conventional utilitarian judgments correlated positively with the U parameter and negatively with the D parameter. As before, the U and D parameters were not significantly correlated with one other. Moreover, we replicated the gender differences mentioned previously. Turning to the "real-world" utilitarianism items, individuals with high U parameters were marginally more likely to disapprove of failing to help poor children. The D parameter correlated positively with desire to save the drowning child in the Pond dilemma, and with the view that the West should help poorer nations. No other

relations reached significance.

Regarding real-world harm, the U parameter marginally correlated positively with acceptance of torture to save lives, whereas the D parameter correlated positively with disapproval of abortion. No other relations reached significance. Regarding donations, the negative correlation between utilitarian judgments and hypothetical donations appeared entirely driven by a positive correlation between the D parameter and donations; there was no significant correlation, positive or negative, between the U parameter and donations. Finally, as in previous studies, regression analyses confirmed that these effects generally held when simultaneously using both parameters to predict each measure, controlling for gender and age (see Table S3).

5.3. Discussion

Altogether, these findings reveal weak and inconsistent relationships involving Kahane and colleagues' measures of "real-world" utilitarian commitments and "real-world" harm. A few plausible relations did emerge, however: The D parameter correlated positively with greater concern for victims (helping poor, hypothetical donations), and the U parameter was marginally positively correlated with disapproving of failing to help the poor, and with approval of torture aimed at saving innocent lives. Nevertheless, the majority of "real-world" utilitarianism and harm items failed to correlate with either parameter or with conventional dilemma judgments. We note, however, that few of these "real-world" items correlated positively with *one another*. Indeed, the reliabilities for both measures were very low ($\alpha = .37$ and $-.05$, respectively). Therefore, we suggest that these "real-world" utilitarianism and harm items do not reflect a single underlying construct relating to utilitarian thinking, or anything else—at least not when presented to ordinary people.

We suspect that these measures demand too much philosophical acumen from the general public. As noted above, many people engage in impartial cost-benefit reasoning, and some people are more likely than others to favor such reasoning when it competes with other moral considerations, as in trolley dilemmas. There is, however, a big difference between employing impartial cost-benefit reasoning in some cases and transforming this tendency into classic utilitarian stances on foreign aid and animal rights. To employ such measures as indices of utilitarian thinking assumes that ordinary people spontaneously make the kinds of philosophical connections that Peter Singer makes in his most influential works (Singer, 1972; Singer, 1975). Consequently, we are not surprised that these measures of "real-world" utilitarian moral commitment exhibit no internal coherence and appear largely unrelated to dilemma judgments, at least when assessing lay populations (we suspect that results may be different if assessing professional philosophers). Importantly, these results provide no clear support for Kahane and colleagues' claims that utilitarian judgments primarily reflect antisocial tendencies.

6. Study 5

Here we examined responses evaluating people's failures to maximize good overall outcomes, following Kahane and colleagues' Study 4. As before, we used process dissociation to better distinguish between prosocial and antisocial motivations for making utilitarian judgments in response to sacrificial dilemmas. We also employed measures of psychopathy and empathic concern, as in Study 2.

6.1. Method

6.1.1. Participants

We recruited 183 American participants via Amazon's Mechanical Turk. They received $1.50 for participating. We excluded 12 participants for failing the attention check, leaving a final sample of 171 (91 males, 80 females, $M_{age} = 34.25$, $SD = 10.91$).

6.1.2. Procedure

Participants first responded to seven "greater good" scenarios, as in Kahane et al.'s (2015) Study 4. Each scenario describes a person who must choose among actions favoring, to varying degrees, self-interest, parochial interest, or broad societal interest. Participants indicated how wrong it would be for the person to *fail* to pursue the greater good on scales from 1 (*not at all wrong*) to 7 (*very wrong*). For example, scenarios involved deciding whether to buy a car or donate to charity, deciding whether to donate to a local versus international charities, and deciding whether to visit one's mother or volunteer for Habitat for Humanity (see Kahane et al., 2015, Supplementary Material for complete wording). We combined responses to these vignettes into a single measure assessing the wrongness of failing to prioritize the greater good ($\alpha = .78$).[11] Participants then completed the same measures of psychopathy ($\alpha = .92$) and empathic concern ($\alpha = .91$) as in Study 2, and the Conway and Gawronski (2013) dilemma battery.

We note that Kahane and colleagues' greater good dilemmas are critically different from the moral dilemmas typically used in psychological research. These dilemmas are not designed to assess commonsense cost-benefit thinking, but rather to assess commitments to a strong version of philosophical utilitarianism, according to which, maximizing aggregate well-being is not only *permissible*, but *mandatory*. By contrast, most researchers (ourselves included) have employed moral dilemmas that ask whether the utilitarian option is *appropriate*, *morally permissible*, or *morally acceptable*. Research indicates that very few ordinary people are so thoroughly committed to utilitarianism (Lombrozo, 2009; Royzman et al., 2015). Therefore, we had no strong expectation of finding a positive correlation between the U parameter and condemnation of failures to uphold the greater good. We take this as an opportunity to replicate our findings concerning psychopathy and empathic concern.

6.2. Results

6.2.1. Conventional analyses

We began by again correlating each measure in the study with conventional utilitarian dilemma judgments (see Table 4). Whereas Kahane and colleagues found no significant relation between utilitarian judgment and condemnation of failing to support the greater good, we found a negative correlation. In other words, people in our study who indicated that causing harm to maximize outcomes is *permissible* were actually less likely than others to regard it as *mandatory*. Indeed, few people indicated such harm was mandatory—on a 7-point scale, the mean was 2.14 ($SD = .94$), with only 11% of participants ($n = 19$) scoring above the mid-point. We also found that conventional utilitarian judgment correlated positively with psychopathy and negatively with empathic concern, replicating the results of Study 2 and Kahane and colleagues' Study 1. Again, men tended to make more conventional utilitarian judgments, and this time, younger people did as well.

6.2.2. Process dissociation analysis

As before, the U parameter correlated positively with conventional utilitarian judgments, whereas the D parameter correlated negatively, and the two parameters were unrelated (see Table 4). We also replicated the previously documented gender effects. As with conventional utilitarian judgments, we found a negative correlation between the U parameter and the greater good measure. The greater good measure did not correlate with the D parameter. However, we

[11] Kahane et al. (2015) conducted a principle component analysis on the greater good dilemmas, arguing they load on two dimensions: self-sacrifice and impartiality. We conducted a similar principle component analysis using the more conservative direct oblimin rotation, yet nonetheless found that all greater good dilemmas loaded on a single factor (eigenvalue 3.17, 45.26% of variance explained). Therefore, we analyzed all greater good dilemmas together. Results remain similar when we divide them into the two dimensions examined by Kahane and colleagues.

Table 4
Correlations between conventional utilitarian vs. deontological judgments, the utilitarian and deontological process dissociation parameters, wrongness of failing to prioritize greater good, psychopathy, empathic concern, gender, and age, Study 5.

	Conventional utilitarian vs. deontological judgments	Utilitarian PD parameter	Deontology PD parameter	Wrongness of failing to prioritize greater good	Psychopathy	Empathic concern	Age
Utilitarian PD parameter	**.49*****						
Deontology PD parameter	**−.70*****	**.23****					
Wrongness of failing to prioritize greater good	**−.20*****	**−.28****	.04				
Psychopathy	**.22****	**−.17***	**−.36*****	.09			
Empathic concern	**−.25*****	.02	**.24****	.11	**−.68*****		
Gender ($m = 1$, $f = 2$)	**−.23****	.02	**.22****	.01	**−.36****	**.31*****	
Age	**−.16***	**.18***	**.32*****	−.10	−.25**	.12	.05

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

replicated the results from Study 2 concerning psychopathy, empathic concern, gender, and age. Although psychopathy correlated positively with conventional utilitarian judgments, it once again correlated negatively with the U parameter. Psychopathy also correlated negatively with the D parameter (again, these two negative relationships result in partial suppression of the effect on conventional analyses).

Although empathic concern correlated negatively with conventional utilitarian judgments, we once again found no correlation with the U parameter. Instead, empathic concern correlated positively with the D parameter, explaining the negative correlation between empathic concern and conventional utilitarian judgments (also consistent with Conway & Gawronski, 2013). The increased conventional utilitarian judgments among men were driven primarily by higher D parameter scores among women. Younger people tended to make more conventional utilitarian judgments, but age was, in fact, positively correlated with both the U and D parameters, indicating that the negative relation between age and utilitarianism reflects the strength difference between these effects. Once again, we present only the correlational analyses here, but the regression analyses indicated a very similar pattern (see Table S4 in the Supplementary Material).

6.3. Discussion

Consistent with Studies 2–4, Study 5 indicated that conventional utilitarian judgments often reflect prosocial concern for the greater good, as we replicated our previous finding that the U parameter correlated negatively with psychopathy. Likewise, we found once again that the reduced empathy associated with conventional utilitarian judgments is best understood as "un-deontological" (negatively correlated with the D parameter), rather than truly (Level-3) utilitarian (i.e., correlated with the U parameter). The results of Study 5 using Kahane and colleagues' "greater good" dilemmas make it clear that lay responses to moral dilemmas do not reflect a thoroughgoing commitment to utilitarianism (Level 4), whereby actions that maximize the greater good are viewed as mandatory. Yet, as noted earlier, no researchers to our knowledge have ever claimed that substantial numbers of ordinary people have such commitments. If ordinary people are not committed to utilitarian principles, the question remains whether their thinking is best describes as merely cost-benefit thinking (Level 2), or whether they show evidence of concern for the greater good (Level 3). We address this question further in Study 6.

7. Study 6

Studies 2–5 provide evidence that the U parameter correlates negatively with antisocial tendencies. This finding indicates that people who score high on U parameter have prosocial, moral motivations behind their utilitarian judgments, even if they are not generally committed to utilitarian values and therefore do not make Level-4 utilitarian judgments. Studies 2–5 suggest that the U parameter reflects Level-3 utilitarian thinking, reflecting moral concern for the wellbeing of others, as opposed to mere Level-2 cost-benefit reasoning with no genuine moral concern behind it. In the current study, we employed a number of additional measures of moral thinking, not previously employed by Kahane and colleagues, with the aim of further clarifying the link between the U parameter and prosocial moral motivation.

One suggestive line of research comes from Janoff-Bulman, Sheikh, and Hepp (2009), who found that people think differently about *moral proscriptions*—which entail avoiding moral violations—and *moral prescriptions*—which entail pursuing moral excellence. Although lay people view both proscriptions and prescriptions as part of morality, they typically view proscriptions as more important and obligatory than prescriptions. In other words, moral attitudes seem to follow the general psychological dictum that 'bad is stronger than good' (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001), such that avoiding doing bad deeds (e.g. not stealing $10) is more important and obligatory than doing good deeds (e.g. donating $10 to charity).

Applying these insights to research using moral dilemmas suggests that lay people's utilitarian thinking may focus more on proscriptions than prescriptions, but in a way that is focused on consequences rather than actions. That is, people may place substantial weight on the goal of minimizing aggregate harm—that is, preventing a major decrease in well-being from a perceived psychological default—and they may not place equal importance on the goal of improving net well-being—that is, promoting a substantial increase in well-being beyond a perceived psychological default. If so, this may explain why we, like Kahane et al. (2015), found no relationship between the U parameter and measures of utilitarian beneficence, such as giving money to charity. Likewise, this may explain why there is no positive correlation between the U parameter and Kahane and colleagues' greater good measures, which ask whether serving the greater good should be regarded as mandatory, i.e., whether it's wrong for people to not serve the greater good when doing so conflicts with other reasonable values. Here, we revised Kahane and colleagues' greater good vignettes to assess judgments about permissibility rather than obligation (i.e. the wrongness of failing to act).

We also employed several additional measures to better understand the moral commitments of participants with high U parameter scores. First, we adapted Skitka and colleagues' measures of "moral conviction" (e.g., Skitka et al., 2005) to assess whether participants' decisions to minimize harm reflect their core moral values and beliefs. Moral convictions, as defined by Skitka and colleagues, pertain to a specific issue or domain and reflect a belief that the relevant moral considerations apply to all people at all times and in all places. As such, people with strong moral convictions are upset by indications that others disagree with them. Moral convictions, so defined, powerfully predict relevant behaviors and actions, above and beyond attitude strength (certainty and extremity). Here we assessed both moral conviction and attitude

strength regarding harm. This enabled us to ask whether the U parameter is associated with moral conviction in this specific sense.

Second, after participants completed the PD dilemma battery, we asked them to report how much their thinking focused on the individual to be sacrificed and (separately) how much it focused on the group to be protected (Robinson et al., 2015). If the U parameter reflects concern about minimizing overall harm, it should correlate specifically with focus on the group. Likewise, the D parameter should correlate specifically with focus on the individual to be sacrificed. Third, we included a measure of moral identity internalization and symbolization (Aquino & Reed, 2002). Here, our aim was to determine whether people with high U scores are more likely to view morality as part of their self-concept (internalization) and whether they are more likely to advertise their moral commitments to others (symbolization).

7.1. Method

7.1.1. Participants

We recruited 192 American participants via Amazon's Mechanical Turk. Each was paid $1.50. We excluded six people for failing the attention check, leaving a final sample of 186 (101 males, 84 females, 1 other, M_{age} = 34.94, SD = 11.47).

7.1.2. Procedure

Participants completed the moral identity measure, measures of moral conviction and attitude strength, revised versions of Kahane and colleagues' "greater good" dilemmas, and measures of individual and group focus, before filling out the Conway and Gawronski dilemma battery and providing demographic information.

7.1.2.1. Moral identity. Participants completed the Aquino and Reed (2002) measure of moral identity, tapping the centrality of morality to the self-concept. This measure provides participants with nine moral adjectives (e.g., honest, honorable, trustworthy), and an opportunity to agree with 10 statements regarding these terms (e.g., *Having these traits is not important to me*) on scales anchored at 1 (*does not describe me at all*) and 7 (*describes me completely*). The measure breaks down into two subscales: internalization (α = .85), reflecting morality as the core of the self-concept, and symbolization (α = .92), reflecting morality as demonstrated socially to others. Crucially, internalization tends to be a far better predictor of prosocial behavior than symbolization (e.g., Aquino & Reed, 2002; Reed & Aquino, 2003).

7.1.2.2. Moral conviction and attitude strength about harm. We adapted Skitka and colleagues' four-item moral conviction measure to assess moral convictions regarding harm (see Skitka & Morgan, 2014). Participants responded on 7-point scales (1 = *not at all*, 7 = *very much*). Each question began by asking participants, *Think about harming another person*. We then asked the four moral conviction questions (α = .74): *To what extent is your position on harm… a reflection of your core moral beliefs and convictions?…connected to your beliefs about fundamental right and wrong? …based on moral principle?… based on a moral stance?* We also asked the three attitude strength questions ((α = .70) with the same prompt: *How strongly do you feel about harm? How important is harm to you personally? How much does harm relate to how you see yourself as a person?* This allowed us to examine whether moral conviction about harm predicts dilemma decisions above and beyond mere attitude strength regarding harm. We note, however, that none of these items clarify which position participants take (for or against causing harm). Thus, we also asked participants to indicate *To what extent is it wrong to cause harm overall?* We assumed that, unlike the contentious issues Skitka and colleagues assess, all of our participants would score at or above the midpoint of this scale, indicating consensus that causing harm is wrong in general. Results indicated this was true of all but one participant, and removing this individual from analysis had no discernable impact on results. We included this item in the correlational analysis.

7.1.2.3. Reworded greater good dilemmas. The greater good scenarios developed by Kahane et al. (2015, Study 4) examined moral condemnation of failures to maximize the greater good in ways that go beyond conventional expectations. In other words, they assess perceptions that maximizing outcomes is mandatory—a position that few people endorse (Lombrozo, 2009; Royzman et al., 2015). Therefore, we adjusted the wording of these scenarios to assess beliefs about *acceptability* rather than strict *obligation* (see Barbosa & Jiménez-Leal, 2017). We anticipated that this change would eliminate the negative correlation found between this measure and the U parameter in Study 5. We also modified these vignettes to increase clarity (see Appendix B for full materials).

7.1.2.4. Moral dilemma battery. All participants completed the same dilemma battery used in Studies 2–5.

7.1.2.5. Individual/group focus scale. Finally, we asked participants to separately indicate how much they focused on (a) the individual to be sacrificed and (b) the overall group when answering moral dilemmas (Robinson et al., 2015). This scale consists of four items, each beginning with *When answering the dilemmas, how much were your judgments affected by…* Two items assessed focus on the individual: *…the welfare of the person being sacrificed?* and *…how the person being sacrificed would feel?* and two assessed focus on the group: *…the welfare of all the people involved as a whole?* and *…what you thought would be best for the group as a whole?* Participants responded on scales from 1 (*Didn't affect my judgments at all*) to 7 (*Affected my judgments strongly*). We combined the two individual (α = .69) and group items (α = .87). Robinson and colleagues correlated these measures with dilemma judgments and found, as one would expect, that utilitarian judgments correlated negatively with focus on the victim and positively with focus on the group; They also found that the two focus measures (individual vs. group) were uncorrelated.

7.2. Results and discussion

7.2.1. Conventional analyses

As before, we examined the correlations between conventional utilitarian judgments and the other variables in the study (see Table 5). Conventional utilitarian judgments correlated with neither moral identity internalization nor symbolization. Nor did conventional utilitarian judgments correlate with moral conviction about harm or higher ratings for the wrongness of causing harm. However, conventional utilitarian judgments correlated negatively with attitude strength regarding harm, and positively with the judgment that ignoring the greater good is acceptable. People who made more conventional utilitarian judgments also indicated that they focused less on the individual and more on the group (replicating Robinson et al., 2015). We also replicated the previously observed age and gender effects.

7.2.2. Process dissociation analysis

Once again, we present only the correlational analyses here, as regression analyses predicting measures using both the U and D parameters simultaneously, while controlling for age and gender, obtained a very similar pattern of results (see Table S5 in the Supplementary Material). As before, the U and D parameters correlated as expected with conventional utilitarian judgment and did not significantly correlate with one another. The U and D parameters correlated positively with moral identity internalization (another suppression effect), whereas the U parameter correlated negatively with symbolization (also consistent with Conway & Gawronski, 2013). This pattern indicates that people who reject causing harm and people who strive to minimize aggregate harm both care deeply about being moral. However, people who care about demonstrating their morality to others

Table 5
Correlation between conventional utilitarian vs. deontological judgments, the utilitarian and deontological process dissociation parameters, moral identity internalization and symbolization, moral conviction and attitude strength about harm, overall wrongness of harm, acceptability of failing to prioritize the greater good, individual and group focus, gender, and age, Study 6.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Conventional utilitarian vs. deontological judgments												
2. Utilitarian PD parameter	**.51*****											
3. Deontology PD parameter	**−.77*****	.11										
4. Moral identity internalization	−.04	**.19***	**.16***									
5. Moral identity symbolization	−.11	**−.21****	.00	**.32*****								
6. Moral conviction about harm	−.07	**.17***	**.18***	**.38*****	.08							
7. Attitude strength about harm	**−.16***	.09	**.26*****	**.20****	.11	**.62*****						
8. Harm wrongness	−.07	**.20****	**.20****	**.41*****	.06	**.38*****	**.29*****					
9. Acceptability of failing to prioritize greater good	**.20****	.02	**−.22****	−.08	−.13	.02	−.05	−.14				
10. Individual focus	**−.21****	.11	**.29*****	**.32*****	**.20****	**.17***	**.20****	**.32*****	**−.14**†			
11. Group focus	**.28*****	**.30*****	−.13	**.23****	.02	.09	−.03	**.14**†	.11	**.18***		
12. Gender (m = 1, f = 2)	**−.23****	**−.15***	**.15***	**.20****	**.18***	.10	.05	**.15***	−.06	.06	−.08	
13. Age	**−.24****	−.04	**.26****	.11	−.08	.11	**.21****	**.14**†	−.09	.07	**−.21****	.02

† $p < .06$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

appear to care less about minimizing bad outcomes.

Moral conviction about harm correlated positively with the U parameter, whereas attitude strength about harm correlated positively with the D parameter. Moral conviction about harm also correlated with the D parameter, but this effect was reduced in regressions using moral conviction to predict each parameter, controlling for attitude strength; All other effects remained significant.[12] Harm wrongness ratings correlated positively with both the U and D parameters—again, these simultaneous positive effects cancelled out for conventional analyses (i.e., a suppression effect). Overall, this pattern of results indicates that people who have strong moral convictions against causing harm nevertheless tend to accept harmful utilitarian sacrifices because these sacrifices minimize aggregate harm.

Next, we examined the greater good scenarios, revised to elicit judgments concerning acceptability rather than strict obligation. Here there was no correlation between responses on the greater goods scenarios and the U parameter, and a negative correlation between such responses and the D parameter. This pattern mirrors that of willingness to accept business violations in Study 2 and is the opposite of that observed for empathic concern in Studies 2, 3, and 5. It suggests that people who consistently reject causing harm in dilemmas and people who tend to empathize with others tend to have negative attitudes toward people who pursue self- or parochial-interest. Conversely, people who favored minimizing harm in sacrificial dilemmas are neither more nor less likely to endorse pursuing self-interest instead of serving the greater good in unexpected but admirable ways. This finding fits with the conceptualization of the U parameter as tracking a commitment to the local minimization of harm rather than a global pursuit of the greater good that goes beyond conventional expectations.

As expected, focus on the individual in sacrificial dilemmas correlated selectively with the D parameter, whereas focus on the group correlated selectively with the U parameter. It is worth emphasizing that the opposite relations did not appear: The U parameter did not correlate negatively with focus on the individual, nor did the D parameter correlate negatively with focus on the group. This pattern of results—which is precisely predicted by the dual-process theory—underscores our general conclusion that utilitarian judgments often reflect genuine concerns about aggregate outcomes, concerns that compete with concerns about causing harm. Finally, consistent with past findings, women tended to have higher D parameter scores while men tended to have higher U parameter scores. Older people tended to score higher on the D parameter.

Taken together, these findings indicate that ordinary people's utilitarian judgments are by no means purely antisocial. Instead, many such judgments appear to be Level-3 utilitarian, reflecting genuine concern for the greater good within the context of the decision. People with high U parameter scores have strong moral convictions about harm, score highly on moral identity internalization (but low on symbolization), and express a focus on group welfare. They place a great deal of moral weight on minimizing harm in dilemma contexts, but they are not generally committed to maximizing global well-being in a way that goes beyond conventional expectations. Likewise, people with high D parameter scores also engage in moralized thinking, as they also score highly on moral identity internalization, express strong attitudes about harm, express a focus on the individual who could be sacrificed in dilemmas, and indicate that selfishness is unacceptable.

We emphasize the role that process dissociation has played in revealing these patterns. Sacrificial dilemmas, by design, pit utilitarian concerns for the greater good against deontological concerns about causing harm in a zero-sum fashion. Thus, any traits that are associated with both kinds of moral concern are likely to go undetected if one focuses exclusively on judgments in response to such cases. Or, if they are detected, they are likely to be mistakenly attributed to only one kind of moral concern. Process dissociation, by distinguishing prosocial utilitarian concerns for minimizing harm from un-deontological indifference to causing harm, gives us a clearer view of the traits and psychological processes behind sacrificial utilitarian judgments.

8. General discussion

Kahane et al. (2015) have argued against the use of sacrificial moral dilemmas in psychological research and against the dual-process theory. They claimed that ordinary people's sacrificial utilitarian judgments "do not reflect impartial concern for the greater good" (p. 193) and instead "merely express a calculating yet selfish mindset" (p. 197). The seven studies presented here provide evidence against these sweeping claims. Studies 1a and 1b affirmed the connection between consequentialist normative ethical stances and sacrificial dilemma judgments in two samples of practicing philosophers: Philosophers who identified as accepting or leaning toward consequentialism were

[12] When we regressed the utilitarian parameter on both moral conviction and attitude strength simultaneously, moral conviction remained a significant predictor, $B = .03$, $SE = .01$, $p = .038$, whereas attitude strength became non-significant, $B = -.004$, $SE = .01$, $p = .714$. Conversely, when we regressed the deontology parameter on both predictors simultaneously, attitude strength remained significant, $B = .03$, $SE = .01$, $p = .010$, but moral conviction did not, $B = .01$, $SE = .01$, $p = .709$.

overwhelmingly more likely to make utilitarian sacrificial dilemma judgments, and vice versa. These studies provide evidence against Kahane's (2015) claim that sacrificial dilemmas, including trolley cases, do not reflect an important tension in Western moral philosophy between consequentialism and deontology.

Studies 2–6 replicated and extended the studies in Kahane et al. (2015), but instead of measuring dilemma responses using conventional techniques, we employed process dissociation to independently assess the utilitarian and deontological response inclinations underpinning sacrificial dilemma judgments (Conway & Gawronski, 2013). Conventional analyses appeared to support Kahane and colleagues' conclusions, as we replicated associations between conventional utilitarian responses and various measures of antisociality, such as psychopathy and egoism. Likewise, we found null or negative relations between conventional utilitarian responses and various measures of prosociality, such as charitable donations. However, process dissociation revealed a very different pattern of effects.

According to Greene and colleagues' dual-process theory (Greene, 2007; Greene et al., 2001; Greene et al., 2004, 2013), responses to sacrificial dilemmas reflect competing psychological processes. Specifically, utilitarian judgments may reflect relatively strong inclinations to maximize good outcomes (captured by the PD utilitarian parameter) and/or relatively weak aversions to causing harm (captured by the deontological parameter). Here, process dissociation analyses revealed that nearly all of the associations between conventional sacrificial utilitarian judgment and antisociality reflect negative associations between antisociality and the deontological parameter. Moreover, when assessing the utilitarian parameter independently of such "un-deontological" antisocial impulses, it becomes clear that utilitarian response inclinations do, in fact, reflect genuine moral concern for the greater good (i.e., qualify as Level-3 utilitarian). Not only does the utilitarian parameter correlate negatively with various measures of antisociality, such as psychopathy and egoism; it also correlates positively with various measures of moral thinking, such as moral identity internalization, moral conviction about harm, and concern for group well-being.[13]

This is not to say that Kahane and colleagues studies have provided no useful information, and using conventional analyses, we replicated many of their findings. It has been an open question whether and to what extent ordinary people who make sacrificial utilitarian judgments are generally committed to utilitarian values. Their results indicate that ordinary people are not generally committed to utilitarian values, and our results support this more limited conclusion, which is consistent with, though not required by, the dual-process theory. People who make sacrificial utilitarian judgments often do so out genuine concern for the greater good, though such concerns do not reliably generalize to other contexts, especially those involving the general promotion of well-being rather than the prevention of immediate harm. These genuinely moral motivations for endorsing utilitarian sacrifices are difficult to detect using conventional analyses, because measures of prosociality load positively on both utilitarian and deontological response tendencies, which then cancel out when these tendencies are treated as opposites in conventional analyses. As this limitation is a feature of the way dilemma responses are measured, statistically controlling for antisociality, as Kahane and colleagues did in multiple studies, is not sufficient to overcome this limitation. Such statistical controls do not affect the way dilemma responses are assessed. Thus, it takes a more sensitive measure, such as process dissociation, to reliably detect such effects.

[13] The one caveat to this pattern was that we found no evidence for relationships between the utilitarian parameter and prosociality that involves improving conditions from baseline (i.e., charity donations) instead of preventing the worst possible outcome (i.e., preventing death). We suspect this finding reflects the fact that different psychologies are involved in preventing harm versus promoting optimal well-being (Janoff-Bulman et al., 2009).

8.1. Implications for definitions, theory, and philosophy

8.1.1. Clarifying the definition of utilitarian judgments

In addition to presenting empirical findings, we have presented a new taxonomy of utilitarian judgment that we hope will provide greater clarity. More specifically, we distinguish between five successive levels of utilitarian judgment: A judgment can be defined as utilitarian simply because of its content—because it's the judgment that promotes the greater good and that is therefore required by utilitarianism (Level 1). Beyond this definitional level, it's an empirical question whether a given Level-1 utilitarian judgment also qualifies at a higher level. A judgment can be utilitarian because it reflects aggregate cost-benefit reasoning, either without genuine concern for the greater good (Level 2) or with (Level 3). Finally, a judgment can be utilitarian because of the traits of the judge, reflecting a general commitment to utilitarian values (Level 4), or even an explicit commitment to utilitarianism (Level 5), as in the case of many professional philosophers.

It seems that much of the disagreement between us and Kahane et al. stems from differences in how the term "utilitarian judgment" is used. Kahane et al., as well several others (Rosas & Koenigs, 2014; Royzman et al., 2015; Sheskin & Baumard, 2016), employ a definition that sets a high bar for what counts a "utilitarian" judgment. They assume that for a judgment to be utilitarian, it must flow from a general commitment to utilitarian values (Level 4), if not from an explicit commitment to utilitarianism (Level 5). In other words, they assume that judgments that qualify as utilitarian must reflect a generally utilitarian mindset on the part of the judge. As explained above, we make no such assumption. When we call a judgment "utilitarian" we mean (at minimum) that it is Level-1 utilitarian—that it is the judgment that utilitarianism requires, as a matter of definition (combined with the facts of the case; for further discussion, see Amit & Greene, 2012; Greene, 2007; Greene et al., 2004; Greene et al., 2008; Greene, 2013, 2014). Once again, as we use the term, one can make a utilitarian judgment without being a utilitarian or having utilitarian traits, just as one can make an Italian meal without being Italian or having Italian traits.

Why, then, has there been so much confusion over this seemingly straightforward definitional issue? The answer, we suspect, is related to a second mismatch in understanding, in this case concerning the philosophical writings of Greene (2007, 2013, 2014) and others (de Lazari-Radek & Singer, 2017; Singer, 2005). Greene, Singer, and others have argued that empirical research examining the sacrificial dilemma judgments of ordinary people gives us insight into the "grand questions" of moral philosophy. One might wonder, then, how such research could be philosophically relevant if a judgment can count as "utilitarian" (or "deontological") while making no assumptions about the psychology of the judge. With this in mind, we now consider the relationship between ordinary people's sacrificial dilemma judgments and the "grand questions" of moral philosophy.

8.1.2. Clarifying the relationship between ordinary utilitarian judgment and utilitarian philosophy

Suppose you're a psychologist with an interest in an "ism." Authoritarianism, let's suppose. What do you do? A natural strategy, familiar from personality psychology, is to find some *authoritarians* and study their minds as they say and do authoritarian things. Your authoritarian subjects need not be perfect exemplars of their "ism." They just need to have some recognizably authoritarian traits. However, to ensure that you've got the right subjects, you must administer personality measures to ensure that your focal subjects are likely to say and do authoritarian things across various situations. If it turns out that such people lack general authoritarian tendencies, then your research will be on the "wrong track."

This is the essence of Kahane et al.'s critique, substituting "utilitarian" for "authoritarian." In their most recent work especially, Kahane et al. (2017) describe sacrificial dilemmas as a flawed tool for

identifying lay utilitarians and deontologists—a botched attempt at philosophical personality psychology. In our view, this critique reflects a misunderstanding of the scientific and philosophical rationale for studying sacrificial moral dilemmas. The goal of "trolleyology" is not to understand utilitarian or deontological philosophy by studying the minds of ordinary people who are generally committed to utilitarian or deontological values. But if that is not the goal, then what is? There are two key points of contact between ordinary people's responses to trolley dilemmas (etc.) and the philosophical tension between utilitarianism and deontology. The first and most important connection concerns the nature of deontology and its characteristic objections to utilitarianism. The second concerns the cognitive rudiments of utilitarian thought. We consider each point of contact in turn.

Sacrificial dilemmas, such as the classic *footbridge* case make utilitarian philosophy look bad. Ask a utilitarian philosopher what she truly cares about, and you will hear about fighting poverty, saving animals from needless suffering, and the like. You will not hear about pushing innocent people in front of speeding trolleys. In light of this, studying sacrificial dilemmas in hopes of better understanding utilitarianism may seem wildly misguided. Indeed, as Kahane et al. (2017) claim, these dilemmas seem to miss the positive core of utilitarian philosophy. As one commenter put it, this is like trying to understand and appreciate Italian food by eating veal Milanese—an uninspiring dish (we're told), virtually identical to Austrian veal schnitzel.[14]

This objection misses the point of focusing on sacrificial dilemmas (and reveals the limits of our culinary analogy). If, somehow, veal Milanese were to mortally threaten the general validity of Italian cuisine, then this dish would indeed demand the attention of Italian food lovers. Sacrificial dilemmas, by contrast, do pose a mortal threat to utilitarianism: If it's wrong to push the man off the footbridge, and utilitarianism requires pushing the man off the footbridge, then there must be something wrong with utilitarianism. Sacrificial dilemmas are important for understanding utilitarian philosophy, not because they nicely express what is most appealing about utilitarianism, but because they nicely express what is most *unappealing* about utilitarianism.

Although trolley dilemmas originated as part of an internal debate among deontologists (Foot, 1967; Thomson, 1986), the contemplation of morally questionable utilitarian sacrifices has played an essential role in utilitarianism's philosophical history. For example, John Rawls (1971/2005) famously faulted utilitarianism for its hypothetical willingness to enslave a minority for the "greater good" of the majority. Elizabeth Anscombe (1958), in the famous *Magistrate and the Mob* case, faulted utilitarianism for its hypothetical willingness to sacrifice an innocent person to assuage an angry mob. Bernard Williams (1973/2012), in the famous *Jim and the Indians* case, faulted utilitarianism for requiring the execution of an innocent person to prevent another person from executing many innocent people. Michael Sandel, in his bestselling introduction to ethics, *Justice*, devoted most of a chapter to sacrificial counter-examples to utilitarianism, including Romans throwing Christians to the lions (to increase the pleasure of spectators), and isolating a poor child in wretched solitude to improve the well-being of other townspeople. Likewise, Sandel discusses the *footbridge* case and other trolley variations, as do many other philosophical textbooks when discussing the pros and cons of utilitarianism (e.g. MacKinnon & Fiala, 2014, pp. 99–109; Vaughn, 2012, pp. 84–95).

Proponents of utilitarianism must address these objections. One promising strategy for handling these objections looks to science for a better understanding of the psychology behind them (Greene, 2007, 2013, 2014; Singer, 2005; de Lazari-Radek & Singer, 2017). According to Greene et al.'s (2001, 2004) dual-process theory, characteristically deontological judgments are driven by automatic emotional reactions to causing harm. Such reactions, they argue, function as moral heuristics (Baron, 1994; Sunstein, 2005)—which are useful, but inflexible—warning us against performing actions that are generally bad, but ignoring the situation-specific consequences of such actions. In computational terms, these responses are now thought to reflect *model-free* versus *model-based* modes of learning and decision-making (Crockett, 2013; Cushman, 2013; Greene, 2017), where model-free learning attaches values directly to actions based in part on their physical features, but with no explicit representation of their expected consequences. Model-based learning, by contrast, attaches values to actions based on an explicit representation of their expected consequences. Consistent with this, the emotional responses behind deontological judgments appear to be sensitive to the physical mechanism of harm, such as pushing versus hitting a switch (Greene et al., 2009; Cushman, Gray, Gaffey, & Mendes, 2012; Cushman, Young, & Hauser, 2006).

Drawing on the research described above, Greene (2007, 2013, 2014, 2017) has argued that many of the philosophical objections to utilitarianism reflect inflexible automatic responses that are sensitive to features of actions that most people—and, critically, not just utilitarians—regard as morally irrelevant. If ordinary people's amygdalae recoil at the thought of pushing a man off a footbridge (Glenn, Raine, Schug, Young, & Hauser, 2009; Shenhav & Greene, 2014), it seems likely that Judith Jarvis Thomson's (1985) and Michael Sandel's (2010) amygdalae do something similar, as suggested by the emotionally driven biases observed in philosophers (Schwitzgebel & Cushman, 2012; Schwitzgebel & Cushman, 2015). And likewise for the amygdalae of Anscombe (1958), Williams, (1973/2012), and Rawls (1971/2005)—not to mention countless readers of philosophy textbooks (MacKinnon & Fiala, 2014; Vaughn, 2012)—as they contemplate the myriad ghastly sacrifices that utilitarianism, at least hypothetically, endorses. If such emotional responses are essentially heuristic—generally useful, but also sensitive to morally irrelevant cues and blind to much morally relevant information—then we have reason to put less stock in these objections. In other words, psychology may not tell us what is right and wrong by itself, but it can help us appreciate the heuristic nature of moral intuition and give us insight into when we are placing too much trust in our moral intuitions, including intuitions that push against utilitarianism (Baron, 1994; Greene, 2007, 2013, 2014, 2017; Singer, 2005; de Lazari-Radek & Singer, 2017).

In our view, a major determinant of which philosophers end up in which camp is how they interpret their own emotional responses to the utilitarianism's most unpleasant implications. If a philosopher takes such responses to be the voice of Moral Truth, then that philosopher is not only likely to give characteristically deontological responses to sacrificial dilemmas. That philosopher is also likely to reject utilitarianism altogether and favor an alternative moral philosophy that better accommodates his/her pattern of emotional responses, such as a deontological theory (Kamm, 1998; Thomson, 1986) that attempts to justify trading one life for five in some cases (such as the *switch* case) and not others (such as the *footbridge* case). By contrast, philosophers who are more circumspect about their emotional reactions to specific actions and are relatively more concerned with producing good consequences are more likely to become utilitarians/consequentialists and eventually make (Level-4 to Level-5) utilitarian judgments in response to trolley cases (etc.). The more general point is that a scientific understanding of the strengths and limitations of our emotional responses, which are largely shared by philosophers and ordinary people, may help philosophers do better philosophy (Greene, 2007, 2013, 2014, 2017; Singer, 2005; de Lazari-Radek & Singer, 2017).

The second point of contact between utilitarian philosophy and sacrificial dilemma judgments concerns the psychology behind the utilitarian judgments of ordinary people. In calling a sacrificial judgment "utilitarian," we assume only that it is Level-1-utilitarian, that it is the option favored by utilitarianism. If sacrificial judgments were Level-1 utilitarian and nothing more, there would indeed be no meaningful contact between the psychology and the philosophy. However, according to Greene and colleagues' dual-process theory, ordinary people's Level-1 utilitarian judgments are also Level-2-utilitarian judgments, reflecting aggregate cost-benefit reasoning. This claim is well

[14] Thanks to Jim Everett (personal communication) for articulating this objection.

supported (Paxton, Ungar, & Greene, 2012, 2013; Conway & Gawronski, 2013; Greene, 2013; cf. Gawronski et al., 2017) and not particularly controversial (Kahane et al., 2015, pp. 206–7). What's more, we believe that when people engage in aggregate cost-benefit reasoning, they (unlike selfish, calculating psychopaths) actually care about the greater good within the context of the dilemma. In other words, we believe, as an empirical matter, that people's Level-1 utilitarian judgments rise not just to Level 2, but also to Level 3. The current data provide ample support for this conclusion—notably, the negative relationships between the U parameter and measures of antisociality, such as egoism and psychopathy, coupled with positive relationships between the utilitarian parameter and measures of prosociality, such as moral identity and moral conviction about harm.

The assumption that utilitarian decisions in sacrificial decisions reflect a degree of moral concern, above and beyond mere amoral cost-benefit reasoning, is a key empirical assumption in Greene's (2007, 2013, 2014) and Singer's (2005) related philosophical writings. We did not anticipate that this would become a point of empirical contention (Kahane et al., 2015). It seemed obvious to us that ordinary people, when confronted with sacrificial dilemmas such as the *footbridge* case, feel a genuine moral pull toward the greater good of saving more lives, as well as a genuine moral pull away from violently killing an innocent person. Thanks to the present results, we now have evidence to back up both of these assumptions, which we had taken for granted.

The question, then, is whether ordinary people's Level-3 utilitarian judgments are of any philosophical relevance. In other words, what does the, "modest, unremarkable, and ordinary thought that it is, ceteris paribus, morally better to save a greater number" (Kahane et al., 2015; p. 207) have to do with utilitarianism? Nothing, say Kahane et al. (2015), according to whom the "positive core" of utilitarianism is the "radical and demanding view" that one must maximize well-being at all times. We don't deny that utilitarianism is demanding and in some ways radical, but we believe that its origins, its "positive core," is more familiar. To say that utilitarianism has nothing to do with ordinary, impartial cost-benefit reasoning is like saying that science has nothing to do with the commonsense hypothesis testing employed by auto mechanics, gardeners, and police officers. Just as science is a rigorous systematization of everyday reasoning about cause and effect, utilitarianism is a rigorous systematization of ordinary, impartial cost-benefit reasoning (Mill, 1861/1998). In any case, this is an argument that one can make, and empirical studies using moral dilemmas can bolster that argument.

In sum, research using sacrificial moral dilemmas can help us understand the most compelling objections to utilitarianism, as well as the cognitive building blocks of utilitarian philosophy. Critically, none of this requires that ordinary people who make utilitarian judgments be generally committed to utilitarian values.

8.1.3. Correcting a misleading impression of utilitarianism

Kahane (2015) suggests that the widespread use of sacrificial dilemmas has distorted many people's impression of utilitarianism.[15] Sadly, we agree. As noted above, sacrificial moral dilemmas focus attention on utilitarianism's least appealing feature, namely its endorsement of any action that truly promotes the greater good, no matter how horrifying that action may be. Focusing on these troublesome cases is good, honest methodology for philosophers who wish to defend utilitarianism. But, as noted above, it's very bad public relations for the philosophy.

Utilitarianism was born in eighteenth-century England as a force for moral progress. The original utilitarian philosophers—Bentham (1789/1961), Mill (1861/1998), and Sidgwick (1907/1981)—were social reformers who argued against slavery and for free speech, free markets, public education, environmental protection, prison reform, animal rights, workers' rights, and women's rights (Driver, 2009). Modern utilitarians such as Peter Singer have made strides in reducing global poverty (Singer, 1972) and improving the treatment of animals (1975). Indeed, introductory philosophy courses typically present utilitarianism's prosocial, progressive side before piling on the objections. This is how the present authors think of utilitarianism (e.g., Greene, 2013), and therefore we, too, find it unfortunate that many people immediately associate utilitarianism with pushing people in front of speeding trolleys.

With that said, we do not follow Kahane (2015) in thinking that we should abandon research using sacrificial dilemmas or that we should refuse to label sacrificial judgments that are required by utilitarianism as "utilitarian." Sacrificial dilemmas shine a bright light on some of utilitarianism's least appealing implications, which is a double-edged sword for those of us who are sympathetic toward utilitarian ideals. These dilemmas do little to win utilitarianism quick converts. Nevertheless, understanding the psychology behind these intuitive objections may be an essential step toward defending utilitarian thinking and its more recognizably moral commitments (Greene, 2013; de Lazari-Radek & Singer, 2017).

8.1.4. Implications for the ecological validity of sacrificial dilemmas

We will take this opportunity to address another familiar criticism of sacrificial trolley dilemmas, namely their lack of realism or "ecological validity" (Bauman, McGraw, Bartels, & Warren, 2014; Kahane, 2015). While we agree that realism in psychological probes is often a worthy goal, it is by no means a requirement for conducting illuminating psychological research, as explained by Mook (1983) in his classic article, "In defense of external invalidity." Critics who dismiss trolley dilemmas for their lack of realism tend to misunderstand the scientific strategy behind their use. Trolley dilemmas are especially useful for revealing cognitive structure (Cushman & Greene, 2012). Vision scientists, for example, routinely use stimuli such as flashing black-and-white checkerboards and Gabor patches, not because they are typical visual objects, but because they are high-contrast stimuli that drive the visual system in revealing ways. Likewise, vision researchers prize visual illusions for what they reveal about how we see. Trolley dilemmas can be understood as artificially "high-contrast" moral stimuli (perhaps even illusory),[16] and are useful for precisely this reason. The processes they engage (cost-benefit reasoning, affective responses to harm) are almost certainly very common. What's unusual about trolley dilemmas is that they pit these processes against each other in a stark and reliable way. Creating this contrast requires some artificial maneuvering in the form of hypothetical story-telling, but many laboratory set-ups work in precisely this way, artificially isolating natural processes that are not, in themselves, artificial.[17]

Moreover, it is worth noting that many important real-world moral decisions, especially at the policy level, are very trolley-like, involving options that would appear to violate some right or duty, but that also promise to deliver better outcomes. These real-world problems are more complicated and uncertain than stylized trolley dilemmas, but the underlying tension is the same. Trolley dilemmas were invented in

[15] Again, it is important to note that calling sacrificial judgments 'utilitarian' does not imply that they reflect a general commitment to utilitarian values (Kahane et al., 2017).

[16] Dilemmas are analogous to illusions insofar as they generate a strongly negative affective response to an action that (by artificial stipulation) is overwhelmingly good in terms of its consequences. To a consequentialist, this is a kind of illusion—a good action emotionally disguised as a bad one. (Greene, 2013, pp. 245–54; Greene, 2017).

[17] Bauman et al. (2014) added a new twist to this objection, arguing against the use of trolley dilemmas because people often find them funny—some versions more than others. Although we agree that researchers should be aware of humorousness as a potential confound, this does not hold up as a general criticism of using sacrificial dilemmas. First, their objection is focused solely on the two most familiar trolley cases, even though the research in question uses a wide range of dilemmas that do not involve things like pushing large people off of footbridges (for example, none of the PD dilemmas contain such material). Second, and more importantly, Baumann and colleagues make no effort to explain how humorousness, as a confounding factor, can explain the wide range of dual-process effects obtained using sacrificial dilemmas, especially effects focused on controlled processing (e.g., Baron, Scott, Fincher, & Metz, 2015; Conway & Gawronski, 2013; Greene et al., 2008; Moore et al., 2008; Paxton, Bruni, & Greene, 2014; Paxton et al., 2012; Royzman et al., 2015).

hopes of clarifying the principles behind competing views on abortion (Foot, 1967), and are widely discussed in the literatures on bioethics (Kamm, 1998; Kolber, 2009), the ethics of war (Sandel, 2010), and, most recently, the ethics of robotics and self-driving cars (Wallach & Allen, 2008; Bonnefon, Shariff, & Rahwan, 2016). Judgments about sacrificial dilemmas differ between medical doctors and public health professionals, consistent with their respective goals of promoting individual vs. collective health (Ransohoff, 2011). These findings indicate that sacrificial dilemmas capture psychological tensions that matter for real-world bioethical decision-making. Doubts about the real-world relevance of trolley dilemmas are as old as trolley dilemmas themselves, but these dilemmas have persisted across decades because, over and over, people charged with solving real-world problems find them relevant and illuminating (Edmonds, 2013).

8.1.5. Implications for ordinary utilitarian thought: harm-minimization vs. outcome-maximization

Although no one has claimed that ordinary people's utilitarian judgments reflect a broad and stable commitment to utilitarian values, it remains an open question whether and to what extent these judgments reflect something more than aggregate cost-benefit reasoning in context. To begin, people who regard utilitarian sacrifices as morally acceptable tend not to regard them as morally obligatory (Lombrozo, 2009; Royzman et al., 2015). When it comes to helping people whom one has not harmed, the moral issue naturally concerns obligation rather than permissibility, since few regard voluntarily helping people as morally impermissible. Consistent with this, Kahane and colleagues find that people who approve of utilitarian sacrifices (without necessarily regarding them as obligatory) are not especially likely to be charitable. The present results confirm and clarify this pattern of results.

In Studies 3 and 4, neither conventional utilitarian judgments, nor the utilitarian PD parameter, correlated positively with hypothetical charitable donations, Identification with All Humanity, perceived obligations of the affluent to help the poor, or other indicators of prosocial dispositions. In Study 5, people with higher U scores even judged maximizing the greater good to be less mandatory. It is possible that this surprising result reflects a tendency for high-U individuals to be less judgmental than others when it comes to prosocial behaviors that they regard as optional. We note that in Study 6, when this question was framed as a matter what's acceptable instead of what's wrong, there was no correlation between the U parameter and the acceptability of failing to prioritize the greater good. Consistent with expectations, we found in Study 6 that the U parameter correlated with (a) moral identity internalization, (b) moral conviction about harm, (c) the belief that causing harm is wrong, and (d) reported focus on the group, but not the individual, in dilemmas. Putting these two sets of results together indicates that lay utilitarian thinking pertains, in a consequentialist way, primarily to the domain of moral proscriptions—avoiding infringing on others—rather than the domain of moral prescriptions—improving other's situations (Janoff-Bulman et al., 2009). In other words, lay utilitarian thinking in sacrificial dilemmas appears focused on *minimizing total harm*, rather than truly *maximizing total happiness*.

8.2. Methodological implications

Whether and how sacrificial dilemmas should be used depends on the goal and strategy of the research. Research focused on identifying and characterizing processes theorized to influence only one dilemma response tendency can reliably obtain effects using conventional dilemma analyses. The dual process model claims that deontological responses are driven primarily by automatic affective response to certain action types, whereas utilitarian responses are driven by a more controlled consideration of aggregate consequences. Accordingly, studies manipulating cognitive load to examine the influence of controlled processing (Greene et al., 2008; Trémolière, De Neys, & Bonnefon, 2012; Conway & Gawronski, 2013), or testing patients with affect-related brain lesions to assess the influence of affective processes (Koenigs et al., 2007; Ciaramelli et al., 2007) on sacrificial judgments can succeed using conventional analytic methods. In such studies, the behavioral changes elicited do not by themselves tell us which psychological processes are responsible for the change. For example, a VMPFC patient who reliably approves of utilitarian sacrifices might, in principle, have reduced affective response to causing harm (low D) or increased concern for the greater good (high U), or evince another more complex pattern. Here, researchers can draw on theoretical knowledge of the impact of VMPFC damage to conclude that the first interpretation is the most likely to be correct, and likewise for other cases.

Yet, studies that rely only on conventional analyses suffer from two interpretational concerns. First, they remain ambiguous regarding whether a given manipulation (for example) increases deontological response tendencies, reduces utilitarian response tendencies, or reflects a more complex underlying pattern. Second, they remain insensitive to any effects that simultaneously load on both parameters in the same direction (i.e., suppression). In the current work, we demonstrate many such suppression effects—cases where a given variable simultaneously predicts both deontological and utilitarian response tendencies—and these simultaneous effects cancel out for conventional dilemma analyses that treat deontological and utilitarian response tendencies as opposites. The current data indicate that such cases are far from rare, revealing suppression effects for psychopathy, rational egoism, ethical egoism, moral identity internalization, moral conviction about harm, harm wrongness, and age. In each case, interpretations based on conventional analyses would suggest there is either (a) no effect on sacrificial judgments, or (b) a modest effect in the direction of whichever PD parameter correlated more strongly. Similar suppression effects occur in many other cases for both individual difference and experimental designs. For example, aversion to witnessing harm correlated with both the D and U parameters, which cancelled out for conventional analyses (Miller et al., 2014; Reynolds & Conway, 2018). Likewise, manipulating trust vs. distrust mindsets (Conway et al., 2018)[18] or the language in which dilemmas are presented (Muda et al., 2017) can simultaneously impact both parameters in the same direction—thereby cancelling out for conventional analyses.[19]

Due to such suppression effects, researchers should use caution when interpreting conventional dilemma analyses, as they (like Kahane and colleagues) may erroneously conclude that endorsing utilitarian sacrifices merely reflects antisociality and may fail to appreciate that there are at least two distinct response tendencies behind these judgments, both of which are positively related to prosociality. Even when process dissociation is preferable on theoretical grounds, it will sometimes not be feasible, as it becomes increasingly unreliable when assessing fewer than 20 dilemmas (Conway & Gawronski, 2013).[20] In

[18] In fact, mediation analyses revealed that generalized distrust mindsets increased decisional ambivalence between the deontological and utilitarian dilemma response options, thereby increasing the tendency to select both PD parameters. These simultaneous increases cancelled out for conventional judgments, resulting in a null effect.

[19] Again, merely controlling statistically for antisocial variables cannot serve the same function, as doing so does not address the fact that conventional dilemma judgments reflect a tension between different response tendencies where measures of prosociality that load on both cancel out (i.e., suppression). Hence, controlling for antisociality does nothing to clarify correlations between prosocial measures and this fundamentally ambiguous measure.

[20] For example, Duke and Bègue (2015) found that drunk people make more utilitarian judgments on sacrificial dilemmas. The current findings suggest that this pattern reflects a reduction in concerns about causing harm among the inebriated (i.e., low D), rather than an increase in concern about maximizing good outcomes (i.e., high U). However, PD is unhelpful in this case as it seems unlikely that inebriated people would pay sufficient attention to 20 dilemmas in a row. Thus, this paper represents an excellent example of where researchers should employ conventional dilemmas but interpret findings with caution: the effect might be better described as 'the drunk anti-deontologist' rather than 'the drunk utilitarian,' although both descriptions probably attribute too much philosophical commitment to participants.

such cases, researchers should employ conventional dilemmas while bearing in mind that it is difficult to ascertain the underlying mechanism from behavior alone: John Stuart Mill and Machiavelli may both endorse utilitarian sacrifices, but for radically different reasons, and their respective motivations should not be conflated.

Finally, PD analyses, in addition to replicating past work (Conway & Gawronski, 2013), provide additional refinement of the dual process model than is possible via conventional analyses. Consider the dual-process claim that deontological responses reflect relatively more affective responses to the thought of causing harm, whereas utilitarian responses reflect relatively more deliberative processing focused on overall outcomes. One would expect people who score higher on empathic concern to experience stronger 'alarm-bell' emotions in response to the thought of directly causing harm (Davis,1983), leading to decreased utilitarian judgment in sacrificial dilemmas (Greene, 2007). It's less clear whether and to what extent high empathic concern might also favor utilitarian sacrifices, reflecting concern for the group of people who would benefit from the sacrifice. PD helps resolve this ambiguity. Different combinations of response tendencies can produce the same judgments in response to standard dilemmas, but PD helps determine which combinations are at work. As suggested above, empathic concern could correlate positively with both the U and D parameters, but more strongly with D than with U. Alternatively, empathic concern could relate uniquely to the tendency to reject causing harm in moral dilemmas (the D parameter) independent of any tendency to maximize positive outcomes (the U parameter). The present results indicate that it's the latter (at least among non-philosophers), an insight that would not be possible without PD.

Moreover, in the current work, the D parameter uniquely correlated with concern for the individual in the dilemmas (Study 6), whereas the U parameter uniquely correlated with concern for the overall group. This pattern clearly corroborates the dual-process claim that affective reactions to the thought of harming a single individual in sacrificial dilemmas motivated harm rejection, whereas focusing on the wellbeing of the overall group motivates acceptance of sacrifices. Thus, critics of dual-process theory must explain this nuanced pattern of results. That said, recent work suggests that the dual process model may not be exhaustive, and that other processes may also contribute to dilemma responses above and beyond the two postulated in dual-process theory (e.g., Gamez-Djokic & Molden, 2016; Reynolds & Conway, 2018; Rom & Conway, 2018). A critical challenge for the field will be to clarify which processes most powerfully contribute to dilemma responses beyond the two postulated by dual-process theory.

8.3. Conclusion

Sacrificial moral dilemmas remain useful tools for studying moral judgment and decision-making. What's more, there is nothing wrong with calling the endorsement of harm-minimizing sacrifices "utilitarian." Such judgments qualify as utilitarian not only because they favor the greater good and must be defended by utilitarians, but also because they reflect genuinely moral cost-benefit reasoning, both in the minds of professional philosophers and ordinary people. The disagreements discussed here largely reflect conceptual differences over the definition of "utilitarian judgment," as well as the methodological limitations of conventional analyses, which underestimate the extent to which ordinary people's sacrificial utilitarian judgments have prosocial motivations. By combining sacrificial moral dilemmas with process dissociation, these cognitive probes become more precise tools for assessing individual differences in moral thinking.

Acknowledgements

This research was funded by a Gottfried Wilhelm Leibniz Award by the German Research Foundation (DFG) awarded to Thomas Mussweiler (MU 1500/5-1) and a First Year Assistant Professor Award from Florida State University to the first author.

Appendix A. Process dissociation calculations for moral dilemma judgments

Calculating the deontology and utilitarian PD parameters requires examining responses to both congruent and incongruent dilemmas. Utilitarianism entails maximizing overall outcomes, whereas deontology entails avoiding causing harm regardless of outcomes. Harmful action maximizes overall outcomes in the incongruent, but not congruent, dilemmas. Therefore, utilitarianism and deontology lead to different response patterns across dilemma variants. Consider the processing tree depicted in Fig. 1: The top path illustrates the case where utilitarianism drives the response to a dilemma, which entails rejecting harm for congruent dilemmas but accepting harm for incongruent dilemmas. The second path illustrates the case where deontology drives the response to a dilemma, which entails rejecting harm for both congruent and incongruent dilemmas. Finally, the bottom path represents the case where neither utilitarianism nor deontology drives the response to a dilemma; this case entails accepting harm for both congruent and incongruent dilemmas.

Using the two columns on the right side of the figure, it is possible to work backward to determine which cases led participants to judge harm as acceptable or unacceptable for both congruent and incongruent dilemmas. For congruent dilemmas, harm is unacceptable when either utilitarianism drives the response, U, or when deontology drives the response, $(1 - U) \times D$. Conversely, harm is acceptable on congruent dilemmas when neither utilitarianism nor deontology drives the response, $(1 - U) \times (1 - D)$. For incongruent dilemmas, harm is unacceptable when deontology drives the response, $(1 - U) \times D$. Conversely, harm is acceptable either when utilitarianism drives the response, U, or when neither utilitarianism nor deontology drives the response, $(1 - U) \times (1 - D)$.

By combining these cases, it becomes possible to algebraically represent the probability of a particular judgment. For example, the probability of judging harm as unacceptable for congruent dilemmas is represented by the case where either utilitarianism drives responses or deontology drives responses:

$$p(\text{unacceptable}|\text{congruent}) = \text{U} + [(1-\text{U}) \times \text{D}] \quad \text{(A.1)}$$

Conversely, the probability of judging harm as acceptable in congruent dilemmas is represented by the case that neither utilitarianism nor deontology drives responses:

$$p(\text{acceptable}|\text{congruent}) = (1-\text{U}) \times (1-\text{D}) \quad \text{(A.2)}$$

For incongruent dilemmas, the probability of judging harm as unacceptable is represented by the case that deontology drives responses:

$$p(\text{unacceptable}|\text{incongruent}) = (1-\text{U}) \times \text{D} \quad \text{(A.3)}$$

Conversely, the probability of judging harm as acceptable for incongruent dilemmas is represented by the cases that utilitarianism drives responses, or neither deontology nor utilitarianism drives responses:

$$p(\text{acceptable}|\text{incongruent}) = U + [(1-U) \times (1-D)] \quad (A.4)$$

Once the probabilities of accepting and rejecting harm in congruent and incongruent dilemmas are represented algebraically, it becomes possible to enter a participants' pattern of actual responses across multiple congruent and incongruent dilemmas, and algebraically combine these equations in order to solve for two parameters estimating deontological (*D*) and utilitarian (*U*) inclinations underpinning their responses. In particular, by including Eq. (A.3) into Eq. (A.1), the latter can be solved for *U*, leading to the following formula:

$$U = p(\text{unacceptable}|\text{congruent}) - p(\text{unacceptable}|\text{incongruent}) \quad (A.5)$$

Moreover, by including the calculated value for *U* in Eq. (A.3), this equation can be solved for *D*, leading to the following formula:

$$D = p(\text{unacceptable}|\text{incongruent})/(1-U) \quad (A.6)$$

Together, these formulas enable researchers to obtain parameters that independently estimate the strength of deontological and utilitarian inclinations underlying conventional moral dilemma judgments.

Appendix B. Updated greater good vignettes

Here we present versions of the Greater Good vignettes adapted from Kahane et al. (2015, Study 4) that we employed in Study 5. Whereas in Study 4 we employed the original vignette and question wording employed by Kahane and colleagues, we made two key changes to these versions. First, we reworded each question to assess whether or not participants view performing the described action as acceptable, despite clearly specified reduced overall outcomes. Second, we clarified the wording and details of each question to reduce the possibility that participants would make additional inferences that undermine the intent of the question. For example, in the Janet question, we clarified that forsaking volunteer work to attend one's mother ensures that net negative outcomes will occur (whereas previously one might surmise that the best overall outcome entails visiting one's mother if doing so makes her happy and the charity work is of little import).

Instructions: Please read each of the following scenarios and indicate how acceptable you think the action is. There are no right or wrong answers; we are just interested in your personal opinion. Response scale: 1 = *Not at all acceptable*; 7 = *Completely acceptable.*

B.1. Benjamin's car or donation

Benjamin is a college student who has been taking the bus for years. He always wanted to have his own car, but could never afford it while paying for his education. One semester, Benjamin's uncle passed away and left him $7000—enough money to purchase a decent used car (spending less is unwise as cheap cars soon require expensive repairs). On his way to the used car dealership, Benjamin heard a news broadcast about a Tsunami in South East Asia that has left thousands dead, wounded, and homeless. There is a website where Benjamin could make a donation to help these victims. Benjamin contemplates donating the $7000. He knows that such a donation would save the lives of several of the hospitalized tsunami victims. However, he would have to continue living without a car.

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

B.2. Anti-Malaria Foundation

John works as a waiter at a local family restaurant. One day, John gets a letter in the mail from the "Against Malaria Foundation," telling him that a donation of $100 will save a child's life in Africa. John considers sending in a donation, but he has also been saving up for a new mobile phone (his old one doesn't work well anymore). In his last paycheck, John finally made the last $100 of spending money he needs for the phone.

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

B.3. Janet and her mother

Janet's mother is feeling really lonely and would like Janet to come and spend the day with her. Janet would like to cheer up her mother, but she had already agreed to spend the day volunteering for Habitat for Humanity, a charity organization that builds houses for poor families. As Head Engineer, if Janet does 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, Janet would helping a family in need for a long time, rather than cheering up her mother for one day.

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

B.4. Vegetarianism

Kathleen recently read a book describing the methods of modern factory farming in gruesome detail. The book persuaded her 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 Kathleen considers becoming vegetarian. On the other hand, Kathleen thinks that humans are more important than animals, it is natural for humans to eat meat, and Kathleen really likes eating meat.

How acceptable would it be for Kathleen to continue eating meat, even though she knows this means animals will continue to suffer for her dining pleasure?

B.5. One vs. many donation

Mark is an American businessman who plans to donate $1000 to help sick children. He is currently deciding which of two charities he should

donate to-he 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 Mark donated $1000 to the American charity, his 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 Mark donated $1000 to the Sub-Saharan charity, his donation would purchase drug treatment regimens that would save the lives of ten underprivileged African children with measles.

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

B.6. Veronica's comfortable lifestyle

Veronica has written several successful novels, and now has an ample income. She has worked hard for this income, and feels proud of her accomplishments. Veronica has enough money to live a comfortable life, with money left over for things like vacations, personal staff, and the latest gadgets. She views these luxuries as the fruit of many years of labor. Veronica realizes that she could give away large sums of money to charity and still be reasonably happy if she reduced the number of luxuries in her life. If she did so, Veronica would save dozens of underprivileged people from poverty, illness, and even death. However, this means Veronica would have to give up many of the special luxuries she worked so hard to achieve.

How acceptable would it be for Veronica to continue enjoying her luxurious lifestyle instead of giving away large sums of money to charity, even though this means that many underprivileged people will continue to suffer from poverty, illness, and death?

B.7. Firefighter

Albert is a firefighter who is rescuing people from a burning building. The building is about to collapse, so the firefighters are rushing out. Albert is the last firefighter in the building, and will only have time to rescue one more person on his way out. In the last room, Albert finds two people trapped. He immediately recognizes 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, Albert realizes that the second person is his own mother—a poor, uneducated housekeeper. Albert's mother is not important for world peace, but she was always a good mother to him. Now Albert must choose whether he should he save the famous peace negotiator or his own mother in the few seconds before the building collapses.

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

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2018.04.01.

References

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions.* Sage.

Amazon (2015). https://requester.mturk.com/ Retrieved on April 1, 2015.

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 23*, 861–868. http://dx.doi.org/10.1177/0956797611434965.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy, 33*, 1–19.

Aquino, K., & Reed, A., II (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*, 1423–1440. http://dx.doi.org/10.1037//0022-3514.83.6.1423.

Armstrong, J., Friesdorf, R., & Conway, P. (2018). Clarifying gender differences in moral dilemma judgments: The complementary roles of harm aversion and action aversion. *Social Psychological and Personality Science* doi:1948550618755873.

Barbosa, S., & Jiménez-Leal, W. (2017). It's not right but it's permitted: Wording effects in moral judgement. *Judgment and Decision Making, 12*, 308–313.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences, 17*(1), 1–10.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*, 265–284.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154–161. http://dx.doi.org/10.1016/j.cognition.2011.05.010.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*, 536–554.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323.

Bentham, J. (1789/1961). An introduction to the principles of morals and legislation. In *Utilitarianism* (pp. 7–398). Garden City, NY: Doubleday.

Bernhard, R. M., Chaponis, J., Siburian, R., Gallagher, P., Ransohoff, K., Wikler, D., et al. (2016). Variation in the oxytocin receptor gene (OXTR) is associated with differences in moral judgment. *Social Cognitive and Affective Neuroscience*. http://dx.doi.org/10.1093/scan/nsw103.

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

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies, 170*, 465–500.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116.

Carney, D. R., & Mason, M. F. (2010). Decision making and testosterone: When the ends justify the means. *Journal of Experimental Social Psychology, 46*, 668–671.

Christov-Moore, L., Conway, P., & Iacoboni, M. (2017). Deontological judgments in moral dilemmas are grounded in sensorimotor representations of harm to others. *Frontiers in Integrative Neuroscience, 11*. http://dx.doi.org/10.3389/fnint.2017.00034.

Ciaramelli, E., Muccioli, M., Làdavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social cognitive and affective neuroscience, 2*, 84–92.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision-making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235. http://dx.doi.org/10.1037/a0031021.

Conway, P., Weiss, A., Burgmer, P., & Mussweiler, T. (2018). Distrusting your moral compass: The impact of distrust mindsets on moral dilemma processing and judgments. *Social Cognition*.

Cooper, M. J., & Pullig, C. (2013). I'm number one! Does narcissism impair ethical judgment even for the highly religious? *Journal of Business Ethics, 112*, 167–176.

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

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

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

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

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social neuroscience, 7*(3), 269–279.

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

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113–126. http://dx.doi.org/10.1037/0022-3514.44.1.113.

de Lazari-Radek, K., & Singer, P. (2017). *Utilitarianism: A very short introduction.* Oxford University Press.

Djerioual, H., & Trémolière, B. (2014). The Dark Triad of personality and utilitarian moral judgment: The mediating role of Honesty/Humility and Harm/Care. *Personality and Individual Differences, 67*, 11–16.

Driver, J. (2009). The history of utilitarianism. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from: < http://plato.stanford.edu/archives/

sum2009/entries/utilitarianism-history/ > .
Duke, A. A., & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition, 134*, 121–127.
Edmonds, D. (2013). *Would you kill the fat man?: The trolley problem and what your answer tells us about right and wrong.* Princeton University Press.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.
Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin, 42*, 696–713. http://dx.doi.org/10.1177/0146167215575731.
Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., et al. (2010). Gender-related differences in moral judgments. *Cognitive processing, 11*, 219–226.
Gamez-Djokic, M., & Molden, D. (2016). Beyond affective influences on deontological moral judgment: The role of motivations for prevention in the moral condemnation of harm. *Personality and Social Psychology Bulletin, 42*, 1522–1537.
Gawronski, G., Conway, P., Friesdorf, R., Armstrong, J., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology* Early Online Access.
Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLOS ONE, 8*, 1–9. http://dx.doi.org/10.1371/journal.pone.0060418.
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.
Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.). *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development.* MIT Press.
Greene, Joshua. (2013). *Moral tribes: Emotion, reason, and the gap between us and them.* New York, USA: Penguin.
Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics, 124*(4), 695–726.
Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*.
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.
Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154. http://dx.doi.org/10.1016/j.cognition.2007.11.004.
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*, 389–400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.
Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking more or feeling less? Explaining the foreign-language effect on moral judgment. *Psychological Science, 28*, 1387–1397.
Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.
Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*, 521.
Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*. http://dx.doi.org/10.1080/17470919.2015.1023400.
Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., et al. (2017). Beyond sacrificial harm: A Two-dimensional model of utilitarian psychology. *Psychological Review*. http://dx.doi.org/10.1037/rev0000093 Advance online publication.
Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209. http://dx.doi.org/10.1016/j.cognition.2014.10.005.
Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & language, 25*, 561–582.
Kamm, F. M. (1998). *Morality, mortality, vol. I: Death and whom to save from it.* New York, USA: Oxford University Pres.
Kant, I. (1785/1959). *Foundation of the metaphysics of morals* (L. W. Beck, Trans.). Indianapolis: Bobbs-Merrill.
Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2011). Utilitarian moral judgment in psychopathy. *Social cognitive and affective neuroscience, 7*, 708–714.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446*, 908–911. http://dx.doi.org/10.1038/nature05631.
Kolber, A. (2009). The organ conscription trolley problem. *The American Journal of Bioethics, 9*, 13–14.
Lee, J. J., & Gino, F. (2015). Poker-faced morality: Concealing emotions leads to utilitarian decision making. *Organizational Behavior and Human Decision Processes, 126*, 49–64.
Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of personality and social psychology, 68*, 151.
Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33*, 273–286.
MacKinnon, B., & Fiala, A. (2014). *Ethics: Theory and contemporary issues*. Nelson Education.
McFarland, S., Webb, M., & Brown, D. (2012). All humanity is my ingroup: A measure and studies of identification with all humanity. *Journal of Personality and Social Psychology, 103*, 830–853.
Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology, 18*, 193–197.
Mill, J. S. (1861/1998). *Utilitarianism*. In R. Crisp (Ed.), New York: Oxford University Press.
Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion, 14*, 573.
Montoya, E. R., Terburg, D., Bos, P. A., Will, G. J., Buskens, V., Raub, W., et al. (2013). Testosterone administration modulates moral judgments depending on second-to-fourth digit ratio. *Psychoneuroendocrinology, 38*, 1362–1369.
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379.
Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*, 549–557. http://dx.doi.org/10.1111/j.1467-9280.2008.02122.x.
Muda, R., Niszczota, P., Bialek, M., & Conway, P. (2017). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 44*, 321–326. http://dx.doi.org/10.1037/xlm0000447.
Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*, 530–542. http://dx.doi.org/10.1016/j.cognition.2005.07.005.
Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisfying to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.
Park, G., Kappes, A., Rho, Y., & Van Bavel, J. J. (2015). *At the heart of morality lies neurovisceral integration: Lower cardiac vagal tone predicts utilitarian moral judgment.* Available at SSRN 2662845.
Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience, 9*, 94–107.
Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5*, 501.
Paxton, J. M., Bruni, T., & Greene, J. D. (2013). Are 'counter-intuitive'deontological judgments really counter-intuitive? An empirical reply to. *Social cognitive and affective neuroscience, 9*, 1368–1371.
Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are 'counter-intuitive' deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Social Cognitive and Affective Neuroscience, 9*, 1368–1371.
Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177.
Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *European Review of Social Psychology, 20*, 272–314. http://dx.doi.org/10.1080/10463280903162177.
Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science, 13*, 265–270.
Ransohoff, K. (2011). *Patients on the trolley track* (Undergraduate thesis)Cambridge, MA: Harvard University.
Rawls, J. (1971/2005). *A theory of justice*. Cambridge, MA: Belknap Press.
Reed, A., II, & Aquino, K. F. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of Personality and Social Psychology, 84*, 1270.
Reynolds, C. J., & Conway, P. (2018). Not just bad actions: Affective concern for bad outcomes contributes to moral condemnation of harm in moral dilemmas. *Emotion*.
Robinson, J. S., Joel, S., & Plaks, J. E. (2015). Empathy for the group versus indifference toward the victim: Effects of anxious and avoidant attachment on moral judgment. *Journal of Experimental Social Psychology, 56*, 139–152.
Rom, S., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*.
Rosas, A., & Koenigs, M. (2014). Beyond "utilitarianism": Maximizing the clinical impact of moral judgment research. *Social neuroscience, 9*(6), 661–667.
Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science, 39*, 325–352.
Sandel, M. J. (2010). *Justice: What's the right thing to do?* London: Penguin Books.
Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27*, 135–153.
Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141*, 127–137.
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.
Sheskin, M., & Baumard, N. (2016). Switching away from utilitarianism: The limited role of utility calculations in moral judgment. *PloS one, 11*, e0160084.
Sidgwick, H. (1907/1981). *The methods of ethics* (7th ed.). Indiana, US: Hackett Publishing.
Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs, 1*, 229–243.
Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals.* New York: Review/Random House.
Singer, P. (2005). Ethics and intuitions. *Journal of Ethics, 9*, 331–352.
Skitka, L. J. & Morgan, G. S. (2014). The social and political implications of moral

conviction. In H. Lavine (Ed.), *Advances in Political Psychology, 35*, 95–110.
Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*, 895.
Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences, 28*, 531–573.
Thomson, J. J. (1986). Killing, letting die, and the trolley problem. In W. Parent (Ed.), *Rights, restitution, and risk: Essays in moral theory*. Cambridge: Harvard University Press.
Thomson, J. J. (2008). Turning the trolley. *Philosophy & Public Affairs, 36*(4), 359–374.
Trémolière, B., De Neys, W., & Bonnefon, J. F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition, 124*, 379–384.
Vaughn, L. (2012). *Doing ethics: Moral reasoning and contemporary issues* (3rd ed.). New York, NY: W.W. Norton & Co.
Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate reflects decreased aversion to harming in counterintuitive utilitarian judgment. *Cognition, 126*, 364–372.
Williams, B. (1973/2012). A critique of utilitarianism. In G. Sher (Ed.), *Ethics: Essential readings in moral theory*. Routledge.