Reasoning Supports Utilitarian Resolutions to Moral Dilemmas Across Diverse Measures

Indrajeet Patil, Micaela Maria Zucchelli, Wouter Kool, Stephanie Campbell, Federico Fornasier, Marta Calò, Giorgia Silani, Mina Cikara, and Fiery Cushman


CITATION
Reasoning Supports Utilitarian Resolutions to Moral Dilemmas Across Diverse Measures

Indrajeet Patil
Harvard University

Micaela Maria Zucchelli
University of Bologna

Wouter Kool
Harvard University and Washington University in St. Louis

Stephanie Campbell
Harvard University

Federico Fornasier and Marta Calò
University of Trieste

Giorgia Silani
University of Vienna

Mina Cikara and Fiery Cushman
Harvard University

Sacrificial moral dilemmas elicit a strong conflict between the motive to not personally harm someone and the competing motive to achieving the greater good, which is often described as the “utilitarian” response. Some prior research suggests that reasoning abilities and deliberative cognitive style are associated with endorsement of utilitarian solutions, but, as has more recently been emphasized, both conceptual and methodological issues leave open the possibility that utilitarian responses are due instead to a reduced emotional response to harm. Across 8 studies, using self-report, behavioral performance, and neuroanatomical measures, we show that individual differences in reasoning ability and cognitive style of thinking are positively associated with a preference for utilitarian solutions, but bear no relationship to harm-relevant concerns. These findings support the dual-process model of moral decision making and highlight the utility of process dissociation methods.

Keywords: reasoning, dual-process model, utilitarianism, process dissociation, moral dilemmas

Supplemental materials: http://dx.doi.org/10.1037/pspp0000281.supp

Moral dilemmas are a central feature of the human condition: too often, we can only prevent a large harm by committing a smaller one. Dilemmas of this kind arise in personal settings (Whether to abort a fetus to save the mother), corporate settings (Should products be tested on live animals to prevent possible harm to consumers?), and political settings (Should we risk civilian lives to bring a swifter end to war?). These dilemmas admit of two basic solutions, each associated with a broad school of philosophical thought: the utilitarian response that favors maximizing welfare by any means (Mill, 1998) and the deontological response that often forbids causing harm, especially instrumentally (Kant, 2005).

Over the last two decades, psychologists have devoted intense theoretical and empirical effort to understanding the processes underlying these competing motives, as well as the process that adjudicates between them. These studies typically use hypothetical...
vignettes that pose a dilemma between harming a few people to save a larger number of individuals from harm (Christensen & Gomila, 2012). Most prominently, the dual-process model (DPM; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) posits that two systems support moral decision-making: (a) an automatic and emotionally grounded system that favors deontological decisions and (b) a deliberative reasoning system that supports utilitarian decisions. The process underlying deontological decisions is assumed to be automatic, affective, and resource independent, whereas the process underlying utilitarian responses is assumed to be deliberate, cognitive, and effortful. In the specific case of sacrificial dilemmas, these systems render competing motives, resulting in a response conflict (Greene et al., 2004). It is important to note, however, that the DPM proposes that these two processes operate independently and are not inversely proportional to each other (Conway & Gawronski, 2013). Thus, one can find it morally acceptable to endorse the utilitarian option within sacrificial moral dilemmas even if they are better at cognitive deliberation (e.g., abstract reasoning, problem solving) or because they have reduced harm aversion (Conway & Gawronski, 2013).

Although much evidence implicates emotional arousal in supporting nonutilitarian/deontological moral judgments (for a review, see Greene, 2014; R. Miller & Cushman, 2013), the evidence implicating reasoning in utilitarian moral judgments is far less consistent (see below). This is the point of departure for our research: as a matter of fact, are individual differences in the tendency to render utilitarian moral judgments due to individual differences in the capacity and propensity for reasoning, or not? If yes, is this association independent of harm-relevant concerns?

To pose this question requires a basic operational understanding of “reasoning.” Although there are many variants of DPMs (Dolan & Dayan, 2013; Epstein, 1994; Evans, 2008; Evans & Stanovich, 2013; Kahneman & Frederick, 2005; Sloman, 1996), generically, all models distinguish between an automatic, parallel, and effortless thinking system (intuitive/heuristic system) and a deliberate, sequential, and effortful thinking system (analytical/deliberative/reflexive system). Here, we use the term reasoning to refer to explicit, reflective, and conscious information processing (Evans & Stanovich, 2013). Reasoning is a mode of thinking that has domain-general involvement in solving novel problems, making decisions, or arriving to a conclusion in the absence of simple evolutionary programs (instincts/reflexes) or a previously learned response (habits; Evans, 2017). In the context of moral dilemmas, reasoning may involve either using consciously available rules (e.g., “always try to do what is best for the most people”; Bennis, Medin, & Bartels, 2010; Nichols & Mallon, 2006) or a systematic cost-benefit analysis (five lives saved is better than one”; Bartels, 2008). Our strategy is to investigate reasoning in this broad and general sense by aggregating across diverse measures. This is an important but preliminary step that, if successful, would lay the foundation for more detailed analyses of the contribution of reasoning to moral judgment.

In sum, we investigate the association between the capacity and propensity for domain-general reasoning and the judgment of sacrificial moral dilemmas pitting the welfare-maximizing option against physical harm to another. Across a series of eight studies (N = 4,204), we find that reasoning is associated with utilitarian resolutions to moral dilemmas but is independent of an aversion to carrying out harm. We begin by reviewing the extant evidence that bears on this question, noting several conceptual and methodological limitations.

**Manipulations of Reasoning**

Several prior studies explore the role of reasoning in utilitarian moral judgment by taxing executive resources necessary for deliberation (Patil & Trémolière, 2018; Trémolière, De Nys, & Bonnefon, 2018). This is done in several ways:

1. **Time pressure manipulation**: limiting the amount of time available to provide moral judgments (Cummins & Cummins, 2012; Rosas & Aguilar-Pardo, 2019; Suter & Hertwig, 2011).

2. **Cognitive load manipulation**: taxing working memory capacity through another concurrent cognitively demanding task (Conway & Gawronski, 2013; De Nys & Bialek, 2017; Gawronski, Armstrong, Conway, Friesdorf, & Hütter, 2017; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008), or exhausting cognitive resources via sleep deprivation (Killgore et al., 2007; Tempesta et al., 2012) or with a prior sequential cognitive depletion task (Timmons & Byrne, 2019), or leading people to think about their mortality (Trémolière, Nys, & Bonnefon, 2012), or, the reverse, easing up the cognitive load by presenting efficient kill-save ratios (Trémolière & Bonnefon, 2014).

3. **Cognitive priming manipulation**: nudging participants to use a deliberative thinking mode (vs. “feelings thinking mode”; Li, Xia, Wu, & Chen, 2018), priming analytical thinking mode by asking them to solve mathematical puzzles before performing the moral judgment task (Kvaran, Nichols, & Sanfey, 2013), or presenting dilemmas written in hard-to-read (disfluent) fonts (Spears, Fernández-Linsenbarth, Okan, Ruz, & González, 2018) to trigger analytic thinking.

The DPM predicts that, by diminishing reliance on the cognitive system, these manipulations will reduce utilitarian moral judgments. Yet existing studies exhibit a mixed pattern: Some studies find the predicted pattern (Białek & De Neys, 2016, 2017; Byrd & Conway, 2019; Conway & Gawronski, 2013; Cummins & Cummins, 2012; De Nys & Bialek, 2017; Kvaran et al., 2013; Li et al., 2018; Spears et al., 2018; Suter & Hertwig, 2011; Timmons & Byrne, 2019; Trémolière & Bonnefon, 2014), whereas others do not (Bago & De Neys, 2019; Baron & Gürcay, 2017; Gawronski et al., 2017; Greene et al., 2008; Gürcay & Baron, 2017; Killgore et al., 2007; Lane & Sulikowski, 2017; Rosas & Aguilar-Pardo, 2019; Tempesta et al., 2012; Tinghög et al., 2016). Thus, evidence from extraneous manipulation of cognitive resources provides inconsistent support for the reasoning–utilitarian association.

**Individual Differences in Reasoning**

Some prior research also asks how individual differences in self-reported cognitive style relate to utilitarian moral judgments. The DPM predicts that higher scores on such measures will be associated with utilitarian responses to moral dilemmas.
To this end, various self-report measures have been used, along with a few behavioral tasks. These include Need for Cognition (NFC; Cacioppo & Petty, 1982), the Cognitive Reflection Test (CRT; Frederick, 2005), Actively Open-Minded Thinking (AOT; Stanovich & West, 1997), and Rational Decision-Making Style (Scott & Bruce, 1995), among others. We will discuss these measures in more detail later, but for now it suffices to say that the DPM prediction has received mixed support from this line of inquiry as well. Many studies do find a positive association between reasoning measures and utilitarian tendencies (Aktas, Yilmaz, & Bahcekapili, 2017; Bartels, 2008; Byrd & Conway, 2019; Conway, Goldstein-Greenwood, Polacek, & Greene, 2018; Paxton, Bruni, & Greene, 2014; Paxton, Ungar, & Greene, 2012; Royzman, Landy, & Leeman, 2015; Wiech et al., 2013), but others do not (Attie & Knobe, 2019; Kahane et al., 2018; McNair, Okan, Hadjichristidis, & de Bruin, 2018), and some provide mixed or inconsistent findings (Baron, Scott, Fincher, & Emlen Metz, 2015; McPhetres, Conway, Hughes, & Zuckerman, 2018).

**Neuroimaging Functional and Structural Correlates of Reasoning**

The dorsolateral prefrontal cortex (dlPFC) has been argued to play a critical role in controlled and abstract reasoning processes (Kroger et al., 2002; MacDonald, Cohen, Stenger, & Carter, 2000; E. K. Miller & Cohen, 2001). More specifically, the dlPFC has been implicated in various aspects of domain-general reasoning: executive functioning (Barbey, Colom, & Grafman, 2013; Menon & Uddin, 2010; Minzenberg, Laird, Thelen, Carter, & Glahn, 2009; Nowrangi, Lyketsos, Rao, & Munro, 2014), cognitive control (Cieslik et al., 2013; Cohen, 2005; Greene et al., 2004; Metuki, Sela, & Lavidor, 2012; E. K. Miller & Cohen, 2001; Weissman, Perkins, & Woldorff, 2008; Wiegand, Sommer, Nieratschker, & Plevnia, 2019), goal-directed planning (Asplund, Todd, Snyder, & Marois, 2010; Botvinick & An, 2009; Kaller, Rahm, Spreeer, Weiller, & Unterrainer, 2011), cost-benefit analysis (Basten, Biele, Heekeren, & Fiebach, 2010), problem solving (Ruh, Rahm, Unterrainer, Weiller, & Kaller, 2012), counterfactual reasoning (Van Hoec, Watson, & Barbey, 2015), model-based control (Fermin et al., 2016; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013), and so on.

More specifically, in the domain of moral judgment and decision-making, the dlPFC has been associated with the ability to (a) facilitate abstract reasoning (e.g., cost-benefit analyses), and (b) provide cognitive control to override strong social-emotional responses elicited by the aversive nature of moral dilemmas (Buckholz & Marois, 2012; Greene et al., 2004; Tassy, Oullier, Cermolacce, & Wicker, 2009). Some prior fMRI studies have shown that the dlPFC exhibits increased activation during utilitarian (vs. nonutilitarian) moral judgments (Cushman, Murray, Gordon-McKeon, Wharton, & Greene, 2012; Glenn, Raine, Schug, Young, & Hauser, 2009; Greene et al., 2004; but see Hutcherson, Montaser-Khouhsari, Woodward, & Rangel, 2015), whereas neurostimulation studies have provided mixed evidence about dlPFC’s causal role in utilitarian moral judgment (Jeutissen, Sack, Roebroeck, Russ, & Pascual-Leone, 2014; Kuehne, Heimrath, Heinze, & Zaehele, 2015; Tassy et al., 2012; Zheng, Lu, & Huang, 2018).

Of particular importance to the current work, interrupting dlPFC activity (using neurostimulation techniques) leads to worsened performance on tasks requiring analytical thinking, such as the CRT and model-based control (Oldrati, Patricelli, Colombo, & Antonietti, 2016; Smittenaar et al., 2013). Similarly, individuals with higher cortical thickness or greater gray matter volume at dlPFC are better at strategic moral reasoning (Yamagishi et al., 2016) and exhibit better performance on the CRT (Yang et al., 2014) and executive function tasks (Yuan & Raz, 2014). Thus, there is convergent evidence from functional and structural MRI studies to implicate dlPFC in domain-general reasoning in both moral and nonmoral domains.

It is currently unknown, however, how individual anatomical differences in the dlPFC relate to utilitarian moral judgment. Recent work indicates that interindividual differences in measures of brain anatomy (such as gray matter volume, cortical thickness, etc.) can be used as a predictor for human behavior and cognition (Breukelaar et al., 2017; Kanai & Rees, 2011; Lerch et al., 2017; Nastro, Müller, Reid, & Eickhoff, 2017; Seidlitz et al., 2018; Takeuchi et al., 2017; Wagstyl & Lerch, 2018). Accordingly, we also explore the neuroanatomical basis of individual differences in utilitarian moral judgment.

**Conceptual Challenges**

Given the volume of empirical work inspired by the dual process model, it is remarkable that we remain so uncertain about several of its core claims. This owes, in part, to underlying conceptual issues prevalent in the existing literature. As noted above, a major conceptual issue in this body of work is that utilitarian and deontological inclinations are treated as mutually exclusive psychological processes, when in fact the theory states that they are independent and nonexclusive processes (Conway & Gawronski, 2013). (The “conflict” between them in many studies is forced only by the structure of the sacrificial dilemmas used as stimuli.) This point has two important consequences. First, even if reasoning supports utilitarian moral judgment, individual differences in the propensity for utilitarian moral judgment in sacrificial dilemmas could be determined mostly by individual differences in the competing motive to avoid direct, personal harm. Second, and of greatest relevance to the present work, evidence that individual differences in reasoning correlate with individual differences in utilitarian moral judgments in sacrificial dilemmas need not imply that reasoning contributes principally to cost-benefit analysis. An alternative possibility is that reasoning inhibits the strong emotional response to the harm involved in the moral dilemma (Cushman, Gray, Gaffey, & Mendes, 2012). In other words, the existing work argues that better reasoners are more likely to endorse utilitarian responses on moral dilemmas because they are more adept at weighing the cost of harming someone with the benefit of saving more lives. But the manner in which these studies have indexed utilitarianism makes these findings compatible with the alternative explanation that better reasoners are simply less emotionally responsive to the aversive nature of harm (Bartels & Pizarro, 2011). We resolve this conceptual ambiguity by relying on the process dissociation (PD) method (Conway & Gawronski, 2013; Conway et al., 2018; Gawronski et al., 2017; Reynolds & Conway, 2018) in several of the studies reported here. This method (discussed in depth later) allows researchers to independently assess individual differences in cost-benefit analysis versus aversion to harm.
A second conceptual issue is the narrow definition of utilitarianism often implicitly adopted in the contemporary psychological literature (Gawronski & Beer, 2016; Kahane et al., 2018). In many contexts, utilitarianism is operationalized exclusively as a willingness to engage in instrumental harm for the greater good. This is because utilitarianism has been studied principally in the context of sacrificial dilemmas (e.g., asking participants if they would push a man to his death in order to save five others from a runaway train). A second component of utilitarianism is impartial beneficence, which signifies that we should help others as much as we can from a completely impartial perspective, giving no special weight to ourselves or to our family or friends. The present article involves many studies of instrumental harm, but no direct study of impartial beneficence; thus, our results hold implications for individual differences in impartial beneficence only indirectly (Capraro, Everett, & Earp, 2019).

Methodological Issues

In addition to these conceptual issues, several methodological issues are present in existing research, each of which we attempt to overcome here:

1. Often a limited number of moral dilemmas is used, which makes it difficult to generalize observed results to the general class of dilemmas from which specific examples are drawn.

2. Related to (1), general linear models are used to draw inferences for a population of subjects, but not items. Given that the behavioral data from moral dilemma tasks have a multilevel or nested structure (items within conditions within participants), generalized linear mixed-effects models should be used. Mixed-effects models correctly handle the inherent dependencies in nested designs and reduce probability of Type I error due to reduced effective sample size (Aarts, Verhage, Veenvliet, Dolan, & van der Sluis, 2014; Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013; Judd, Westfall, & Kenny, 2012).

3. Classical moral dilemmas are typically used, with a single scale that forces utilitarian and deontological inclinations to be inversely related.

4. Self-report measures of individual differences in cognitive/reasoning ability dominate the current literature.

In the current work, we try to address all of these concerns by carrying out studies with (relatively) large sample sizes, using a variety of both self-report and behavioral task measures of reasoning, and using mixed-effects models to generalize the results not only across subjects, but also across items. Consistent with the diversity of measures that we use, our goal is not to investigate which specific reasoning measure is associated with utilitarian inclinations, but rather whether reasoning—defined and assessed broadly—is correlated with such inclinations at the aggregate level across individual measures. Of course, the results for individual measures will be informative for future meta-analyses with a more focused inquiry (e.g., “Are individual differences in CRT predictive of differences in utilitarian judgments?”). But we hesitate to make claims about any specific reasoning measure (e.g., CRT) having greater or lesser predictive utility for individual differences in utilitarian choices given that our goal is to provide evidence for the most general form of relationship between the constructs of domain-general reasoning and moral utilitarianism. Finally, although the battery of measures we use here is not exhaustive, it is drawn from the DPM literature where these measures have been used across a variety of subfields to index various aspects of reasoning (please see Ball & Thompson, 2019; Blanchette, 2013; Trémolière et al., 2018, for discussion of how several of these measures relate to one another and the broader construct of reasoning).

General Methods

Data Availability Statement

All data files and R analysis scripts are available from the Open Science Framework: https://osf.io/jdzfs/.

Data Analysis

Statistical analysis was conducted using R programming language (https://www.r-project.org/). As we note above, because the behavioral data had a multilevel or nested structure (i.e., items within conditions within participants in Study 1), we used generalized linear mixed-effects models with maximal random effects structure.

Our studies were not designed to characterize a detailed pattern of associations between individual reasoning measures and utilitarian moral judgment (e.g., some specific correlations would be stronger than others), but instead to firmly establish the general form of this association. Therefore, we carried out random-effects meta-analyses (Borenstein, Hedges, Higgins, & Rothstein, 2009) using regression estimates (and the associated standard errors) across measures for each study and assessed if the meta-analytic effect was significantly different than 0. We also provide an index of variability ($I^2$) in effect size estimates that is due to statistical heterogeneity and not due to sampling error (Borenstein et al., 2009).

In addition to providing details from null hypothesis significance testing NHST approach, we also compute Bayes factors for the random-effects meta-analysis (Heck, Gronau, & Wagenmakers, 2017). The prior distribution for the mean effect was assumed to be normal, $d \sim N(0, 0.3)$, whereas the prior for the standard deviation of study effects $\tau$ in the random-effects meta-analysis was assumed to follow half-Cauchy distribution (scale = 0.5). This analysis is especially crucial for establishing evidence in favor of the null hypothesis ($BF_{01}$), where the null hypothesis significance testing approach is uninformative (Aczel et al., 2018). Note that, where relevant, we provide natural logarithm values for Bayes Factors, that is, $\log_e(BF_{01})$, which need to be exponentiated to get the $BF_{01}$.

For Study 1, these were standardized estimates from a linear mixed-effects regression model, while for Study 3 these were standardized regression estimates from a simple linear regression model.
Our conclusions are based on these meta-analytic summary effects. Note that under a random-effects model it is assumed that the true effect size varies from study to study and the meta-analytic summary effect is an estimate of the mean for the population of effect sizes (Borenstein et al., 2009). This was indeed the case for the current study where we acknowledge that the precise strength of the relationship between any given measure of reflective reasoning and utilitarianism might vary from measure to measure. As such, we test whether the mean association is significantly different than zero in both frequentist and Bayesian random-effects meta-analysis.

### Data Visualization and Reporting

For brevity, many statistical parameters are included in the figures rather than the main text (an approach adopted in the R package ggstatsplot; Patil, 2018). In addition, for the sake of brevity, demographic details for all studies (age summary statistics and gender breakdown) and details about experimental design for the studies are provided in Table 1. Finally, more exhaustive details about the questionnaires are provided in the online supplemental materials (Supplemental Text S1), while the detailed text of the scenarios is reported in Supplemental Text S6.

### Ethics Statement

Across all studies, participants provided written informed consent before any study procedure was initiated. The studies conducted in Italy (1b and 2) were approved by the Ethics Committee of Scuola Internazionale Superiore di Studi Avanzati (Trieste, Italy) and the Hospital ‘Santa Maria della Misericordia’ (Udine), respectively. Studies 3b–3d were carried out on Amazon Mechanical Turk and were approved by the Ethics Committee of Harvard University under the umbrella protocol (IRB14-2016). Study 1a was carried out on the Moral Sense Test platform, while Study 3a was carried out on Harvard Digital Lab for the Social Sciences platform (Strange, Enos, Hill, & Lakeman, 2019). Both studies were approved by the Ethics Committee of Harvard University.

### Studies 1a and 1b: Reasoning and “Utilitarianism” in Classic Dilemmas

In the classic approach to studying utilitarian moral judgments, participants are asked to judge a harmful action as either acceptable or unacceptable, which is taken to be an endorsement of either the utilitarian or deontological principle, respectively. In this study, we assessed associations between self-reported measures of reasoning and “utilitarian” moral judgment as indexed by a single measure that assumes utilitarian and deontological tendencies are inversely related. As we note above, this approach is problematic (Conway & Gawronski, 2013; Conway et al., 2018), but this is how utilitarianism has typically been assessed and therefore we wanted to see if we could find a positive association between reasoning and utilitarian inclination when it is assessed in this manner.

### Participants

See Table 1.

### Measures

We included the following questionnaires (for more, see Supplemental Text S1 in the online supplemental materials):

- **NFC** (Cacioppo, Petty, & Kao, 1984), which assesses intrinsic motivation to engage in cognitive deliberation;
- **Rational Ability (RA)**, which indexes ability to think logically and analytically, and **Rational Engagement (RE)**, which gauges reliance on and enjoyment of thinking in an analytical, logical manner—these are subscales of the Rational Experiential Inventory (REI; Pacini & Epstein, 1999);
- **AOT** (Baron, 1993; Baron et al., 2015), which assesses individual differences in disposition to be fair toward different conclusions even if they go against one’s initially favored conclusion, to spend enough time on a problem before giving up, and to consider carefully the opinions of others in forming one’s own;
- **Belief Bias (BB)**, syllogisms taken from Morley [née Lambell], Evans & Handley, 2004; Thompson & Evans, 2012; Baron et al., 2015), which measures the tendency to uncritically accept evidence when it favors prior beliefs, and to dismiss or discredit evidence that supports a conclusion that one does not agree with;
- **CRT** (Frederick, 2005), which measures individuals’ ability to suppress an intuitive response in favor of a more reflective and deliberative response.

### Table 1

**Details About Experimental Design and Participant Demographic Information for All Studies**

<table>
<thead>
<tr>
<th>Study</th>
<th>Paradigm</th>
<th>Platform used</th>
<th>Sample size</th>
<th>Average age</th>
<th>Gender (% female)</th>
<th>Measure/manipulation</th>
<th>Number of dilemmas</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>Classical behavioral</td>
<td>MoralSenseTest</td>
<td>1,127</td>
<td>29.50</td>
<td>52%</td>
<td>NFC, AOT, BB, REI</td>
<td>13</td>
<td>Likert (1–7)</td>
</tr>
<tr>
<td>1b</td>
<td>Classical behavioral</td>
<td>Lab-based</td>
<td>109</td>
<td>23.95</td>
<td>61%</td>
<td>CRT</td>
<td>4</td>
<td>Likert (1–7)</td>
</tr>
<tr>
<td>2</td>
<td>Classical neuroimaging</td>
<td>MRI scanner</td>
<td>50</td>
<td>23.06</td>
<td>64%</td>
<td>Cortical thickness</td>
<td>30</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>3a</td>
<td>PD behavioral</td>
<td>DLABSS</td>
<td>1,029</td>
<td>—</td>
<td>54%</td>
<td>NFC, AOT, BB, REI, CRT</td>
<td>20</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>3b</td>
<td>PD behavioral</td>
<td>MTurk</td>
<td>302</td>
<td>35.27</td>
<td>42%</td>
<td>w (Daw—nonmoral)</td>
<td>20</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>3c</td>
<td>PD behavioral</td>
<td>MTurk</td>
<td>454</td>
<td>37.22</td>
<td>57%</td>
<td>w (Kool—nonmoral)</td>
<td>20</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>3d</td>
<td>PD behavioral</td>
<td>MTurk</td>
<td>198</td>
<td>34.39</td>
<td>41%</td>
<td>w (Kool—moral)</td>
<td>20</td>
<td>Dichotomous</td>
</tr>
<tr>
<td>3e</td>
<td>PD behavioral</td>
<td>MTurk</td>
<td>142</td>
<td>35.42</td>
<td>49%</td>
<td>w (Daw—moral)</td>
<td>20</td>
<td>Dichotomous</td>
</tr>
</tbody>
</table>

**Note.** AOT = Actively Open-Minded Thinking; BB = Belief Bias; CRT = Cognitive Reflection Test; DLABSS = The Harvard Digital Lab for the Social Sciences; MTurk = Amazon Mechanical Turk; NFC = Need for Cognition; PD = Process Dissociation; REI = Rational-Experiential Inventory; w = weighting parameter from two-step reinforcement-learning task. Age for Study 3a is not available because it was measured in terms of range on the respective platform.
Note that CRT and BB scores are our indices of reasoning ability, while NFC, AOT, and REI scores are taken to gauge cognitive style of thinking.

Procedure

In Study 1a, participants were randomly assigned to complete one of the following questionnaires: NFC, REI, or AOT plus BB. In Study 1b, participants completed only the CRT. Good internal reliability was observed for all subscales of different questionnaires (see Supplemental Table S1 in the online supplemental materials).

For Study 1a, we recontacted\(^2\) (in 2017) participants who, in the past (2005–2017), had completed the (high-conflict) moral dilemma battery (13 items) as part of Harvard University’s Moral Sense Test platform (Cushman, Young, & Hauser, 2006). For Study 1b, participants came to the lab and completed the six-item CRT (among other questionnaires) and the moral dilemma task (with four items).

Results

The linear mixed-effects regression (see Figure 1; also see Supplemental Text S2 and Supplemental Figures S1–S10) revealed that the estimates were positive for all measures (except for NFC). Despite this, random-effects meta-analysis showed that the summary estimate was not significantly greater than zero. More importantly, the Bayes factor in favor of the null hypothesis (summary estimate not being significantly different from 0) was 2.45 (see Figure 1), which corresponds to inconclusive evidence to support the null (Jarosz & Wiley, 2014). In other words, even though self-reported reasoning scores were on average associated with increased likelihood of viewing the “utilitarian” course of action to be more permissible, the evidence to support this claim was inconclusive. Therefore, we cannot conclude that there is a meta-analytic association between reasoning measures and utilitarian responses, nor can we conclude that there is not, which, as discussed in the introduction, is in line with the mixed findings in the existing literature utilizing classical moral dilemmas.

Study 2: Neural Basis of Individual Differences in Reasoning and Utilitarian Moral Judgments

In this study, we tested whether individual differences in cortical thickness of dIPFC (a neuroanatomical measure associated with deliberative reasoning, complex planning, and goal-directed reasoning; Koechlin, Ody, & Kouneiher, 2003; E. K. Miller & Cohen, 2001; Ramnani & Owen, 2004) correlated with utilitarian moral judgments. Again, our prediction was that the higher cortical thickness at dIPFC would be associated with more utilitarian moral judgments.

Participants

A total of 50 healthy community members (32 female) were recruited to participate in this study and were financially compensated for their time and travel. The average age was 23.06 years (SD = 3.08), with a range of 18 to 35. All participants provided written informed consent. The study was approved by the ethics committee of the Santa Maria della Misericordia Hospital in Udine, Italy (for more details, see Supplemental Text S5 in the online supplemental materials).

Materials

Experimental stimuli were 30 text-based, hypothetical vignettes (20 involved classical moral dilemmas, whereas 10 were control scenarios with no conflict). Participants’ responses to two questions: behavior or choice of action (“Would you [nature of action] in order to [outcome of the proposed action]?”) and judgment (“Is it appropriate for you to [nature of action] in order to [outcome of the proposed action]?”). The affirmative answer always corresponded to commission of sacrificial action and was deemed “utilitarian.” Because we used standard dilemmas rather than a PD method, in this study we intend utilitarian to mean “characteristically utilitarian”—a description of the response, but with some uncertainty about the underlying motivation (Greene, 2014; for more details, see Supplemental Figure S1 and Supplemental Table S1 in Supplemental Text S5 in the online supplemental materials).

Data Acquisition

High-resolution structural images were acquired as 190 T1-weighted transverse images with 3D ultrafast gradient echo sequence on a 3T Philips Achieva scanner equipped with an 8-channel SENSE head coil. The following parameters were used: voxel size = 1 × 1 × 1 mm, slice thickness = 1 mm, TR/TE = 8.2/3.7 ms, matrix size = 240 × 240 mm, field of view = 19 cm, flip angle = 8°, no overcontiguous slices.

Cortical Thickness Analysis

Cortical thickness analysis was carried out using Computational Anatomy Toolbox (CAT12) for SPM (r914; http://www.neuro.uni-jena.de/cat12/; Righart et al., 2017). CAT12 relies on the projection-based thickness method to estimate cortical thickness in a fully automated manner (Dahnke, Yotter, & Gaser, 2013). For more exhaustive methodological details, see Supplemental Text 5 in the online supplemental materials.

Statistical Analyses

Smoothed cortical thickness (CT) maps were entered in a second-level GLM for each hemisphere separately. In all models, we included the following covariates: age, age squared (to account for possible quadratic influences), handedness, and gender. As recommended, total intracranial volume was not added as one of the variables (CAT12 manual, see http://www.neuro.uni-jena.de/cat12/CAT12-Manual.pdf). The dependent variable in every model was vertex-specific cortical thickness values and the independent variable of interest was average utilitarian moral judgment scores. We computed utilitarian moral judgment scores by averaging utilitarian choices on (all 20) moral dilemmas.

---

\(^2\) Note that we assume that reasoning ability/capacity is relatively stable over time. There is some evidence to support this claim (Chesney, Bjelkebring, & Peters, 2015; Enkavi et al., 2019; Raoelison & De Neys, 2019; Stagnaro, Pennycook, & Rand, 2018).
Two contrasts were created for each model that regressed CT on the moral judgment scores, one tracking positive association, the other tracking negative: (a) positive ([0, 1]; greater CT associated with increased utilitarian choices), and (b) negative ([0, −1]; greater CT associated with reduced utilitarian choices). Again, we predicted that a positive association between cortical thickness and average utilitarian moral judgment scores would be found at dlPFC.

Whole-brain analyses were thresholded at $p < .05$, with family wise error (FWE) corrected at the cluster-level. The cluster-forming threshold was set to $p < .0001$, $k = 10$ (cf. Eklund, Nichols, & Knutsson, 2016; for more details, see Supplemental Text S5 in the online supplemental materials).

Results

We observed only one significant effect: a positive association between CT at right middle frontal gyrus (or right dlPFC) and utilitarian moral judgment scores ($x = 30, y = 17, z = 46; k = 133, p < .0001$ (primary threshold), $p$ (FWE-corrected) = 0.037). In other words, as predicted, individuals who were more likely to say that they would cause harm when this was necessary for producing greater general welfare also had thicker right dlPFC.

Discussion

In this study, we tested whether individual differences in cortical thickness were associated with making utilitarian choices in moral dilemmas. Indeed, increased thickness in dorsolateral prefrontal cortex (dlPFC) was associated with greater frequency of utilitarian decisions. This converges with our previous results identifying a positive association between greater deployment of reasoning-related activity, arguably stemming from greater computational efficiency afforded by thicker cortical surface (Kanai & Rees, 2011; Wagstyl & Lerch, 2018), and the propensity to downregulate a competing harm-averse response if this is necessary to obtain a better overall outcome (for more detailed discussion, see Supplemental Text S5 in the online supplemental materials).

Studies 3a–3e: Reasoning and Utilitarian Inclinations

As we note above, the classic approach to assessing how people resolve moral dilemmas conflates selecting the deontological...
choice with rejecting the utilitarian choice (and vice versa). However, the DPMs maintain that the deontological and utilitarian inclinations derive from conceptually independent processes, rendering it possible for them to produce conflicting inclinations in high-conflict moral dilemmas (Koenigs et al., 2007). Conway and Gawronski (2013) have outlined a PD approach that allows independent measurement of individual differences in the strength of deontological and utilitarian tendencies.

**Process Dissociation Method**

The central idea underlying PD is to compare responses on incongruent trials, in which the underlying processes lead to divergent response tendencies, as well as responses on congruent trials, in which the underlying processes lead to the same response tendency. Therefore, participants are administered two types of dilemmas: incongruent and congruent moral dilemmas (10 items of each type). Incongruent dilemmas pit deontological against utilitarian inclinations (traditionally known as high-conflict moral dilemmas): for example, is it acceptable to sacrifice animals for scientific experimentation if this can lead to discovery of a life-saving drug? On the other hand, congruent variants of the same dilemmas are designed to avoid this tension between deontological and utilitarian inclinations, and so both moral concerns converge on the same solution: for example, is it acceptable to sacrifice animals for scientific experimentation if this can lead to discovery of a cosmetic product to maintain smooth skin? Comparing responses on incongruent trials (when the two underlying processes compete) to responses on congruent trials (when the two underlying processes converge), the independent contributions of deontological and utilitarian inclinations, and so both moral concerns converge on the same solution: for example, is it acceptable to sacrifice animals for scientific experimentation if this can lead to discovery of a cosmetic product to maintain smooth skin? Comparing responses on incongruent trials (when the two underlying processes compete) to responses on congruent trials (when the two underlying processes converge), the independent contributions of deontological and utilitarian inclinations to responses on moral dilemmas can be dissociated. It is worth highlighting here that PD parameters aren’t exactly homologous to the two processes described by the DPM, but rather represent response tendencies/inclinations, which result from the combined influence of diverse underlying processes (Conway et al., 2018). For example, utilitarian response tendency can be a result of a deliberative reasoning process (Conway & Gawronski, 2013), processes related to concerns about moral identity (Conway & Gawronski, 2013), and so forth, while deontological response tendency can be a result of an outcome/empathic aversion process (Christov-Moore, Conway, & Iacoboni, 2017; Conway et al., 2018), an action aversion process (Reynolds & Conway, 2018), and so on.

In the current work, we test whether deliberative reasoning is associated with a utilitarian response tendency but not associated with a deontological response tendency. This latter aspect of our investigation is crucial because the same processes can contribute to both response tendencies. For example, prior work shows that processing of religious concerns contribute to both utilitarian and deontological response tendencies (Conway & Gawronski, 2013). Thus, it is important for us to not only establish that reasoning processes contribute to utilitarian concerns about outcome-maximization, but also to show that the same processes do not contribute to deontological concerns about harm avoidance.

Therefore, in Study 3, we assessed utilitarian and deontological inclinations separately using the PD approach and hypothesized that individual differences in cognitive ability and deliberation would be positively associated specifically with utilitarian inclinations, which represent a favorable attitude for harm considering the cost-benefit analysis. We predicted no such relationship between reasoning measures and deontological inclinations, which reflect concerns about avoiding harm to others.

**Participants**

See Table 1.

**Measures**

**Study 3a.** Study 3a assessed all the same measures from Studies 1a and 1b (AOT, REL, NFC, BB, and CRT) along with participants’ utilitarian and deontological inclinations, as quantified by the PD method. Participants were randomly assigned to complete one of the reasoning measures in addition to the dilemmas.

**Study 3b–3e.** Because all the measures (except for CRT and BB) we used thus far have been self-report questionnaires, there is still a possibility that any positive association between analytic thinking and utilitarian inclinations could be aspirational or driven by self-representational concerns. To rule out this possibility, Studies 3b–3e also included behavioral tasks that index effortful thinking, derived from a class of recent reinforcement-learning (RL) tasks that measure the degree to which people use either a more automatic, habitual strategy or a more deliberative, goal-directed strategy in decision making.

Reinforcement learning is a computational approach to modeling value-guided learning and decision making (Sutton & Barto, 1998). Although value-guided decision making is just one part of a person’s cognitive style, it is an especially useful part to interrogate for our purposes. In this framework, the habitual strategy, model-free RL, simply reinforces actions that previously led to good outcomes (Thorndike, 1898). This is a relatively “cheap” strategy, because it needs only to associate experienced rewards with the actions that produced them. On the other hand, the goal-directed strategy, model-based RL, uses sophisticated search algorithms to plan over a representation (or model) of the environment toward goals. This is a computationally demanding strategy, but also a more accurate one, because it can incorporate sudden changes into its causal model, without having to learn from trial-and-error. Because we have well-specified computational models of each approach, we can construct decision problems that reliably elicit unique signatures of either model-free or model-based RL methods. This allows us to infer an important element of a person’s cognitive style in an unobtrusive manner, avoiding aspirational or self-representational biases.

Specifically, model-free and model-based methods can be inferred from participants’ responses to a family of so-called “two-step” tasks (i.e., tasks in which participants earn rewards by sequencing appropriate behaviors in two steps). Here, we focus on two variations of the two-step task (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Kool, Cushman, & Gershman, 2016), both of which are able to distinguish contributions of model-free and model-based systems by introducing situations in which the
model-based system can use its explicit model of the task structure to overcome the limitations of using direct action-reward associations. We included tasks of both types to provide convergent evidence for our hypothesis about a relationship between cognitive style and moral judgment; we did not predict systematic differences between the tasks in this relationship.

The logic of each task and their distinguishing features have been extensively catalogued elsewhere (Kool, Cushman, & Gershman, 2018). In essence, both tasks take advantage of certain situations in which a person takes an action that later leads to reward. Model-free algorithms selectively draw the local inference that the particular action should be repeated (“If A was repeated, choose A again”), whereas model-based algorithms draw global inferences that sometimes favor different actions based on the internal logic of the task (“If A was rewarding, this implies that B must be just as good, or better!”).

Most crucially for our purposes, model-based control in these tasks requires effortful thinking. For example, people show a reduced capacity for model-based control when they are under cognitive load (Otto, Raio, Chiang, Phelps, & Daw, 2013), or when functioning of their frontal cortex is disrupted using transcranial magnetic stimulation (Smittenaar et al., 2013). On the other hand, people increase model-based control when the potential for reward is temporarily amplified (Kool, Gershman, & Cushman, 2017). Moreover, individual differences in the mixture or weighting parameter (w) on the two-step task predict those measures from tasks that require executive functioning, such as the Stroop task (Otto, Skatova, Madlon-Kay, & Daw, 2015) and the CRT (Don, Goldwater, Otto, & Livesey, 2016), and it also predicts steeper temporal discounting (Shenhav, Rand, & Greene, 2017).

We measured participants’ degree of model-based control using both the Daw et al. (2011) and Kool et al. (2016) versions of the two-step task, applying each of these in two different framings. First, following prior research, we embedded (a nonmoral version of) each task in a space travel cover story (Figure 2A and 2C). In these tasks, participants chose between spaceships to travel to planets where they encountered aliens from whom they could earn points in the form of treasure pieces (Decker, Otto, Daw, & Hartley, 2016; Kool et al., 2016). Second, we used versions of both two-step tasks that were embedded in a cover story with a moral dimension (Figure 2B and 2D). In these tasks, participants chose between doors that transitioned to rooms where they pressed buttons that administered painful electric shocks of varying intensities to other people (hypothetically). In this task, each outcome was paired with an image of a human face with a proportional painful expression. For all of these studies, we used the dual-system RL model to measure participants’ relative balance between model-free and model-based control (see Supplemental Text S1 in the online supplemental materials for more detail; scripts for all RL tasks available at: https://github.com/wkool/shockgames).

Across these four studies, we recontacted (after a few months) participants who had completed the four variants of two-step tasks for us in the past. Participants who responded to the follow-up were then asked to complete the moral dilemma task.

**Results**

Collapsing across Studies 3A–3E, we observed the predicted positive correlation between the utilitarian parameter and measures of reasoning. Although this association was statistically significant only for a subset of the measures, the random-effects meta-analysis showed that the summary effect differed significantly from zero (see Figure 3). By contrast, we found no support for an association between reasoning measures and the deontological parameter. We found similar results by repeating the same analysis with robust linear regression (see Supplemental Figure S1 in Supplemental Text S5 in the online supplemental materials).

These results are consistent with the DPM’s proposal that individuals with greater reasoning abilities/tendencies (as assessed here using the two-step tasks, CRT, and BB) and a greater preference for a cognitive style of thinking (indexed using self-report measures NFC, AOT, and REI) are more likely to use cost-benefit thinking to override a contradictory intuitive, emotional response (see Figure 3). Importantly, as would be expected from the DPM, reasoning ability or style of thinking did not predict the magnitude of the putative emotional response (for more detailed visualizations, see Supplemental Figures S1–S9 in Supplemental Text S3 in the online supplemental materials).

**General Discussion**

The DPM of morality maintains that utilitarian responding to sacrificial dilemmas requires reasoning to conduct a cost-benefit analysis and also to override an emotional aversion to harm (Greene et al., 2004). Thus, individuals with a greater ability to reason, or propensity for reasoning, are predicted to make utilitarian moral judgments more often in sacrificial dilemmas. Although some prior work has supported this prediction, other work has not. Overall, this literature has been hampered by both conceptual challenges (on account of not using the PD approach to index utilitarian inclinations) and methodological weaknesses. In the current work, we use a PD approach to properly assay utilitarian attitudes and overcome prior methodological limitations to show that better reasoning skills are indeed on average associated with a preference for utilitarian outcomes in sacrificial moral dilemmas. Across eight studies and a variety of self-report, behavioral, and neuroanatomical measures, we find that reasoning ability and a tendency toward deliberative thinking are positively associated with a preference for utilitarian solutions, independent of individual differences in harm aversion.

In recent years psychologists have paid increasing attention to the importance of conducting rigorous, principled tests of influential theories (e.g., see Lai et al., 2014, for a parallel approach in the domain of IAT interventions). Here, our approach has been to incorporate several different measures of reasoning drawn from diverse subfields of psychological research and employing a variety of methods. We have also used a large set of moral dilemmas, including both traditional sacrificial dilemmas and those developed for PD. Generalizing across these dimensions of variability, we find a consistent relationship between reasoning and utilitarian moral judgments. An important consequence of this finding is that researchers may now seek to better characterize the precise nature of this relationship with increased confidence that, in the most general sense, some important relationship exists. Although this series of studies provides an intriguing data set documenting the relative strength of the correlations between diverse measures of reasoning and utilitarian moral judgments, it was not designed to characterize a detailed pattern of associations (e.g., some specific
Figure 2 (opposite)
correlations would be stronger than others) but instead to firmly establish the relationship’s general form.

Our study does inform a contemporary debate over the cause of “utilitarianism” as a dimension of individual difference. Several studies convincingly argue that individuals with cold-hearted attitudes toward others’ wellbeing, psychopathic traits, and antisocial tendencies tend to favor utilitarian actions on moral dilemmas (Bartels & Pizarro, 2011; Djeriouat & Trémolière, 2014; Duke & Bègue, 2015; Francis et al., 2016; Gavronski et al., 2017; Gleichgerrcht & Young, 2013; Kahane, Everett, Earp, Farias, & Savulescu, 2015; Marshall, Watts, & Lilienfeld, 2018; McDonald, Defever, & Navarrete, 2017; Park, Kappes, Rho, & Van Bavel, 2016; Patil, 2015; Patil, Melsbach, Hennig-Fast, & Silani, 2016; Patil & Silani, 2014; Reynolds & Conway, 2018; Tassy, Deruelle, Mancini, Leistedt, & Wicker, 2013). This has been taken to mean that a “utilitarian” response on the moral dilemma is diagnostic of dark traits (Bartels & Pizarro, 2011) rather than an impartial concern for the wellbeing of the actors involved in the dilemma (Kahane et al., 2015). But the relevant studies have used a single index to represent moral judgments of moral dilemmas, which conflates disregard for deontic prohibitions and endorsement of utilitarian principles. Once utility maximization and harm avoidance concerns are separately assessed, however, individuals who perceive themselves to be more rational (as reported on questionnaires like the NFC, REI, and AOT) and who perform well on cognitively demanding behavioral tasks (e.g., two-step RL tasks, BB, and CRT) prefer utilitarian solutions to moral dilemmas because of their elevated concerns for maximizing utility, and not due to reduced concerns about harm avoidance. Moreover, other studies using the PD approach show that dark personality traits in fact predict reduced deontological inclinations (i.e., a reduced aversion to harm), not increased utilitarian inclinations (Conway et al., 2018; Conway, Reynolds, Celestin, & Pizarro, 2019; Gavronski et al., 2017; Reynolds & Conway, 2018). In sum, utilitarianism in moral dilemmas is not a “mismeasure of morals” (Bartels & Pizarro, 2011)—rather, it partially but meaningfully reflects concerns about achieving the greater good (Greene, 2013; Greene et al., 2001).

In addition to these theoretical contributions, we have also demonstrated the utility of taking a multimodal approach of using self-report, performance, computational, and neural assessments of a construct. For the first time, we have explored the role of model-based control in utilitarian reasoning. Prior formal, computational frameworks (Crockett, 2013; Cushman, 2013) propose two systems: (a) a model-based, outcome-focused algorithm that forms value representations through an internally represented causal model of the world, selects actions based on the value assigned to their likely outcomes, and supports a utilitarian response, and (b) a model-free, action-focused algorithm that assigns value to actions intrinsically based on their reinforcement history and supports deontological harm aversion. These accounts predict that individual differences in model-based control would be correlated with utilitarian judgments on sacrificial dilemmas, a prediction modestly supported by the data (Studies 3b–3e). Importantly, the strength of this association was sensitive to the ecological context (the nature of the task, rewards, etc.) and this can be a starting point for future studies on this topic.

Although there is a large body of evidence to implicate the dIPFC in various aspects of domain-general reasoning (Asplund et al., 2010; Cieslik et al., 2013; MacDonald et al., 2000; Menon & Uddin, 2010; E. K. Miller & Cohen, 2001; Minzenberg et al., 2009; Smittenaar et al., 2013; Van Hoeck et al., 2015; Weissman et al., 2008; Yamagishi et al., 2016) and utilitarian moral reasoning (Cushman, Murray, et al., 2012; Greene et al., 2004; Jeurissen et al., 2014; Tassy et al., 2009; Zheng et al., 2018), few studies have investigated whether individual differences in neurofunctional or neuroanatomical correlates of dIPFC correlate with utilitarian responses. In the current study, we provided evidence for the first time for an association between individual differences in a neural measure (cortical thickness at dIPFC) and “utilitarian” judgments. Given that we are relying on reverse-inference (Poldrack, 2006) to attribute a functional role (i.e., “derivative reasoning”) to dIPFC in the context of utilitarian moral judgment, we refrain from discussing the exact mechanistic pathway for how thicker cortex at dIPFC can lead to greater utilitarian inclinations (for our speculative proposal, see Supplemental Text S5 in the online supplemental materials).

Limitations and Future Scope

A notable limitation of this study is that, despite using several diverse measures of reasoning, we are not well-powered to determine whether some show greater relationships to utilitarian moral

Figure 2 (opposite). The transition structures of the four two-step tasks (Study 3b–3e). We used two different task structures/paradigms (Daw and Kool, respectively) that were both framed in a nonmoral and moral context. The nonmoral versions of the tasks involved ‘space travel,’ where participants choose spaceships to fly to planets with ‘aliens’ that offer an opportunity to win ‘space treasure.’ The moral versions of the tasks involved choosing between doors that lead to rooms with buttons that administered shocks to a hypothetical agent. In the Daw versions of the two-step task (A, B), each trial begins with a first stage choice between two options. Each first-stage choice has a high (70%) probability of transitioning to one of two second-stage states and a low (30%) probability of transitioning to the other. Each second-stage choice was associated with a probability of earning a certain outcome (between 0.25 and 0.75) that slowly drifted according to a random Gaussian walk ($\sigma = 0.025$). In the Kool versions of the two-step task (C, D), each trial begins in one of two first-stage states, each of which offers a choice between two options. Each choice triggers a deterministic transition to one of two second-stage states, which in turn are associated with a scalar outcome (between 0 and 3) that drifted according to a random Gaussian walk ($\sigma = 2$). For each of these tasks, model-free and model-based decision making can be dissociated, because the model-based system can use its representation of the transition structure to plan toward goals, whereas the model-free system relies on slower trial-and-error learning to inform choice (E). Trial sequence in the nonmoral version of the Kool two-step task. On this trial, the participant selects the turquoise spaceship, and therefore transitions to the red planet, where she earns four pieces of space treasure ($f$). Trial sequence in the moral version of the Kool two-step task. On this trial, the participant opens the brown door, and transitions to the purple room, where the button press leads to a maximally painful shock for the hypothetical agent. All individuals depicted in the figures contained in this article granted full permission for their likenesses to appear herein (Mende-Siedlecki, Qu-Lee, Goharzad, & Drain, 2019). See the online article for the color version of this figure.
Future research will benefit from assessing the relationship between different, more specific components of reasoning (problem solving, counterfactual thinking, etc.) and utilitarian preferences. For example, a recent study replicated the correlation between measures of reflection and increased utilitarian choices on dilemma judgments, but also revealed that arithmetic reflection, such as performance on the CRT, predicts only utilitarian tendencies, whereas logical reflection, such as a test measuring ability to solve logical syllogisms, predicts both deontological and utilitarian tendencies (Byrd & Conway, 2019).

Relatedly, the precise mechanistic pathway that connects reasoning to utilitarianism remains unclear. Although we have argued that reasoning supports cost-benefit analysis, it is possible that better reasoners are also better at down-regulating their emotional response to harm. This is an important avenue to explore in future research.

---

**Figure 3.** Standardized regression coefficients from linear regression models with process dissociation (PD) parameters regressed on reasoning measures scores from Study 3a–3c. The estimates were consistently positive (except for one) for the utilitarian parameter, and the meta-analytic summary estimate was significantly different than zero. No such effect was observed for the deontological parameter. In other words, higher self-report scores and behavioral performance on analytic thinking measures was associated with increased concerns for the utility maximization, and not with harm minimization. The caption for each plot also shows results from random-effects Bayesian meta-analysis and heterogeneity tests. The latter shows that there is a substantial heterogeneity ($I^2 = 60\%$) for effect sizes (correlation coefficients) between reasoning measure and deontological parameter, while there is little variation for correlation coefficients between reasoning measure and utilitarian parameter ($I^2 = 31\%$), where the effect sizes are consistently positive in magnitude. Error bars indicate 95% confidence intervals. Error bars indicate 95% confidence intervals. w = mixture or weighting parameter. See the online article for the color version of this figure.
Lastly, the present studies are purely correlational in nature and cannot speak about a causal relationship between these two variables. Future work can rigorously attempt to use either extraneous load manipulations or neurostimulation techniques to assess causal influence of deliberative reasoning in utilitarian inclinations.

Implications

Dilemmas are interesting to study precisely because they are so difficult to resolve. Outside of the laboratory, people routinely face difficult decisions about when harm to few is an allowable path to benefiting many. Is it fair to take a toy from one toddler so it can be shared with the class? Is it right to abort a fetus for the welfare of its mother? How many civilian deaths are permitted in order to win a just war?

Although no one thinks that descriptive facts about human moral psychology are sufficient to answer these difficult questions, many people find them informative (Greene, 2013; Greene & Cohen, 2004; Kohlberg, 1971). To the extent that ordinary people’s utilitarian moral judgments arise from callous indifference to others’ suffering, we might reasonably approach such judgments (and those who advance them) with skepticism by default (Bartels & Pizarro, 2011; Kahane, 2015). Alternatively, to the extent that they arise from deliberation about how to impartially minimize such suffering (consistent with our findings), less skepticism is warranted. Similarly, it has been influential to argue that reasoning plays no substantive role in moral judgment beyond post hoc rationalization (Haidt, 2001). The possibility that reasoning is inert, and reasons post hoc, holds significant implications for how we should approach moral disagreement and personal or societal moral improvement. Our finding that reasoning has an important and systematic role in resolving moral dilemmas is an important counterweight to this view, implying the possibility of moral change by rational thought (Hannikainen, Machery, & Cushman, 2018; Pinker, 2012, 2019).

Conclusion

The influential DPM of moral judgment makes a basic prediction about individual differences: those who reason more should tend to make more utilitarian moral judgments. Nearly 20 years after the theory was proposed, this empirical connection remains disputed. Here, we assemble the largest and most comprehensive empirical survey to date of this putative relationship, and we find strong evidence in its favor.

References


Breukelaar, I. A., Antees, C., Grieve, S. M., Foster, S. L., Gomes, L.,
Williams, L. M., & Korgaonkar, M. S. (2017). Cognitive control net-
work anatomy correlates with neurocognitive behavior: A longitudinal
hbm.23401
Cognitive and neural foundations of social norms and their enforcement.
Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arith-
metic reflection predicts utilitarian tendencies, but logical reflection
predicts both deontological and utilitarian tendencies. Cognition, 192,
103995. http://dx.doi.org/10.1016/j.cognition.2019.06.007
10.1037/0022-3514.42.1.116
http://dx.doi.org/10.1207/s15327752ja4803_13
Capraro, V., Everett, J. A. C., & Earp, B. D. (2019). Priming intuition
decorrelation of instrumental harm but not impartial beneficence.
Frontiers in Experimental Social Psychology, 83, 142–149. http://dx.doi.org/
10.1016/j.esper.2019.04.006
work anatomy correlates with neurocognitive behavior: A longitudinal
work anatomy correlates with neurocognitive behavior: A longitudinal
Cushman, F. (2013). Action, outcome, and value: A dual-system frame-
work for morality. Personality and Social Psychology Review, 17, 273–
292. http://dx.doi.org/10.1177/1088868313495594
Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating
murder: The aversion to harmful action. Emotion, 12, 2–7. [http://dx.doi.org/10.1037/a0025071](http://dx.doi.org/10.1037/a0025071)
Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene,
J. D. (2012). Judgment before principle: Engagement of the frontopari-
tal control network in condemning harms of omission. Social Cognitive
and Affective Neuroscience, 7, 888–895. http://dx.doi.org/10.1093/scan/nst072
reasoning and intuition in moral judgment: Testing three principles of
j.1467-9220.2006.01834.x
10.1016/j.neuroimage.2012.09.050](http://dx.doi.org/10.1016/j.neuroimage.2012.09.050)
Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J.
(2011). Model-based influences on humans’ choices and striatal predic-
2011.02.027
creatures of habit to goal-directed learners: Tracking the developmental
emergence of model-based reinforcement learning. Psychological Sci-
cence, 27, 848–858. http://dx.doi.org/10.1177/0956797616639301
De Neys, W., & Bialek, M. (2017). Dual processes and conflict during
moral and logical reasoning: A case for utilitarian intuitions? In B.
Trémolière & J. F. Bonnefon (Eds.), Moral inferences (pp. 123–136).
Hove, UK: Psychology Press.
Djamin, R., & Trémolière, B. (2014). The dark triad of personality and
utilitarian moral judgment: The mediating role of honesty/humility and
80, 312–325. http://dx.doi.org/10.1016/j.neuron.2013.09.007
abstraction, model-based choice, and cognitive reflection. Psychonomic
016-0102-y
concentration predicts utilitarian responses in moral dilemmas. Cogni-
tion, 134, 121–127. [http://dx.doi.org/10.1016/j.cognition.2014.09.006](http://dx.doi.org/10.1016/j.cognition.2014.09.006)
fMRI inferences for spatial extent have inflated false-positive rates.
Proceedings of the National Academy of Sciences of the United States of
America, 113, 7900–7905. http://dx.doi.org/10.1073/pnas.1602413113
Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon,
test-retest reliabilities of self-regulation measures. Proceedings of the
National Academy of Sciences of the United States of America, 116,
5472–5477. http://dx.doi.org/10.1073/pnas.1818430116
Epstein, S. (1994). Integration of the cognitive and the psychodynamic
0003-066X.49.8.709](http://dx.doi.org/10.1037/0003-066X.49.8.709)
Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judg-
[http://dx.doi.org/10.1146/annurev.psych.59.103006.093629](http://dx.doi.org/10.1146/annurev.psych.59.103006.093629)
Evans, J. S. B. T. (2017). Thinking and reasoning: A very short introduc-
1093/actrade/9780198787259.001.0001](http://dx.doi.org/10.1093/actrade/9780198787259.001.0001)


Patil, I., Melsbach, J., Hennig-Fast, K., & Silani, G. (2016). Divergent roles of autistic and alexithymic traits in utilitarian moral judgments in adults