Journal of Personality and Social Psychology

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

Online First Publication, January 9, 2020. http://dx.doi.org/10.1037/pspp0000281

American Psychological Association

Journal of Personality and Social Psychology:
Personality Processes and Individual Differences

ISSN: 0022-3514

http://dx.doi.org/10.1037/pspp0000281

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

Indrajeet Patil, Department of Psychology, Harvard University; Micaela Maria Zucchelli, Department of Psychology, University of Bologna; Wouter Kool, Department of Psychology, Harvard University, and Department of Psychological and Brain Sciences, Washington University in St. Louis; Stephanie Campbell, Department of Psychology, Harvard University; Federico Fornasier and Marta Calò, Department of Life Sciences, University of Trieste; Giorgia Silani, Department of Applied Psychology: Health, Development, Enhancement and Intervention, University of Vienna; Mina Cikara and Fiery Cushman, Department of Psychology, Harvard University.

Indrajeet Patil and Micaela Maria Zucchelli contributed equally to this work.

We would like to acknowledge funding from The Pershing Square Venture Fund for Research on the Foundations of Human Behavior through the Harvard Digital Lab for the Social Sciences. We would additionally like to thank Grant 61061 from the John Templeton Foundation and Dean's Competitive Fund for Promising Scholarship at Harvard for supporting this work.

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

Correspondence concerning this article should be addressed to Indrajeet Patil, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138 or to Micaela Maria Zucchelli, Department of Psychology, University of Bologna, Via Zamboni, 33 - 40126 Bologna, Italy. E-mail: patilindrajeet.science@gmail.com or micaela.zucchelli3@unibo.it

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 *either* because 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/reflective 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 Neys, & 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 Neys & Białek, 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, Neys, & 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 Neys & Białek, 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 & Gürçay, 2017; Gawronski et al., 2017; Greene et al., 2008; Gürçay & 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, & Bahçekapili, 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, & Plewnia, 2019), goal-directed planning (Asplund, Todd, Snyder, & Marois, 2010; Botvinick & An, 2009; Kaller, Rahm, Spreer, Weiller, & Unterrainer, 2011), cost-benefit analysis (Basten, Biele, Heekeren, & Fiebach, 2010), problem solving (Ruh, Rahm, Unterrainer, Weiller, & Kaller, 2012), counterfactual reasoning (Van Hoeck, 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 (Buckholtz & 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-Kouhsari, Woodward, & Rangel, 2015), whereas neurostimulation studies have provided mixed evidence about dlPFC's causal role in utilitarian moral judgment (Jeurissen, Sack, Roebroeck, Russ, & Pascual-Leone, 2014; Kuehne, Heimrath, Heinze, & Zaehle, 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; Nostro, 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[1] 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 τ 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}.

[1] 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

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

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 in the online supplemental materials; Supplemental Text S2 and Supplemental Figures S1 to 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 than 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 dlPFC (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 dlPFC 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, Bjalkebring, & Peters, 2015; Enkavi et al., 2019; Raoelison & De Neys, 2019; Stagnaro, Pennycook, & Rand, 2018).

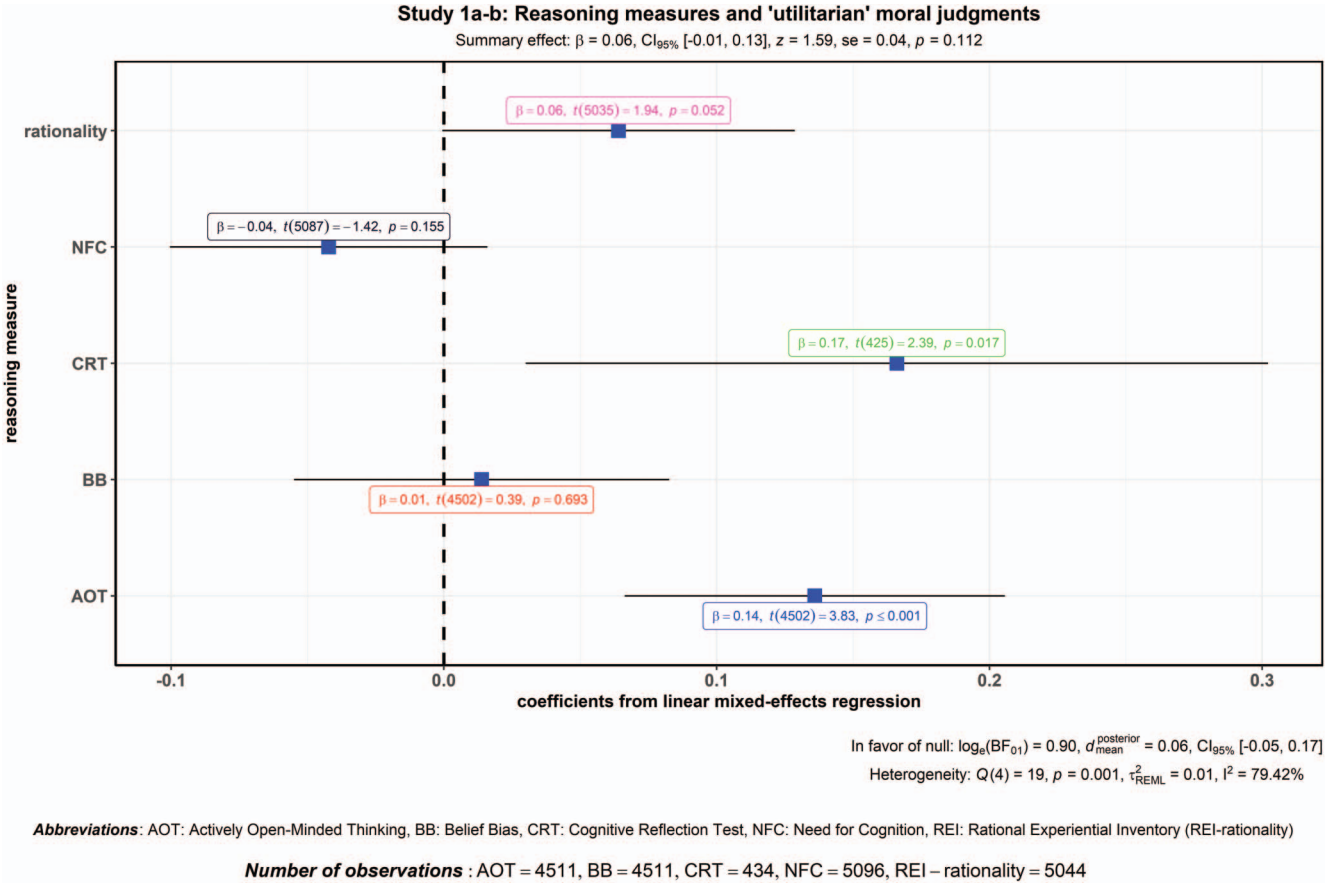

Figure 1. Regression coefficients for analytic thinking measures from linear mixed-effects regression analyses carried out separately for each reasoning measure (Study 1a–b). Although all—except NFC—regression coefficients were positive, the meta-analytic estimate of regression coefficient estimates did not significantly differ from 0. Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

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 to responses on moral dilemmas can be dissociated.[3] 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, REI, 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

[3] The exact algorithmic details of how these parameters are quantified is beyond the scope of this article and are discussed at length elsewhere (Conway & Gawronski, 2013; Conway et al., 2018; Gawronski et al., 2017; Reynolds & Conway, 2018).

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

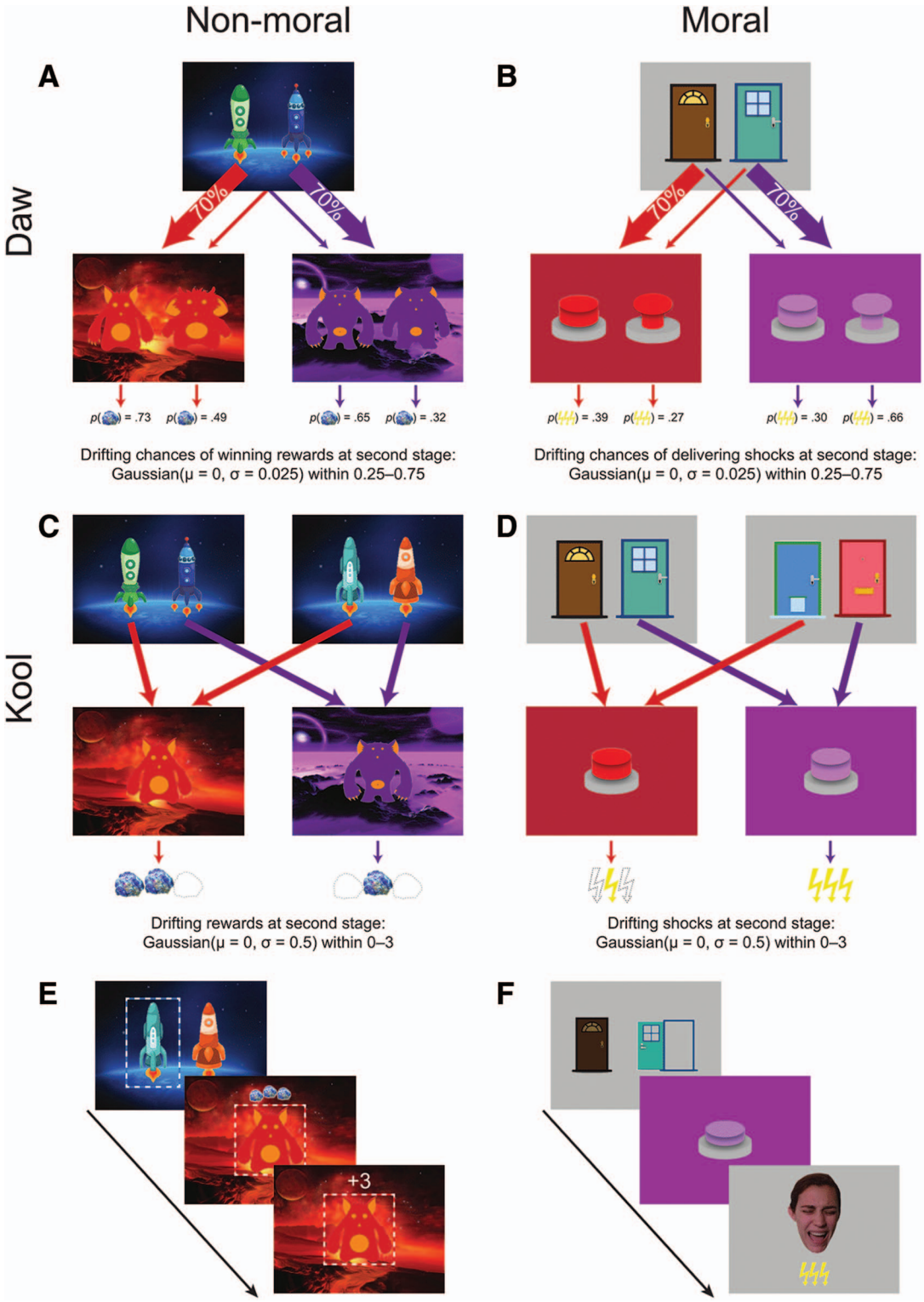

Figure 2 (opposite)

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; Gawronski 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 assayed, 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; Gawronski 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 dlPFC 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 dlPFC 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 dlPFC) and "utilitarian" judgments. Given that we are relying on reverse-inference (Poldrack, 2006) to attribute a functional role (i.e., "deliberative reasoning") to dlPFC in the context of utilitarian moral judgment, we refrain from discussing the exact mechanistic pathway for how thicker cortex at dlPFC 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.

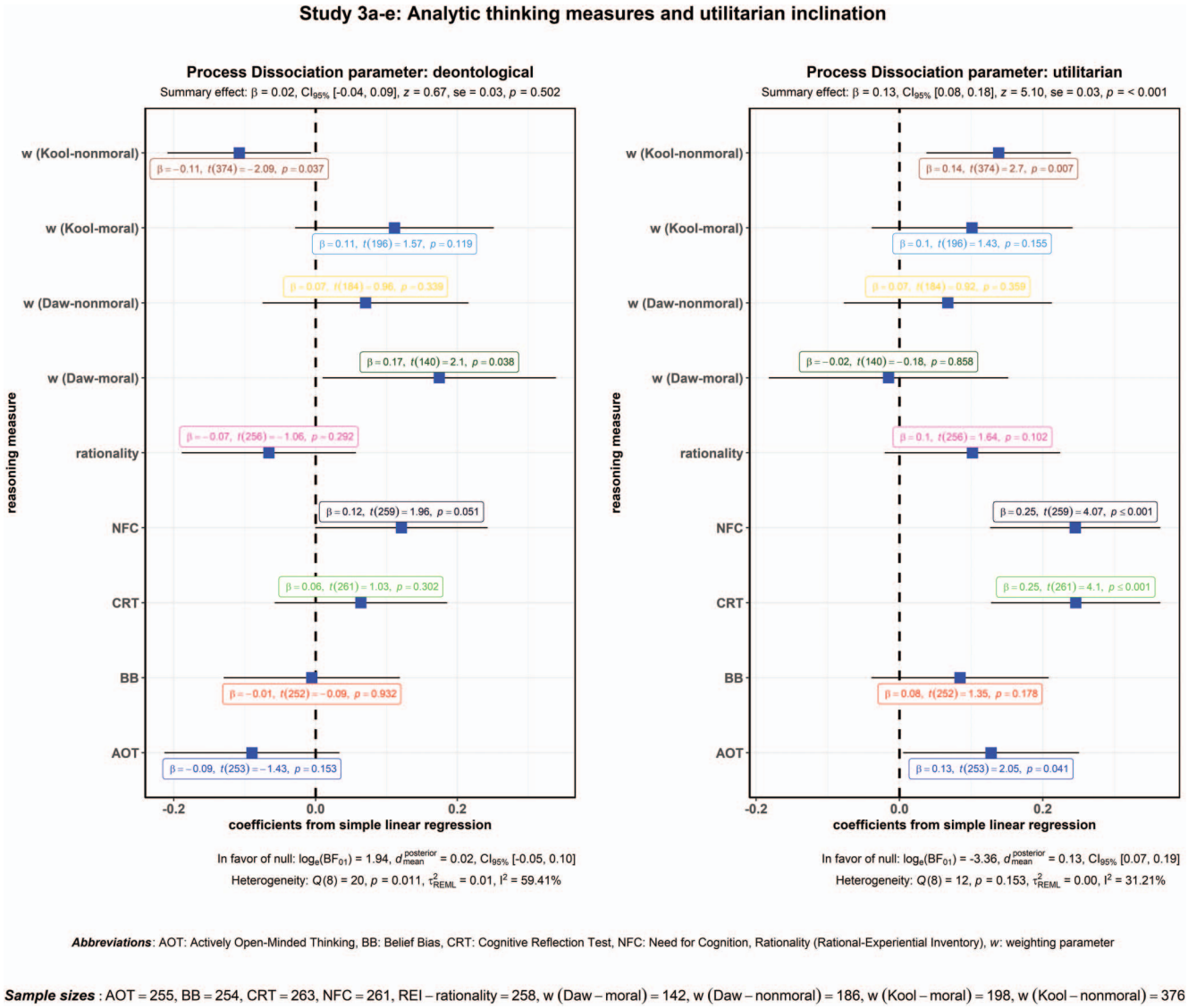

Figure 3. Standardized regression coefficients from linear regression models with process dissociation (PD) parameters regressed on reasoning measures scores from Study 3a–3e. The estimates were consistently positive (except for one) for the utilitarian parameter, and the meta-analytic summary estimate 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[4] heterogeneity ($I^2 \sim 60\%$) 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 ($I^2 \sim 31\%$), where the effect sizes are consistently positive in magnitude. Error bars indicate 95% confidence intervals. w = mixture or weighting parameter. See the online article for the color version of this figure.

judgment than others. 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.

[4] See http://handbook-5-1.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm.

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 benefitting 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 influentially argued 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

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience, 17,* 491–496. http://dx.doi.org/10.1038/nn.3648

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., . . . Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science, 1,* 357–366. http://dx.doi.org/10.1177/2515245918773742

Aktas, B., Yilmaz, O., & Bahçekapili, H. G. (2017). Moral pluralism on the trolley tracks: Different normative principles are used for different reasons in justifying moral judgments. *Judgment and Decision Making, 12,* 297–307. Retrieved from http://journal.sjdm.org/17/17207/jdm17207.pdf

Asplund, C. L., Todd, J. J., Snyder, A. P., & Marois, R. (2010). A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature Neuroscience, 13,* 507–512. http://dx.doi.org/10.1038/nn.2509

Attie, M., & Knobe, J. (2019). Replication of Study 1 by Paxton, J. M., Ungar, L., and Greene, J. D. *Cognitive Science*. Advance online publication. https://osf.io/mjvsd/

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General, 148,* 1782–1801.

Ball, L. J., & Thompson, V. B. (Eds.). (2019). *International handbook of thinking and reasoning* (1st ed.). New York, NY: Psychology Press.

Barbey, A. K., Colom, R., & Grafman, J. (2013). Architecture of cognitive flexibility revealed by lesion mapping. *NeuroImage, 82,* 547–554. http://dx.doi.org/10.1016/j.neuroimage.2013.05.087

Baron, J. (1993). Why teach thinking? An essay. *Applied Psychology: An International Review, 42,* 191–214. http://dx.doi.org/10.1111/j.1464-0597.1993.tb00731.x

Baron, J., & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition, 45,* 566–575. http://dx.doi.org/10.3758/s13421-016-0686-8

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory & Cognition, 4,* 265–284. http://dx.doi.org/10.1016/j.jarmac.2014.09.003

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108,* 381–417. http://dx.doi.org/10.1016/j.cognition.2008.03.001

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

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 21767–21772. http://dx.doi.org/10.1073/pnas.0908104107

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science, 5,* 187–202. http://dx.doi.org/10.1177/1745691610362354

Białek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology, 28,* 631–639. http://dx.doi.org/10.1080/20445911.2016.1156118

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making, 12,* 148–167. Retrieved from http://journal.sjdm.org/17/17224/jdm17224.pdf

Blanchette, I. (Ed.), (2013). *Emotion and reasoning* (1st ed.). New York, NY: Psychology Press. http://dx.doi.org/10.4324/9781315888538

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley, Ltd. http://dx.doi.org/10.1002/9780470743386

Botvinick, M., & An, J. (2009). Goal-directed decision making in prefrontal cortex: A computational framework. *Advances in Neural Information*

Processing Systems, 21, 169–176. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25258502

Breukelaar, I. A., Antees, C., Grieve, S. M., Foster, S. L., Gomes, L., Williams, L. M., & Korgaonkar, M. S. (2017). Cognitive control network anatomy correlates with neurocognitive behavior: A longitudinal study. *Human Brain Mapping, 38,* 631–643. http://dx.doi.org/10.1002/hbm.23401

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience, 15,* 655–661. http://dx.doi.org/10.1038/nn.3087

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition, 192,* 103995. http://dx.doi.org/10.1016/j.cognition.2019.06.007

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116–131. http://dx.doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307. http://dx.doi.org/10.1207/s15327752jpa4803_13

Capraro, V., Everett, J. A. C., & Earp, B. D. (2019). Priming intuition disfavors instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology, 83,* 142–149. http://dx.doi.org/10.1016/j.jesp.2019.04.006

Chesney, D., Bjalkebring, P., & Peters, E. (2015). How to estimate how well people estimate: Evaluating measures of individual differences in the approximate number system. *Attention, Perception, & Psychophysics, 77,* 2781–2802. http://dx.doi.org/10.3758/s13414-015-0974-6

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews, 36,* 1249–1264. http://dx.doi.org/10.1016/j.neubiorev.2012.02.008

Christov-Moore, L., Conway, P., & Iacoboni, M. (2017). Deontological dilemma response tendencies and sensorimotor representations of harm to others. *Frontiers in Integrative Neuroscience, 11,* 34. http://dx.doi.org/10.3389/fnint.2017.00034

Cieslik, E. C., Zilles, K., Caspers, S., Roski, C., Kellermann, T. S., Jakobs, O., . . . Eickhoff, S. B. (2013). Is there "one" DLPFC in cognitive action control? Evidence for heterogeneity from co-activation-based parcellation. *Cerebral Cortex, 23,* 2677–2689. http://dx.doi.org/10.1093/cercor/bhs256

Cohen, J. D. (2005). The vulcanization of the human brain: A neural perspective on interactions between cognition and emotion. *The Journal of Economic Perspectives, 19,* 3–24. http://dx.doi.org/10.1257/089533005775196750

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., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition, 179,* 241–265. http://dx.doi.org/10.1016/j.cognition.2018.04.018

Conway, P., Reynolds, C., Celestin, F., & Pizarro, D. (2019). *Antisociality does not predict utilitarian responses: Clarifying the relationships between psychopathy, prosociality, and dilemma judgments via process dissociation*. Manuscript under review.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17,* 363–366. http://dx.doi.org/10.1016/j.tics.2013.06.005

Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology, 3,* 328. http://dx.doi.org/10.3389/fpsyg.2012.00328

Cushman, F. (2013). Action, outcome, and value: A dual-system framework 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

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2012). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience, 7,* 888–895. http://dx.doi.org/10.1093/scan/nsr072

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,* 1082–1089. http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x

Dahnke, R., Yotter, R. A., & Gaser, C. (2013). Cortical thickness and central surface estimation. *NeuroImage, 65,* 336–348. 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 prediction errors. *Neuron, 69,* 1204–1215. http://dx.doi.org/10.1016/j.neuron.2011.02.027

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science, 27,* 848–858. http://dx.doi.org/10.1177/0956797616639301

De Neys, W., & Białek, 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.

Djeriouat, 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. http://dx.doi.org/10.1016/j.paid.2013.12.026

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80,* 312–325. http://dx.doi.org/10.1016/j.neuron.2013.09.007

Don, H. J., Goldwater, M. B., Otto, A. R., & Livesey, E. J. (2016). Rule abstraction, model-based choice, and cognitive reflection. *Psychonomic Bulletin & Review, 23,* 1615–1623. http://dx.doi.org/10.3758/s13423-016-1012-y

Duke, A. A., & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition, 134,* 121–127. http://dx.doi.org/10.1016/j.cognition.2014.09.006

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why 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, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of 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 unconscious. *American Psychologist, 49,* 709–724. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8092614. http://dx.doi.org/10.1037/0003-066X.49.8.709

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278. http://dx.doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2017). *Thinking and reasoning: A very short introduction*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/actrade/9780198787259.001.0001

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8,* 223–241. http://dx.doi.org/10.1177/1745691612460685

Fermin, A. S. R., Yoshida, T., Yoshimoto, J., Ito, M., Tanaka, S. C., & Doya, K. (2016). Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Scientific Reports, 6,* 31378. http://dx.doi.org/10.1038/srep31378

Francis, K. B., Howard, C., Howard, I. S., Gummerum, M., Ganis, G., Anderson, G., & Terbeck, S. (2016). Virtual morality: Transitioning from moral judgment to moral action? *PLoS ONE, 11*(10), e0164374. http://dx.doi.org/10.1371/journal.pone.0164374

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25–42. http://dx.doi.org/10.1257/089533005775196732

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology, 113,* 343–376. http://dx.doi.org/10.1037/pspa0000086

Gawronski, B., & Beer, J. S. (2016). What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience, 12,* 626–632. http://dx.doi.org/10.1080/17470919.2016.1248787

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLoS ONE, 8*(4), e60418. http://dx.doi.org/10.1371/journal.pone.0060418

Glenn, A. L., Raine, A., Schug, R., Young, L., & Hauser, M. (2009). Increased DLPFC activity during moral decision-making in psychopathy. *Molecular Psychiatry, 14,* 909–911. http://dx.doi.org/10.1038/mp.2009.76

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, NY: Penguin Press.

Greene, J. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics, 124,* 695–726. http://dx.doi.org/10.1086/675875

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 359,* 1775–1785. http://dx.doi.org/10.1098/rstb.2004.1546

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. http://dx.doi.org/10.1016/j.neuron.2004.09.027

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. http://dx.doi.org/10.1126/science.1062872

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning, 23,* 49–80. http://dx.doi.org/10.1080/13546783.2016.1216011

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–834. http://dx.doi.org/10.1037/0033-295X.108.4.814

Hannikainen, I. R., Machery, E., & Cushman, F. A. (2018). Is utilitarian sacrifice becoming more morally permissible? *Cognition, 170,* 95–101. http://dx.doi.org/10.1016/j.cognition.2017.09.013

Heck, D. W., Gronau, Q. F., & Wagenmakers, E.-J. (2017). metaBMA: Bayesian model averaging for random and fixed effects meta-analysis (Version 0.3.9). *CRAN*. Retrieved from https://cran.r-project.org/package=metaBMA

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *The Journal of Neuroscience, 35,* 12593–12605. http://dx.doi.org/10.1523/JNEUROSCI.3402-14.2015

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving, 7*(1). http://dx.doi.org/10.7771/1932-6246.1167

Jeurissen, D., Sack, A. T., Roebroeck, A., Russ, B. E., & Pascual-Leone, A. (2014). TMS affects moral judgment, showing the role of DLPFC and TPJ in cognitive and emotional processing. *Frontiers in Neuroscience, 8,* 18. http://dx.doi.org/10.3389/fnins.2014.00018

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69. http://dx.doi.org/10.1037/a0028347

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience, 10,* 551–560. http://dx.doi.org/10.1080/17470919.2015.1023400

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125,* 131–164. http://dx.doi.org/10.1037/rev0000093

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134,* 193–209. http://dx.doi.org/10.1016/j.cognition.2014.10.005

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment.ord display. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 167–293). NewYork, NY: Cambridge University Press. Retrieved from http://psycnet.apa.org/record/2005-09680-012

Kaller, C. P., Rahm, B., Spreer, J., Weiller, C., & Unterrainer, J. M. (2011). Dissociable contributions of left and right dorsolateral prefrontal cortex in planning. *Cerebral Cortex, 21,* 307–317. http://dx.doi.org/10.1093/cercor/bhq096

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience, 12,* 231–242. http://dx.doi.org/10.1038/nrn3000

Kant, I. (2005). *The moral law: Groundwork of the metaphysic of morals* (2nd ed.). London, UK: Routledge.

Killgore, W. D. S., Killgore, D. B., Day, L. M., Li, C., Kamimori, G. H., & Balkin, T. J. (2007). The effects of 53 hours of sleep deprivation on moral judgment. *Sleep: Journal of Sleep and Sleep Disorders Research, 30,* 345–352. http://www.ncbi.nlm.nih.gov/pubmed/17425231. http://dx.doi.org/10.1093/sleep/30.3.345

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science, 302,* 1181–1185. http://dx.doi.org/10.1126/science.1088545

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446,* 908–911. http://dx.doi.org/10.1038/nature05631

Kohlberg, L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel (Ed.), *Cognitive development and epistemology* (pp. 151–235). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-498640-4.50011-1

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology, 12*(8), e1005090. http://dx.doi.org/10.1371/journal.pcbi.1005090

Kool, W., Cushman, F., & Gershman, S. J. (2018). Competition and

cooperation between multiple reinforcement learning systems. In R. W. Morris, A. Bornstein, & A. M. Shenhav (Eds.), *Goal-directed decision making: computations and neural circuits* (pp. 153–178). London, UK: Academic Press. http://dx.doi.org/10.1016/B978-0-12-812098-9.00007-3

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science, 28,* 1321–1333.

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: A parametric study of relational complexity. *Cerebral Cortex, 12,* 477–485. http://dx.doi.org/10.1093/cercor/12.5.477

Kuehne, M., Heimrath, K., Heinze, H.-J., & Zaehle, T. (2015). Transcranial direct current stimulation of the left dorsolateral prefrontal cortex shifts preference of moral judgments. *PLoS ONE, 10*(5), e0127061. http://dx.doi.org/10.1371/journal.pone.0127061

Kvaran, T., Nichols, S., & Sanfey, A. (2013). The effect of analytic and experiential modes of thought on moral judgment. *Progress in Brain Research, 202,* 187–196. http://dx.doi.org/10.1016/B978-0-444-62604-2.00011-3

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143,* 1765–1785. http://dx.doi.org/10.1037/a0036260

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

Lerch, J. P., van der Kouwe, A. J. W., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K. L., . . . Sotiropoulos, S. N. (2017). Studying neuroanatomy using MRI. *Nature Neuroscience, 20,* 314–326. http://dx.doi.org/10.1038/nn.4501

Li, Z., Xia, S., Wu, X., & Chen, Z. (2018). Analytical thinking style leads to more utilitarian moral judgments: An exploration with a process-dissociation approach. *Personality and Individual Differences, 131,* 180–184. http://dx.doi.org/10.1016/j.paid.2018.04.046

Marshall, J., Watts, A. L., & Lilienfeld, S. O. (2018). Do psychopathic individuals possess a misaligned moral compass? A meta-analytic examination of psychopathy's relations with moral judgment. *Personality Disorders: Theory, Research, and Treatment, 9,* 40–50. http://dx.doi.org/10.1037/per0000226

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288,* 1835–1838.

McDonald, M. M., Defever, A. M., & Navarrete, C. D. (2017). Killing for the greater good: Action aversion and the emotional inhibition of harm in moral dilemmas. *Evolution and Human Behavior, 38,* 770–778. http://dx.doi.org/10.1016/j.evolhumbehav.2017.06.001

McNair, S., Okan, Y., Hadjichristidis, C., & de Bruin, W. B. (2018). Age differences in moral judgment: Older adults are more deontological than younger adults. *Journal of Behavioral Decision Making, 32,* 47–60. http://dx.doi.org/10.1002/bdm.2086

McPhetres, J., Conway, P., Hughes, J. S., & Zuckerman, M. (2018). Reflecting on God's will: Reflective processing contributes to religious peoples' deontological dilemma responses. *Journal of Experimental Social Psychology, 79,* 301–314. http://dx.doi.org/10.1016/j.jesp.2018.08.013

Mende-Siedlecki, P., Qu-Lee, J., Goharzad, A., & Drain, A. (2019). The Delaware pain database: A set of painful expressions and corresponding norming data. *PsyArXiv*. http://dx.doi.org/10.31234/osf.io/kjez5

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure & Function, 214,* 655–667. http://dx.doi.org/10.1007/s00429-010-0262-0

Metuki, N., Sela, T., & Lavidor, M. (2012). Enhancing cognitive control components of insight problems solving by anodal tDCS of the left dorsolateral prefrontal cortex. *Brain Stimulation, 5,* 110–115. http://dx.doi.org/10.1016/j.brs.2012.03.002

Mill, J. S., & Crisp, R. (Eds.) (1998). *Utilitarianism*. New York, NY: Oxford University Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202. http://dx.doi.org/10.1146/annurev.neuro.24.1.167

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass, 7,* 707–718. http://dx.doi.org/10.1111/spc3.12066

Minzenberg, M. J., Laird, A. R., Thelen, S., Carter, C. S., & Glahn, D. C. (2009). Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Archives of General Psychiatry, 66,* 811–822. http://dx.doi.org/10.1001/archgenpsychiatry.2009.91

Morley, N. J., Evans, J. S. B. T., & Handley, S. J. (2004). Belief bias and figural bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A, 57,* 666–692. http://dx.doi.org/10.1080/02724980343000440

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

Nostro, A. D., Müller, V. I., Reid, A. T., & Eickhoff, S. B. (2017). Correlations between personality and brain structure: A crucial role of gender. *Cerebral Cortex, 27,* 3698–3712. http://dx.doi.org/10.1093/cercor/bhw191

Nowrangi, M. A., Lyketsos, C., Rao, V., & Munro, C. A. (2014). Systematic review of neuroimaging correlates of executive functioning: Converging evidence from different clinical populations. *The Journal of Neuropsychiatry and Clinical Neurosciences, 26,* 114–125. http://dx.doi.org/10.1176/appi.neuropsych.12070176

Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. *Neuropsychologia, 91,* 499–508. http://dx.doi.org/10.1016/j.neuropsychologia.2016.09.010

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 20941–20946. http://dx.doi.org/10.1073/pnas.1312011110

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience, 27,* 319–333. http://dx.doi.org/10.1162/jocn_a_00709

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76,* 972–987. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10402681. http://dx.doi.org/10.1037/0022-3514.76.6.972

Park, G., Kappes, A., Rho, Y., & Van Bavel, J. J. (2016). At the heart of morality lies neuro-visceral integration: Lower cardiac vagal tone predicts utilitarian moral judgment. *Social Cognitive and Affective Neuroscience, 11,* 1588–1596. http://dx.doi.org/10.1093/scan/nsw077

Patil, I. (2015). Trait psychopathy and utilitarian moral judgement: The mediating role of action aversion. *Journal of Cognitive Psychology, 27,* 349–366. http://dx.doi.org/10.1080/20445911.2015.1004334

Patil, I. (2018). ggstatsplot: "ggplot2" Based Plots with Statistical Details. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html

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

with autism. *Scientific Reports, 6,* 23637. http://dx.doi.org/10.1038/srep23637

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5,* 501. http://dx.doi.org/10.3389/fpsyg.2014.00501

Patil, I., & Trémolière, B. (2018). *I think, therefore I forgive: The role of reasoning in forgiving accidental harms*. Retrieved from http://dx.doi.org/10.17605/OSF.IO/8GJTH

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are "counter-intuitive" deontological judgments really counter-intuitive? An empirical reply to Kaha ne et al. (2012). *Social Cognitive and Affective Neuroscience, 9,* 1368–1371. http://dx.doi.org/10.1093/scan/nst102

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36,* 163–177. http://dx.doi.org/10.1111/j.1551-6709.2011.01210.x

Pinker, S. (2012). *The better angels of our nature: Why violence has declined*. New York, NY: Penguin Books.

Pinker, S. (2019). *Enlightenment now: The case for reason, science, humanism, and progress*. New York, NY: Penguin Books.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10,* 59–63. http://dx.doi.org/10.1016/j.tics.2005.12.004

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience, 5,* 184–194. http://dx.doi.org/10.1038/nrn1343

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making, 14,* 170–178. Retrieved from http://journal.sjdm.org/18/181112/jdm181112.html

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, 18,* 1009–1023. http://dx.doi.org/10.1037/emo0000413

Righart, R., Schmidt, P., Dahnke, R., Biberacher, V., Beer, A., Buck, D., . . . Mühlau, M. (2017). Volume versus surface-based cortical thickness measurements: A comparative study with healthy controls and multiple sclerosis patients. *PLoS ONE, 12*(7), e0179590. http://dx.doi.org/10.1371/journal.pone.0179590

Rosas, A., & Aguilar-Pardo, D. (2019). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning*. Advance online publication. http://dx.doi.org/10.1080/13546783.2019.1679665

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. http://dx.doi.org/10.1111/cogs.12136

Ruh, N., Rahm, B., Unterrainer, J. M., Weiller, C., & Kaller, C. P. (2012). Dissociable stages of problem solving (II): First evidence for process-contingent temporal order of activation in dorsolateral prefrontal cortex. *Brain and Cognition, 80,* 170–176. http://dx.doi.org/10.1016/j.bandc.2012.02.012

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement, 55,* 818–831. http://dx.doi.org/10.1177/0013164495055005017

Seidlitz, J., Váša, F., Shinn, M., Romero-Garcia, R., Whitaker, K. J., Vértes, P. E., . . . the NSPN Consortium. (2018). Morphometric similarity networks detect microscale cortical organization and predict inter-individual cognitive variation. *Neuron, 97,* 231–247. http://dx.doi.org/10.1016/j.neuron.2017.11.039

Shenhav, A., Rand, D. G., & Greene, J. D. (2017). The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. *Judgment and Decision Making, 12,* 1–18. Retrieved from http://journal.sjdm.org/16/16710/jdm16710.pdf

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22. http://dx.doi.org/10.1037/0033-2909.119.1.3

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80,* 914–919. http://dx.doi.org/10.1016/j.neuron.2013.08.009

Spears, D., Fernández-Linsenbarth, I., Okan, Y., Ruz, M., & González, F. (2018). Disfluent fonts lead to more utilitarian decisions in moral dilemmas. *Psicológica, 39,* 41–63. http://dx.doi.org/10.2478/psicolj-2018-0003

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making, 13,* 260–267. Retrieved from http://journal.sjdm.org/18/18201/jdm18201.pdf

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89,* 342–357.

Strange, A., Enos, R. D., Hill, M., & Lakeman, A. (2019). *Intrinsic motivation at scale: Online volunteer laboratories for social science research*. Working Paper. Retrieved from https://scholar.harvard.edu/files/renos/files/strangeenoshilllakeman.pdf

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119,* 454–458. http://dx.doi.org/10.1016/j.cognition.2011.01.018

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (1st ed.). Cambridge, MA: MIT Press.

Takeuchi, H., Taki, Y., Nouchi, R., Yokoyama, R., Kotozaki, Y., Nakagawa, S., . . . Kawashima, R. (2017). Global associations between regional gray matter volume and diverse complex cognitive functions: Evidence from a large sample study. *Scientific Reports, 7,* 10014. http://dx.doi.org/10.1038/s41598-017-10104-8

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., & Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Frontiers in Human Neuroscience, 7,* 229. http://dx.doi.org/10.3389/fnhum.2013.00229

Tassy, S., Oullier, O., Cermolacce, M., & Wicker, B. (2009). Do psychopathic patients use their DLPFC when making decisions in moral dilemmas? *Molecular Psychiatry, 14,* 908–909. http://dx.doi.org/10.1038/mp.2009.71

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., . . . Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social Cognitive and Affective Neuroscience, 7,* 282–288. http://dx.doi.org/10.1093/scan/nsr008

Tempesta, D., Couyoumdjian, A., Moroni, F., Marzano, C., De Gennaro, L., & Ferrara, M. (2012). The impact of one night of sleep deprivation on moral judgments. *Social Neuroscience, 7,* 292–300. http://dx.doi.org/10.1080/17470919.2011.614002

Thompson, V., & Evans, J. S. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning, 18,* 278–310. http://dx.doi.org/10.1080/13546783.2012.670752

Thorndike, E. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), 1–109.

Timmons, S., & Byrne, R. M. (2019). Moral fatigue: The effects of cognitive fatigue on moral reasoning. *Quarterly Journal of Experimental Psychology, 72,* 943–954.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016). Intuition and moral decision-making: The effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLoS ONE, 11*(10), e0164012. http://dx.doi.org/10.1371/journal.pone.0164012

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient Kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin, 40,* 923–930. http://dx.doi.org/10.1177/0146167214530436

Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2018). Reasoning and moral judgment: A common experimental toolbox. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge international handbook of thinking and reasoning* (1st ed., pp. 575–589). New York, NY: Routledge/Taylor & Francis Group. Retrieved from https://psycnet.apa.org/record/2017-56397-032

Trémolière, B., Neys, W. D., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition, 124,* 379–384. http://dx.doi.org/10.1016/j.cognition.2012.05.011

Van Hoeck, N., Watson, P. D., & Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience, 9,* 420. http://dx.doi.org/10.3389/fnhum.2015.00420

Wagstyl, K., & Lerch, J. P. (2018). *Cortical thickness* (pp. 35–49). New York, NY: Humana Press.

Weissman, D. H., Perkins, A. S., & Woldorff, M. G. (2008). Cognitive control in social situations: A role for the dorsolateral prefrontal cortex. *NeuroImage, 40,* 955–962. http://dx.doi.org/10.1016/j.neuroimage.2007.12.021

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition, 126,* 364–372. http://dx.doi.org/10.1016/j.cognition.2012.11.002

Wiegand, A., Sommer, A., Nieratschker, V., & Plewnia, C. (2019). Improvement of cognitive control and stabilization of affect by prefrontal transcranial direct current stimulation (tDCS). *Scientific Reports, 9,* 6797. http://dx.doi.org/10.1038/s41598-019-43234-2

Yamagishi, T., Takagishi, H., Fermin, A. de, S. R., Kanai, R., Li, Y., & Matsumoto, Y. (2016). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 5582–5587. http://dx.doi.org/10.1073/pnas.1523940113

Yang, W., Liu, P., Wei, D., Li, W., Hitchman, G., Li, X., . . . Zhang, Q. (2014). Females and males rely on different cortical regions in Raven's Matrices reasoning capacity: Evidence from a voxel-based morphometry study. *PLoS ONE, 9*(3), e93104. http://dx.doi.org/10.1371/journal.pone.0093104

Yuan, P., & Raz, N. (2014). Prefrontal cortex and executive functions in healthy adults: A meta-analysis of structural neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 42,* 180–192. http://dx.doi.org/10.1016/j.neubiorev.2014.02.005

Zheng, H., Lu, X., & Huang, D. (2018). tDCS over DLPFC leads to less utilitarian response in moral-personal judgment. *Frontiers in Neuroscience, 12,* 193. http://dx.doi.org/10.3389/fnins.2018.00193

Received February 4, 2019
Revision received November 16, 2019
Accepted November 19, 2019 ■