

Instrumental Harm and Impartial Beneficence Distinctively Frame Cognitive Representations of Moral Decision Problems

Yoonseo Zoh^{1, 2}, Hongbo Yu³, Luis Sebastian Contreras-Huerta^{4, 5}, Annayah M. B. Prosser^{6, 7},
Matthew A. J. Apps⁴, Walter Sinnott-Armstrong^{8, 9}, Steve W. C. Chang², and M. J. Crockett^{1, 2, 10}

¹ Department of Psychology, Princeton University

² Department of Psychology, Yale University

³ Department of Psychological and Brain Sciences, University of California, Santa Barbara

⁴ Centre for Human Brain Health, School of Psychology, University of Birmingham

⁵ Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez

⁶ School of Management, University of Bath

⁷ Department of Psychology, University of Bath

⁸ Department of Philosophy, Duke University

⁹ Kenan Institute for Ethics, Duke University

¹⁰ University Center for Human Values, Princeton University

Utilitarian ethical theories argue that the morality of actions depends on their consequences for impartially maximizing overall welfare. Recent research suggests that individual differences in utilitarian tendencies fall along two dimensions: a permissive attitude toward harming others for greater good (instrumental harm [IH]) and an impartial concern for others' welfare (impartial beneficence [IB]). We hypothesize that these dimensions operate as intuitive theories in the moral domain, framing distinctive patterns of moral judgments and behavior. Using intersubject representational similarity analysis of behavioral data ($N = 254$), we found that when participants shared endorsement of instrumental harm or impartial beneficence, they showed similar patterns of moral judgment and decision making. Intersubject representational similarity analysis of functional neuroimaging data ($N = 68$) revealed that participants with similar endorsement of instrumental harm or impartial beneficence showed similar neural encoding of moral choice attributes, even when they made different choices. Meanwhile, participants with dissimilar endorsement of these dimensions showed distinctive neural encoding of moral choice attributes, even when they made similar choices. These similarity and dissimilarity patterns emerged in distinct brain regions for instrumental harm and impartial beneficence. Together, our findings suggest that instrumental harm and impartial beneficence distinctively frame cognitive representations of moral decision problems, over and above guiding judgments and decisions.

Public Significance Statement

Utilitarianism is an influential ethical theory that judges actions based on their consequences, aiming to impartially maximize overall well-being. Recent research suggests utilitarian thinking has two key dimensions: impartial concern for others' welfare and a willingness to cause harm for the greater good. However, it remains unclear whether these two dimensions of utilitarian thinking shape moral judgments and decisions in different ways. Here we explore whether people who agree with one another on either dimension show more similar patterns of moral judgment, decision making, and brain activity. We show that when people agree on a dimension of utilitarian thinking, they show more similar patterns of moral judgment and more similar patterns of brain activity when making moral decisions—even when they make different decisions. These findings suggest that agreement on a moral principle influences the meaning people make from their moral actions.

Keywords: utilitarianism, moral cognition, moral psychology, neuroimaging, social neuroscience

Supplemental materials: <https://doi.org/10.1037/xge0001820.supp>

Ana Gantman served as action editor.

Yoonseo Zoh  <https://orcid.org/0000-0002-3659-5638>

M. J. Crockett  <https://orcid.org/0000-0001-8800-410X>

These results have not been disseminated or published elsewhere. The authors declare no competing interests. This work was supported by the John Templeton Foundation (Beacons Project and No. 61495), Academy of Medical Sciences (Grant SBF001\1008), John Fell Fund, University of

Oxford, and Wellcome Trust (Grant 204826/Z/16/Z) awarded to M. J. Crockett; Korea Foundation for Advanced Studies awarded to Yoonseo Zoh; Royal Society (Grant NF160700) and Yale University (Theresa Seessel Endowed Fellowship) awarded to Hongbo Yu; Biotechnology and Biological Sciences Research Council David Phillips Fellowship (BB/R010668/2) awarded to Matthew A. J. Apps. Because this research was funded in whole, or in part, by Wellcome Trust (Grant 204826/Z/16/Z), for the purpose of open access, the author has applied a CC BY public copyright license to any author

continued

Utilitarianism is a highly influential ethical theory that argues the moral status of actions depends solely on their consequences for impartially maximizing overall welfare (Bentham, 1983; de Lazari-Radek & Singer, 2017; Mill, 1969; Sidgwick, 1981; Singer, 2011). Over the past 2 decades, psychologists have extensively investigated the neural and cognitive mechanisms that guide endorsement of utilitarian principles. Many of these studies measured responses to sacrificial dilemmas (like the famous “trolley problem”) that probe endorsement of *instrumental harm* (IH): the moral permissibility of harming some people for a greater good (Everett & Kahane, 2020; Kahane et al., 2018). This work has demonstrated that endorsement of instrumental harm is positively associated with deliberative cognitive processing, clinical and subclinical psychopathy, and social and economic conservatism (Bartels, 2008; Bartels & Pizarro, 2011; Capraro et al., 2019; Conway et al., 2018; Conway & Gawronski, 2013; Glenn et al., 2010; Kahane et al., 2015, 2018; Koenigs et al., 2012; Patil et al., 2021; Tassy et al., 2013); it is negatively associated with aversion to harming others, empathic concern, and identification with all of humanity (Gleichgerrcht & Young, 2013; Kahane et al., 2015, 2018; Miller & Cushman, 2013; Patil & Silani, 2014; Takamatsu, 2018). Neurally, endorsement of instrumental harm is associated with both activation and cortical thickness in brain regions involved in reasoning and deliberation, such as the dorsolateral prefrontal cortex (brodmann areas 10 and 46; Greene et al., 2004, 2008; Patil et al., 2021). One interpretation of this literature is that endorsement of instrumental harm relies more on deliberative cognitive processes than emotional processes (Greene, 2009; but see also Kahane et al., 2012).

Other work has sought to characterize a second dimension of utilitarianism called *impartial beneficence* (IB): an impartial concern for maximizing welfare overall, not privileging self or close others over distant strangers. According to the two-dimensional model of utilitarian psychology, instrumental harm and impartial beneficence are psychologically dissociable (Everett & Kahane, 2020; Kahane et al., 2018). In contrast with instrumental harm, endorsing impartial beneficence is positively associated with characteristics such as expanded circle of moral and empathic concern, religiosity, and charitable giving (Amormino et al., 2022; Earp et al., 2024; Fowler et al., 2021; Kahane et al., 2018; Paruzel-Czachura & Charzyńska, 2022; Syropoulos et al., 2023; Tuen et al., 2023). Behaviorally, individuals strongly endorsing impartial beneficence demonstrate more uniform concern for the well-being of others across varying levels of social distance, showing a reduced tendency to discount the importance of outcomes affecting socially distant individuals (Earp et al., 2024; Vekaria et al., 2017). Neurally, greater endorsement of impartial

beneficence is associated with less discriminate encoding of outcomes across self and others at varying social distances, reflected in more uniform representation of others’ welfare in the rostral anterior cingulate cortex and amygdala, along with overlapping neural responses to pain for self and others in the left anterior insula (Brethel-Haurwitz et al., 2018; Rhoads et al., 2023). Overall, this literature has revealed the affective underpinnings of impartial beneficence, highlighting its connections with vicarious emotional responses to motivationally salient outcomes for others.

The two-dimensional model of utilitarian psychology draws attention to the unique psychological characteristics of instrumental harm and impartial beneficence, and thus far the broader literature supports the claim that these dimensions are indeed psychologically and neurally distinctive. However, several open questions remain. First, most past research can only *indirectly* support claims that instrumental harm and impartial beneficence are psychologically and neurally distinct, because it has typically probed these dimensions in isolation from one another. Here, we measure both dimensions in the same participants and allow them to compete for variance in explaining moral judgment, moral decision making, and its neural correlates. Second, the broader consequences of the two-dimensional model for moral life remain unclear. Why does it matter if different people endorse different dimensions of utilitarian ethics? Here, we consider the implications of the two-dimensional model for moral agreement and disagreement, asking whether people who agree on each dimension represent moral scenarios in similar ways and whether those who disagree on each dimension represent the same moral scenarios in more dissimilar ways.

We approach these questions by considering instrumental harm and impartial beneficence as *intuitive moral theories* that guide the formation of beliefs and behaviors in ambiguous moral scenarios (Crockett et al., 2024; Gottlieb & Lombrozo, 2018). To the extent two people share an intuitive moral theory, they should cognitively represent moral situations in a similar way. Below, we briefly review the literature on intuitive moral theories and articulate several exploratory hypotheses that we test with behavioral and neural data in the present work.

Intuitive Moral Theories

Intuitive theories (sometimes called “implicit theories” or “lay theories”) are informal accounts of particular domains used to support prediction and explanation. They emerge without any formal education or instruction, representing our sense of “how things work” in the world (Keil, 2024). Intuitive theories are typically not as comprehensive or explicitly articulated as scientific or philosophical

accepted manuscript version arising from this submission. The authors are grateful to the members of Crockett Lab, Robb Rutledge, Michel-Pierre Coll, and Diana Tamir, for helpful discussions on this work.

Yoonseo Zoh played a lead role in data curation, formal analysis, methodology, software, validation, and visualization and an equal role in conceptualization, writing—original draft, and writing—review and editing. Hongbo Yu played a lead role in data curation and investigation and a supporting role in conceptualization, methodology, resources, software, supervision, and writing—review and editing. Luis Sebastian Contreras-Huerta played a supporting role in investigation and writing—review and editing and an equal role in data curation. Annayah M. B. Prosser played a supporting role in conceptualization, project administration, and writing—review and editing and an equal role in data

curation, investigation, and methodology. Matthew A. J. Apps played a supporting role in conceptualization, methodology, supervision, and writing—review and editing. Walter Sinnott-Armstrong played a supporting role in conceptualization, funding acquisition, methodology, and writing—review and editing. Steve W. C. Chang played a supporting role in supervision and writing—review and editing and an equal role in formal analysis. M. J. Crockett played a lead role in conceptualization, funding acquisition, methodology, project administration, resources, supervision, and writing—review and editing, a supporting role in formal analysis, and an equal role in writing—original draft.

Correspondence concerning this article should be addressed to M. J. Crockett, Department of Psychology, Princeton University, South Drive, Princeton, NJ 08540, United States. Email: mj.crockett@princeton.edu

theories, but they share key features, like specifying domain-specific ontologies of concepts (e.g., mental states like beliefs, desires, and intentions for an “intuitive theory of mind”) and principles that causally connect those concepts (e.g., “people tend to take rational actions based on their desires and beliefs”; Gerstenberg & Tenenbaum, 2017). Ultimately, intuitive theories operate as “belief-generating systems,” structuring our responses to ambiguous situations (Mahr & Csibra, 2021). Some intuitive theories are widely shared among adults within a given culture, while other intuitive theories are more variable across individuals and situations (e.g., “fixed” vs. “growth” theories of personal change; Dweck et al., 1995). When we share an intuitive theory with others, we represent and understand the world in a similar way; when we hold different intuitive theories from others, we see things differently.

In the moral domain, intuitive theories represent informal accounts of “how things ought to be.” They hold some resemblance to philosophical theories in being abstract, rule-based, internally consistent, and responsive to evidence (Gottlieb & Lombrozo, 2018), but intuitive moral theories are not necessarily as comprehensive as full-fledged philosophical theories. For example, as reviewed above, the philosophical theory of utilitarianism encompasses both instrumental harm and impartial beneficence. In the lay population, however, these subcomponents of utilitarianism are psychologically dissociable; endorsing the abstract principle of instrumental harm does not necessarily entail endorsing impartial beneficence, and vice versa (Kahane et al., 2018). Moreover, these subcomponents of utilitarianism might shape moral judgments and behavior implicitly, in ways that are not necessarily easy to articulate. This suggests that instrumental harm and impartial beneficence might be more appropriately conceptualized as intuitive moral theories than full-fledged philosophical theories (Crockett et al., 2024; Everett & Kahane, 2020).

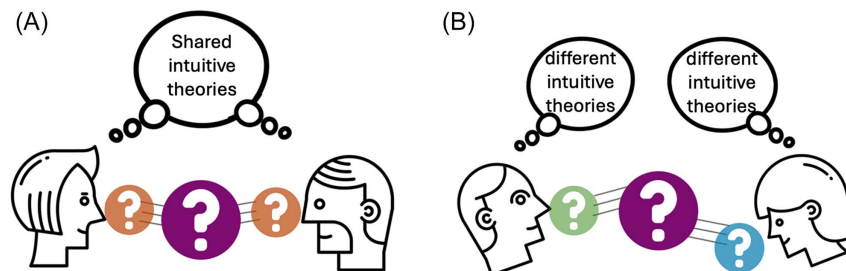
If instrumental harm and impartial beneficence operate as distinct intuitive moral theories, commitment to these principles in the abstract should support moral judgments in concrete cases tapping those same principles. Lombrozo (2009) provided some initial evidence for this by demonstrating that explicit commitments to utilitarian principles predict moral permissibility judgments in sacrificial dilemmas. However, this work did not distinguish between instrumental harm and

impartial beneficence. Kahane et al. (2018) made this distinction and showed that abstract endorsement of instrumental harm predicted concrete moral judgments about the appropriateness of instrumental harm but not impartial beneficence. Endorsement of impartial beneficence in the abstract, however, predicted concrete moral judgments about appropriateness of both impartial beneficence and instrumental harm. Thus, while instrumental harm and impartial beneficence are robustly distinguishable as abstract principles, it remains unclear to what extent they support distinct patterns of concrete moral judgments.

Conceptualizing instrumental harm and impartial beneficence as intuitive moral theories also might illuminate how these constructs relate to moral agreement and disagreement. Because intuitive moral theories are hypothesized to guide how we represent and understand moral situations, to the extent that two people *similarly endorse* an intuitive moral theory, they should process moral scenarios similarly. Conversely, to the extent that two people *differentially endorse* an intuitive moral theory, they should process moral scenarios differently (Figure 1).

Exploring this hypothesis requires a different analytic approach than has typically been used in studies of instrumental harm and impartial beneficence. The typical approach employs univariate correlation to ask, for example, which psychological and neural processes are associated with *higher or lower levels* of instrumental harm or impartial beneficence. By contrast, multivariate approaches are better suited for exploring how *shared or differential endorsement* of instrumental harm or impartial beneficence explains *similar or different* patterns of moral judgment and behavior. One notable approach, intersubject representational similarity analysis (IS-RSA), deploys intersubject correlation to measure the extent to which dyadic similarities or differences on a psychological construct predict dyadic similarities or differences in multivariate patterns of brain activity (Chen et al., 2020; Chen & Qu, 2021; Finn et al., 2020; Kuo et al., 2023; Nastase et al., 2019; Nummenmaa et al., 2012; Stephens et al., 2010; van Baar et al., 2019, 2021; Yeshurun et al., 2017). Rather than testing hypotheses about how high or low levels of psychological constructs predict univariate responses in particular brain regions of interest, this approach asks whether similarities or

Figure 1
The Role of Intuitive Moral Theories in Shared Representation of Moral Problems



Note. Panel (A): Two individuals who share intuitive moral theories approach the same moral decision problem in a similar way, which increases the likelihood of arriving at similar moral judgments. Panel (B): In contrast, two individuals presented with the same moral decision problem but representing it differently guided by their differential endorsement of intuitive moral theories, leading to disagreement in their moral judgment. The human icons in were downloaded from the Noun Project. They are covered under a royalty-free license through subscription plan. See the online article for the color version of this figure.

differences in psychological constructs predict similar or distinctive multivariate responses across the whole brain (although observing where in the brain these correlations manifest can potentially be informative about the underlying mechanisms).

Recent studies have used IS-RSA to examine how shared versus conflicting ideologies predict neural responses to moral and political content. For example, Leong et al. (2020) measured neural activity in liberals and conservatives while they watched the same set of political advertisements about immigration policy. They used IS-RSA to show that liberals showed more similar neural activity to other liberals and conservatives more similar neural activity to other conservatives. This “neural polarization” effect was concentrated in a region of dorsomedial prefrontal cortex associated with processing narrative content. A similar study by de Bruin et al. (2023) found that shared political ideology predicted similar neural responses to processing political concepts, which in turn predicted similar neural responses in the striatum and amygdala to videos of political debates. A third study by van Baar et al. (2019) measured patterns of behavior in economic trust games to identify clusters of participants who spontaneously and consistently employed distinctive “moral strategies” (e.g., guilt aversion or inequality aversion) to guide their decisions. Using IS-RSA, they showed that people with different strategies showed more distinctive patterns of neural activity during decision making, even when they made the same decisions (van Baar et al., 2019). These studies are suggestive that intuitive moral theories could align neural responses to moral scenarios, even though they did not test this question directly.

The Present Studies

We used three sources of data to explore whether endorsement of instrumental harm or impartial beneficence frames distinctive psychological and neural responses. First, we examined patterns of moral judgment in concrete scenarios designed to tap instrumental harm and impartial beneficence. We predicted that differential endorsement of instrumental harm in the abstract would more strongly predict dissimilarity in concrete judgments about instrumental harm than impartial beneficence. Likewise, we predicted that differential endorsement of impartial beneficence in the abstract would more strongly predict dissimilarity in concrete judgments about impartial beneficence than instrumental harm.

Second, we explored how differential endorsement of instrumental harm and impartial beneficence predicts patterns of moral decision making in a task that does *not* probe these constructs directly. The task invites participants to either earn money by inflicting mild electric shocks on themselves or a stranger or forgo money in order to prevent the pain from being delivered (Crockett et al., 2014, 2017). Though the task itself is highly artificial (most of us are not routinely being offered money to inflict electric shocks on another), the structure of the task mirrors a common moral situation we face in daily life, where our own interests trade off with those of others.

The decisions participants make in this task can reflect distinctive utilitarian principles; for example, a participant endorsing impartial beneficence might endeavor to weight their own welfare equally against the stranger’s; a participant endorsing instrumental harm might plan to donate money earned from shocking a stranger to a charity and therefore see that action as morally acceptable or even obligatory. However, unlike the concrete scenarios that serve as our

first source of data, there is not a straightforward 1:1 mapping between endorsement of utilitarian principles and predicted choice patterns in this task. This ambiguity allows us to investigate the extent to which endorsement of instrumental harm and impartial beneficence frames decision making in moral scenarios that do not directly tap these constructs. If we observe distinctive patterns of moral decision making associated with differential endorsement of instrumental harm and impartial beneficence, respectively, this would be evidence that these intuitive moral theories frame participants’ interpretation of moral situations beyond dilemmas that probe them directly. And to the extent that endorsement of instrumental harm and impartial beneficence predict different aspects of moral decision making, this would be evidence that these two dimensions operate as distinctive intuitive theories.

That said, given the ambiguity of the task, it is possible that two people could make the same decisions for very different moral reasons if they subscribe to different intuitive moral theories. This could manifest as distinctive neural patterns when making the same decisions, just as liberals and conservatives showed distinctive brain patterns when watching the same political videos (de Bruin et al., 2023; Leong et al., 2020) and just as people deploying different moral strategies showed distinctive brain patterns during a trust game when making the same decisions (van Baar et al., 2019). To explore this hypothesis, we turn to a third source of data: neural activity during the moral decision task. We used IS-RSA to measure the relationship between divergent endorsement of utilitarian principles and neural encoding of choice attributes, controlling for the decisions participants made. We predicted that pairwise disagreement about instrumental harm and impartial beneficence would predict pairwise neural dissimilarity during moral decision making, over and above the decisions that people make. Furthermore, if instrumental harm and impartial beneficence reflect independent intuitive moral theories, neural response patterns associated with each should be observed in distinct sets of brain areas. Overall, our neuroimaging analyses address a previously unexplored question: When people are making moral decisions, does their abstract endorsement of utilitarian principles shape the way they represent the decision problem—regardless of what choices they ultimately make?

Method

Transparency and Openness

We made the materials, behavioral data, and analysis codes for these studies publicly available on the Open Science Framework (<https://osf.io/ntu79/>). The design and analysis plans for the studies were not preregistered. For Study 1, the sample size was determined based on the available funding for running the study. For Study 2, 80 healthy volunteers between the ages of 18 and 38 years were recruited from the University of Oxford and the local Oxford, United Kingdom, community. The recruitment was designed to achieve a target sample size of 64 participants, accounting for expected participant dropout based on Crockett et al. (2017). This sample size was determined to be sufficient for detecting brain–behavior correlations (r_s) of 0.3 with 80% power. Data collection was terminated once we reached the predetermined sample size ($N = 80$). Study 1 was approved by the Yale Human Subjects Committee (No. 2000022849), and Study 2 was approved by the University of Oxford Research Ethics Committee (No. R50262/RE001).

Overview of Studies

In two studies, we measured abstract commitments to instrumental harm and impartial beneficence using the Oxford Utilitarianism Scale (OUS; Kahane et al., 2018) and examined their relationship with patterns of moral judgment and behavior. In both studies, participants completed informed consent and were paid for their participation.

Study 1 ($N = 192$) was conducted online between May and November 2018. Participants completed the OUS as part of a larger battery of trait questionnaires, moral judgment tasks probing instrumental harm and impartial beneficence, and a hypothetical moral decision-making task.

Study 2 ($N = 79$) was conducted online and in-person between September 2017 and April 2018. In an initial online session, participants completed the OUS as part of a larger battery of trait questionnaires and moral judgment tasks probing instrumental harm and impartial beneficence. At least 1 week later, they attended an in-person lab session where they completed an incentivized moral decision-making task in the functional magnetic resonance imaging (fMRI) scanner. Neuroimaging data from a subset of these participants ($N = 62$) have been previously analyzed and published in a separate study (Yu et al., 2022). This study tested a hypothesis that multivariate patterns of guilt during moral decision making predicted hypocritical blame measured with a separate behavioral task. None of the previously published neuroimaging analyses or behavioral analyses are included in the present article.

Participants

Study 1 participants were recruited from the Oxford community and online. Because we aimed for a sample whose endorsements of utilitarianism spanned the full measurable range, we advertised our study to members of the Effective Altruism (EA) organization, whom we expected to endorse high levels of utilitarianism. EA is “a research field and practical community that aims to find the best ways to help others, and put them into practice” (Effective Altruism, n.d.; MacAskill, 2019). Many individuals in the community pledge to give at least 10% of their lifetime earnings to “effective” charities believed to deliver the “most good” to those who need help. This recruitment process yielded 94 participants who identified as members of the EA community. We then recruited an additional 98 participants who matched the EA sample in age, gender, country of residence, and education level. These participants were recruited from the Oxford community with flyers and online advertisements as well as from a pool of prior study participants who had given permission to contact them about participating in future studies. For most of our analyses, we aggregated data from these two groups for a total of $N = 192$ participants (85 men, 101 women, six nonbinary, $M_{\text{age}} = 44.56$). Due to a coding error, trait psychopathy data are unavailable for $N = 17$ participants; analyses containing this variable are reported based on the 175 participants for which data are available.

Study 2 participants were recruited from the University of Oxford and local subject pools. For fMRI safety purposes, people with a history of neurological or neuropsychiatric disorders and pregnant women were excluded from participation. To limit suspicion about our incentivized moral decision task, people with more than 2 years of psychology education and/or those who had previously taken part in studies involving deception or electric shocks were excluded from participation. A total of 80 participants took part in the study. One

participant was excluded from analysis due to lack of completion of the OUS, leaving $N = 79$ participants for behavioral analysis (36 men, 42 women, one nonbinary, $M_{\text{age}} = 42.49$). For fMRI analysis, eight participants were excluded from analysis due to excessive head motion (>3 mm or $>3^\circ$ within one functional run) in the scanner. One participant’s neuroimaging data were not registered due to technical issues with the scanner. Two participants were excluded from analysis for expressing doubts regarding whether the receiver participant would actually receive the electric shocks. After these exclusions, we were left with $N = 68$ participants (27 men, 40 women, one nonbinary, $M_{\text{age}} = 42.43$).

In both studies, participants were asked to self-report their demographic information. Gender identity was collected by asking, “What is your gender?” Participants could select from the following options: (a) male, (b) female, (c) other/neither/both. Information about participants’ racial and ethnic identity was collected by asking, “What is your race?” with the following options provided: (a) White/Caucasian, (b) Black/African American, (c) Hispanic or Latino, (d) Asian, (e) Native American, (f) Pacific Islander, (g) other (with a free-response box).

For Study 1, the final sample consisted of 192 participants, of whom 85 identified as women, 101 identified as men, and six identified as nonbinary. Participants self-reported their racial and ethnic identities as follows: 131 identified as White, three as Black or African American, 13 as Hispanic or Latino/a/x, 37 as Asian, one as Native American, and seven as “other.” Among those who selected “other,” free responses included one participant identifying as “Ashkenazi/White,” one as “Middle Eastern,” one as “multiracial,” and one as “White/Hispanic.”

For Study 2, the sample for behavioral data analysis consisted of 79 participants, of whom 36 identified as men, 42 identified as women, and one identified as nonbinary. Regarding racial/ethnic identity, 58 participants identified as White, one identified as Black or African American, two identified as Hispanic or Latino/a/x, 13 identified as Asian, and five identified as “other.” Among those who selected “other,” free responses included one participant identifying as “Asian/European,” one as “Middle Eastern Mixed (White/Asian),” and one as “White/Black African.”

Procedure

In Study 1, participants followed a web link to an online Qualtrics survey. After providing informed consent, participants provided demographic information and completed a battery of questionnaires, including the OUS (Kahane et al., 2018), a trait psychopathy measure (Self-Report Psychopathy Scale–Short Form; Paulhus et al., 2016), a trait empathy measure (Questionnaire of Cognitive and Affective Empathy; Reniers et al., 2011), and a trait self-control measure (Brief Self-Control Scale; Tangney et al., 2018). They also completed two moral judgment tasks in a randomized order, one probing instrumental harm (sacrificial dilemmas from Greene et al., 2008) and one probing impartial beneficence (social value orientation dilemmas from Van Lange et al., 1997). In addition, they completed a hypothetical moral decision-making task (Contreras-Huerta et al., 2022; Crockett et al., 2014). Participants also answered several questions about their familiarity and involvement with the EA community.

Study 2 consisted of multiple sessions and was conducted at the Wellcome Centre for Integrative Neuroimaging and the Department of Experimental Psychology, University of Oxford. A week before the fMRI session, participants completed a battery of online measures

that included demographics, the OUS (Kahane et al., 2018), a trait psychopathy measure (Self-Report Psychopathy Scale–Short Form; Paulhus et al., 2016), a trait empathy measure (Questionnaire of Cognitive and Affective Empathy; Reniers et al., 2011), and a trait self-control measure (Brief Self-Control Scale; Tangney et al., 2018). They also completed two moral judgment tasks administered in a randomized order, one probing instrumental harm (sacrificial dilemmas from Greene et al., 2008) and one probing impartial beneficence (impartiality dilemmas presented through social value orientation task from Van Lange et al., 1997).

Upon arrival at the fMRI session, participants provided informed consent and went through a pain thresholding procedure in preparation for an incentivized moral decision-making task (for details, see Crockett et al., 2014). The purpose of the pain thresholding procedure was twofold: (a) to familiarize participants with the experience of the shocks, which they would later take into account in their decision making, and (b) to determine the physical intensity of the shocks so that their subjective intensity was matched across participants. After the thresholding procedure, participants were instructed they would be randomly assigned to roles of either decider or receiver using a procedure that has been described in detail elsewhere (Crockett et al., 2014, 2017). In reality, participants were always assigned to the role of decider, and the role of receiver was played by a confederate.

Following role assignment, participants received instructions for the moral decision task, answered comprehension questions, and practiced outside the scanner for six trials. Participants were informed that their choices would be anonymous and that they would not meet or interact with the receiver. This was done to minimize concerns about reputation or reciprocity in their decision making. They then completed the moral decision task in the fMRI scanner.

Functional MRI scanning was performed on a 3-Tesla Siemens Prisma scanner at the Wellcome Centre for Integrative Neuroimaging at the University of Oxford. Functional images were obtained with multiband T2*-weighted echo-planar imaging sequence. The echo-planar imaging images were acquired in an ascending manner, at an oblique angle (approximately 30°) to the anterior commissure–posterior commissure plane to minimize the signal dropout in the orbitofrontal areas. The following acquisition parameters were used: voxel size = 2 × 2 × 2 mm, echo time (TE) = 30 ms; repetition time = 1,570 ms; flip angle = 90°; field of view = 216 × 216 mm. The structural image was taken using a magnetization prepared rapid gradient echo sequence with 192 slices; repetition time = 1,900 ms; TE = 3.97 ms; field of view = 192 × 192 mm; voxel size = 1 × 1 × 1 mm resolution. We also acquired a field map (short TE = 4.92 ms; long TE = 7.38 ms; repetition time = 482.0 ms; resolution = 2 × 2 × 2 mm; field of view = 219 × 219 mm) to correct distortions in the functional images.

Moral Judgment Tasks

To probe concrete moral judgments concerning instrumental harm, we used three sacrificial dilemmas from Greene et al. (2008): the trolley dilemma, the cruise dilemma, and the soldier dilemma. In each scenario, participants judged whether it is morally permissible to harm one person to save a greater number. Participants responded on a scale from 1 (*not morally wrong at all*) to 7 (*very morally wrong*); for data analysis, their scores were averaged across all three scenarios. The full text of the scenarios is available in the [Supplemental Materials](#).

To probe concrete moral judgments concerning impartial beneficence, we presented impartiality dilemma using the social value orientation task (Van Lange et al., 1997). On each trial, participants choose between three hypothetical allocations of money to themselves and a stranger. The “impartial” option allocates equal amounts to self and other. The “individualist” option allocates a higher amount to self and a lower amount to other, relative to the prosocial option. The “competitive” option allocates a lower amount to self and a comparably even lower amount to other, relative to the prosocial option. Participants completed a total of nine trials (full list is available in the [Supplemental Materials](#)). For data analysis, participants’ responses were rated as a continuous sum across items: The impartial option was assigned a score of 3, the individualistic option a score of 2, and the competitive option a score of 1.

Moral Decision-Making Task

In the moral decision-making task, participants chose between two options that involved different combinations of profit (money) and pain (electric shocks). Across trials, one option (harmful option) always contained more pain and profit, and the other option (helpful option) contained less pain and profit. On half the trials, the pain recipient was the participant (self condition); on the other half of trials, the pain recipient was the “receiver,” an anonymous stranger (other condition). The profit was always for the participant (Figure 2).

In Study 1, participants completed a hypothetical version of the task. They were instructed to imagine a hypothetical experiment where they were randomly assigned to the role of decider, and another unknown participant was assigned to the role of receiver. They were also told to imagine that one trial would be randomly selected and actually implemented. Participants were randomly assigned to one of three trial sets following the procedures described in previous work (Crockett et al., 2014). Participants completed 35 trials in both the self and other conditions, with 70 trials in total.

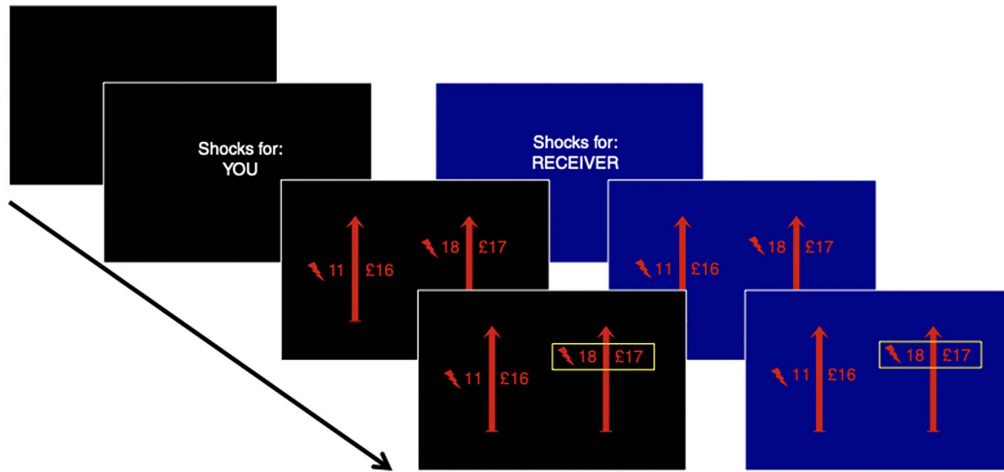
In Study 2, participants completed an incentivized version of the task. They were informed that one of their choices would be randomly selected and implemented at the end of the scanning session. This version had a total of 152 trials. We first created a set of 72 trials according to the criteria reported in previous work (Crockett et al., 2017). In addition to those trials, four “catch” trials were inserted, where the more harmful option containing a smaller amount of money was also inserted. This resulted in 76 trials in the set. Half of the trials were randomly selected to present the more harmful option on the right-hand side of the screen, whereas in the other half of the trials the more harmful option was presented on the left-hand side of the screen. Then the same trial set was duplicated to match the total number of trials and the choice options in two recipient conditions. Trials were then distributed into two scanning runs of 76 trials, each containing equal numbers of trials of self and other conditions. Four different trial sets were created in this way, which were randomly assigned to the participants.

Data Analysis

Computational Modeling of Moral Decision Making

To analyze data from the moral decision-making task, we applied a computational model to quantify the subjective cost of harming self and others. Starting from a previously validated

Figure 2
The Schematic of Harm Aversion Task



Note. In each trial, participants chose between a harmful option containing more profit and pain and a helpful option containing less profit and pain. The selected option was highlighted by a yellow box. The money was always for the participant, but on half of the trials the pain were for the participant (self condition) and on the other half the pain were for the other individual in the receiver role (other condition). See the online article for the color version of this figure.

model (Crockett et al., 2014, 2017), we estimated individual parameters using nonlinear optimization in MATLAB (MathWorks, Inc.) with maximum likelihood estimation. Model variations were compared using the Bayesian information criterion to assess goodness of fit, with the lowest group-level Bayesian information criterion score determining the best fitting model which we used for subsequent analysis (see Supplemental Table S1 for model details).

The selected model included three parameters: separate harm aversion parameters for self and other and an inverse temperature parameter. The model assumes that subjective value differences between the harmful and helpful option are determined by weighted differences in profit and pain between two options. The harm aversion parameter (κ) captured the subjective cost of harming self and other, with higher κ values indicating greater harm aversion and willingness to sacrifice more profit to reduce pain. Conversely, lower κ values suggest reduced harm aversion and a tendency to prioritize profit over harm reduction. The inverse temperature parameter modulates how strongly value differences influence choices, which are transformed into probabilities using a softmax function.

$$\begin{aligned} \Delta V &= (1 - \kappa_{\text{self}})\Delta m - \kappa_{\text{self}}\Delta s \text{ if self trial} \\ \Delta V &= (1 - \kappa_{\text{other}})\Delta m - \kappa_{\text{other}}\Delta s \text{ if other trial} \\ P(\text{choose alternative}) &= \frac{1}{1 + e^{-\beta\Delta V}} \text{ if self trial.} \end{aligned} \tag{1}$$

Behavioral IS-RSA With Dyadic Regression Models

To examine whether participants who differentially endorse utilitarian principles show dissimilar patterns of moral judgments and decisions, we applied IS-RSA using dyadic regression models to behavioral data. These models tested whether the degree to which paired participants differentially endorse instrumental harm and impartial beneficence predicts dissimilarity in their behavior.

In all models, we regressed pairwise dissimilarity in a given behavioral measure against pairwise dissimilarity in instrumental harm, impartial beneficence, trait psychopathy, trait empathy, and trait self-control. This approach enabled us to identify the unique predictive effects attributable to divergent endorsement of each utilitarian principle, over and above divergence in the other principle and related psychological traits. Random intercepts for participants were included in the dyadic regression model to account for statistical dependencies arising from the repeated occurrence of individual data in pairwise observations. The models were implemented in R using code adapted from previous work using the same approach (van Baar et al., 2019).

For the two moral judgment tasks directly assessing impartial beneficence and instrumental harm, we conducted separate dyadic regression models. For both the sacrificial dilemmas and the impartiality dilemmas, the Euclidean distance in responses (as simple difference in values) was modeled as a function of the distance in impartial beneficence and instrumental harm while controlling for trait psychopathy, empathy, and self-control.

For the moral decision-making task, we ran dyadic regression on the harm aversion parameters estimated from the best fitting computational model of behavior. We calculated the Euclidean distance between each pair of participants for both κ_{other} and κ_{self} . To examine the relative relationship between these two parameters (the relative orientation between κ_{other} and κ_{self}), we employed the cosine distance ($1 - \text{cosine similarity}$) of the vector defined by κ_{other} and κ_{self} for each pair. This cosine distance served as the dependent variable in the dyadic regression model, with the distances in impartial beneficence and instrumental harm as predictors while controlling for trait psychopathy, empathy, and self-control.

Finally, as an exploratory analysis, we leveraged Study 1 data that included participants from the EA community. These participants rated the importance of EA involvement to their identity (“My involvement with EA is an important part of my identity”). Participants who were

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

not involved with EA were assigned a score of “0” on this measure. Since EA is a movement grounded in utilitarian principles, we reasoned that identification with the EA community could amplify the above relationships between dissimilar endorsement of utilitarian principles and dissimilarity in moral judgment and decision making. To test this hypothesis, we calculated the pairwise dissimilarity in EA identification between participants and included this as an interaction term in the dyadic regression models described above. Specifically, we allowed pairwise dissimilarity in EA identification to interact with pairwise dissimilarity in instrumental harm and impartial beneficence, while controlling for pairwise dissimilarity in trait psychopathy, trait empathy, and trait self-control.

fMRI Analysis

Preprocessing

MRI data were preprocessed and analyzed using SPM12 (<https://www.fil.ion.ucl.ac.uk/spm>). Functional images were realigned and unwarped with reference to the fieldmap and coregistered to the participant’s own structural image. The structural images underwent routine preprocessing steps, including segmentation, bias correction, and spatial normalization to the Montreal Neurological Institute template. Finally, images were spatially smoothed with an SPM default Gaussian kernel (8-mm full-width at half maximum).

Parametric Model of Decision Parameters

A generalized linear model (GLM) was conducted to obtain β estimates of blood-oxygen-level-dependent (BOLD) activation parametrically responding to the relative amounts of profit and pain between the harmful option and the helpful option, separately for self and other conditions, at the decision onset. BOLD time series were regressed onto a model with two main event regressors indicating the onsets of self trials and other trials. Those two events were modeled with a duration corresponding to the participant’s response time on that trial. Each was associated with two parametric modulators: the relative amount of profit (Δ profit) and pain (Δ pain) between the harmful and the helpful option irrespective of participants’ choices. We included additional event regressors of no interest, indicating the onsets of left button press, right button press, transition to self condition, transition to other condition, as well as six nuisance regressors to control for head motion. The β estimates of the relative amount of pain in other condition (Δ pain other) obtained from running this GLM were collapsed across two runs by averaging and used in subsequent IS-RSA.

Neural IS-RSA

In our dyadic regression approach to neural data, the dependent variable measuring neural distance was derived by calculating the Euclidean distance between pairs of participants based on their parametric BOLD responses encoding choice attributes—specifically, relative pain and profit amount for each voxel—based on the GLM described earlier. Here, we focused our analysis on trials in the other condition since our interest was to examine divergent neurocognitive patterns within the context of moral decision making. For the independent variable, we calculated the Euclidean distance between all pairs of participants for impartial beneficence and instrumental harm scores and all measures that we controlled, including harm aversion

parameter of κ other, trait psychopathy, trait self-control, and trait-empathy.

The dyadic regression model mapped distance between participants in endorsement of utilitarian principles onto brain activity by regressing dissimilarity in β estimates onto dissimilarity in impartial beneficence and instrumental harm scores. We included dissimilarity in κ_{other} as a covariate. This approach enabled us to identify multivariate neural patterns uniquely associated with dissimilarity of each utilitarian principle, controlling for dissimilarity in choices.

Each dyadic regression model was estimated at the voxel level, where observations corresponded to all unique pairs of participants. The resulting β estimates for dissimilarity in impartial beneficence and instrumental harm were subsequently mapped onto 3D brain space, generating a β map. Statistical significance of the β map was tested using voxel-wise thresholding at p (False Discovery Rate correction) $< .05$. We reported clusters that survived this thresholding, requiring a minimum size of five contiguous voxels, consistent with a recent study employing the same method (van Baar et al., 2019).

Results

Differential Endorsement of Utilitarian Principles Predicts Disagreement in Concrete Moral Judgments

We first examined whether dissimilar endorsement of instrumental harm and impartial beneficence in the abstract (as measured by the OUS) predicts divergence in concrete moral judgments about scenarios directly tapping these constructs. Specifically, we used dyadic regression to test whether participants who disagreed in their endorsement of instrumental harm and impartial beneficence in the abstract were more likely to also disagree in their concrete moral judgments. All regression models included trait psychopathy, empathy, and self-control as regressors of no interest.

As expected, we found that differential endorsement of instrumental harm on the OUS predicted disagreement in moral judgments of sacrificial dilemmas, Study 1: $B = 0.20$, $SE = 0.01$, $t(15210) = 24.65$, $p < .001$, confidence interval (CI) [0.19, 0.22]; Study 2: $B = 0.32$, $SE = 0.02$, $t(3046) = 17.26$, $p < .001$, CI [0.29, 0.36]. However, differential endorsement of instrumental harm did not predict dissimilarity in moral judgments of impartiality dilemma, Study 1: $B = 0.01$, $SE = 0.01$, $t(15210) = 0.70$, $p = .482$, CI [−0.01, 0.01]; Study 2: $B = -0.01$, $SE = 0.02$, $t(3075) = -0.43$, $p = .671$, CI [−0.04, 0.03].

Turning to impartial beneficence, we found—as expected—that differential endorsement of impartial beneficence on the OUS predicted dissimilarity in responses to impartiality dilemma, Study 1: $B = 0.04$, $SE = 0.01$, $t(15200) = 5.41$, $p < .001$, CI [0.03, 0.06]; Study 2: $B = 0.03$, $SE = 0.017$, $t(2992) = 1.99$, $p = .047$, CI [−0.01, 0.07]. However, differential endorsement of impartial beneficence did not consistently predict disagreement in moral judgments of sacrificial dilemmas, Study 1: $B = 0.02$, $SE = 0.01$, $t(15220) = 2.02$, $p = .044$, CI [−0.01, 0.03]; Study 2: $B > -0.01$, $SE = 0.02$, $t(3073) = -0.003$, $p = .997$, CI [−0.03, 0.03].

Differential Endorsement of Utilitarian Principles Predicts Disagreement in Moral Decision Making

We next tested whether differential endorsement of instrumental harm and impartial beneficence predicted dissimilar patterns of moral

decision making in a task that does not directly probe these constructs. To this end, dyadic regression models were run to test whether participants who disagreed in their endorsement of instrumental harm and impartial beneficence in the abstract were more likely to also diverge on different dimensions of moral decision making: κ_{other} , κ_{self} , and relative κ ($\kappa_{\text{other}} - \kappa_{\text{self}}$). All models included trait psychopathy, empathy, self-control, and participants' beliefs about the receiver's pain tolerance as regressors of no interest.

Differential endorsement of instrumental harm did not consistently predict dissimilarity in any aspect of moral decision making across studies (Figure 3B). Dissimilarity in instrumental harm predicted dissimilarity in κ_{other} in Study 2 but not Study 1 (Figure 3B), Study 1: $B < 0.01$, $SE = 0.01$, $t(14990) = 0.18$, $p = .859$, $CI [-0.02, 0.02]$; Study 2: $B = 0.05$, $SE = 0.02$, $t(3070) = 2.79$, $p = .005$, $CI [0.02, 0.09]$. Dissimilarity in instrumental harm also predicted dissimilarity in κ_{self} in Study 2 but not Study 1, Study 1: $B < 0.01$, $SE = 0.01$, $t(15220) = 0.46$, $p = .648$, $CI [-0.01, 0.02]$; Study 2: $B = 0.07$, $SE = 0.02$, $t(3068) = 4.09$, $p < .001$, $CI [0.04, 0.10]$. There was no relationship between dissimilarity in instrumental harm and dissimilarity in relative κ in any study, Study 1: $B > -0.01$, $SE = 0.01$, $t(15110) = -0.93$, $p = .351$, $CI [-0.01, 0.01]$; Study 2: $B = -0.02$, $SE = 0.01$, $t(3052) = -1.08$, $p = .282$, $CI [-0.05, 0.01]$.

Turning to impartial beneficence, differential endorsement of this dimension consistently predicted dissimilarity in relative κ across studies (Figure 3B), Study 1: $B = 0.02$, $SE < 0.01$, $t(15100) = 4.60$, $p < .001$, $CI [0.01, 0.03]$; Study 2: $B = 0.08$, $SE = 0.01$, $t(3039) = 5.43$, $p < .001$, $CI [0.05, 0.10]$. However, there was no consistent relationship between dissimilarity in impartial beneficence and dissimilarity in κ_{other} or κ_{self} . Specifically, dissimilarity in impartial beneficence predicted dissimilarity in κ_{other} in Study 1 but not Study 2, Study 1: $B = 0.03$, $SE = 0.01$, $t(15100) = 3.08$, $p = .002$, $CI [0.01, 0.04]$; Study 2: $B = -0.01$, $SE = 0.02$, $t(3071) = -0.61$, $p = .544$, $CI [-0.05, 0.02]$, while the opposite pattern of results was observed for κ_{self} , Study 1: $B > -0.01$, $SE = 0.01$, $t(15210) = -0.19$, $p = .854$, $CI [-0.02, 0.01]$; Study 2: $B = 0.05$, $SE = 0.016$, $t(3056) = 3.23$, $p = .001$, $CI [0.01, 0.07]$.

Group Identification With EA Amplifies the Relationship Between Utilitarian Principles and Moral Cognition

Our behavioral data so far suggest that impartial beneficence and instrumental harm shape distinctive patterns of moral judgment and (to a lesser extent) decision making. As an exploratory analysis, we tested an additional prediction that socially identifying with these principles—in other words, considering them to be core to one's identity—could amplify the extent to which endorsement of those principles shapes moral behavior (Ellemers et al., 2014). We reasoned that if two individuals similarly *identify* with a community that emphasizes utilitarian principles (i.e., as an effective altruist), they should show a stronger correspondence between shared endorsement of those principles and similar patterns of moral judgment and behavior. By contrast, individuals who similarly endorse utilitarian *principles* but differ in *identification* with a group centered around those principles may be less aligned in how they frame and represent moral problems.

To test this, we leveraged a subset of our data from Study 1 collected from the EA community. Specifically, we predicted that dissimilarity in group identification with EA (operationalized as dissimilar answers to the question “My involvement with EA is an

important part of my identity”) would moderate the previously observed relationships between dissimilarity in utilitarian principles and dissimilarity in moral judgment and decision making.

We observed significant interactions between dissimilarity in EA identification and dissimilarity in the endorsement of utilitarian principles in dyadic regressions predicting divergence in moral judgments and decisions. Dissimilarity in EA identification moderated the relationship between divergence in instrumental harm endorsement and divergence in sacrificial dilemma responses, Figure 4A: $B = 0.02$, $SE = 0.01$, $t(15120) = 2.72$, $p = .007$, $CI [<0.01, 0.03]$; the relationship between dissimilarity in impartial beneficence endorsement and dissimilarity in impartiality dilemma responses, Figure 4B: $B = 0.02$, $SE = 0.01$, $t(15100) = 4.13$, $p < .001$, $CI [0.01, 0.04]$; and the relationship between dissimilarity in impartial beneficence endorsement and dissimilarity in κ_{other} , Figure 4C: $B = 0.02$, $SE = 0.01$, $t(15170) = 2.43$, $p = .015$, $CI [<0.01, 0.03]$.

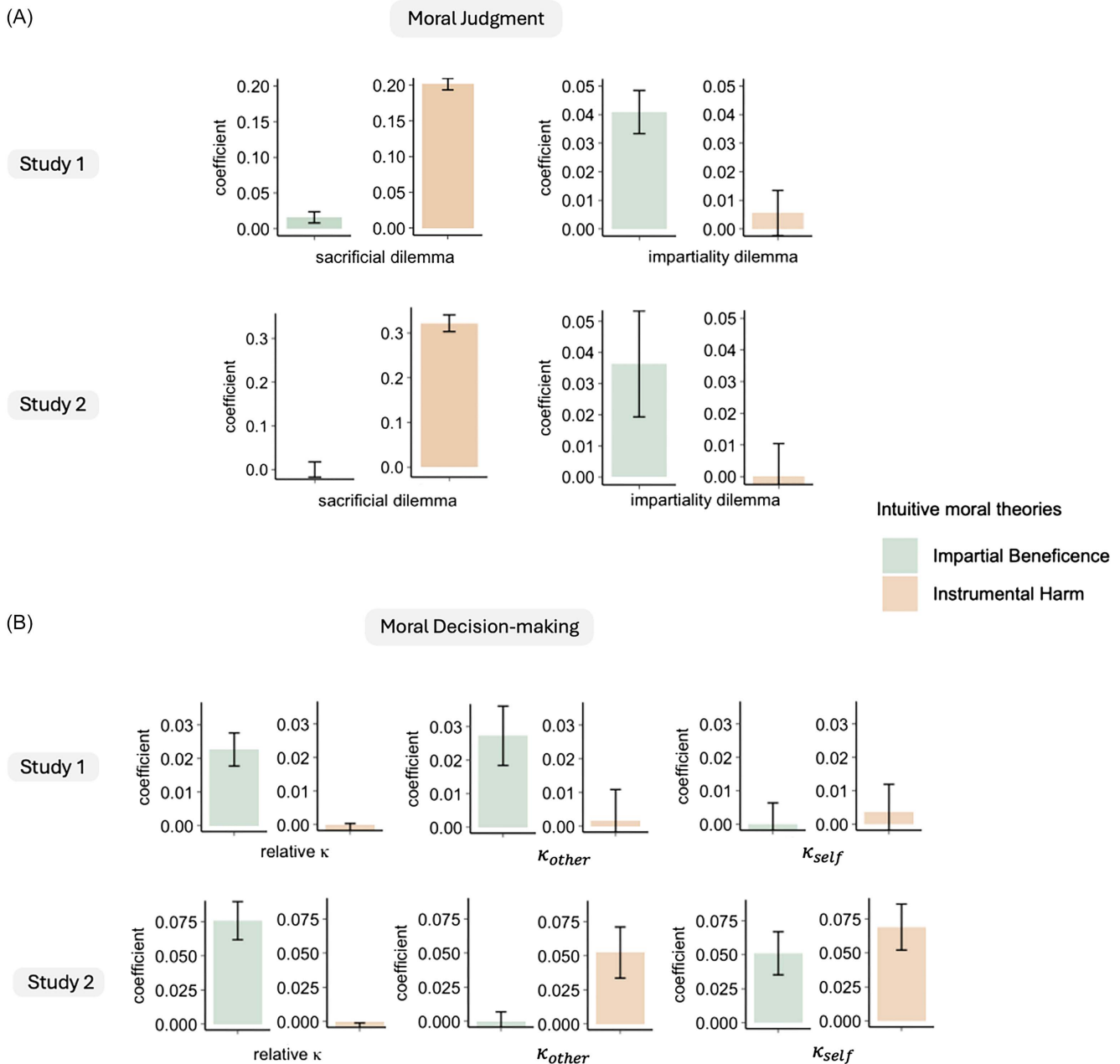
Divergent Endorsement of Utilitarian Principles Predicts Dissimilar Neural Correlates of Moral Decisions

We can see in our behavioral data that participants who differentially endorsed instrumental harm and impartial beneficence did not make dramatically different choices in the moral decision-making task (although there were some differences in directions that make sense given what we know about these dimensions of utilitarianism). We next asked whether differential endorsement of instrumental harm and impartial beneficence predicted dissimilar neural representations of moral decisions. This kind of analysis would not be possible in a task that directly probes instrumental harm and impartial beneficence, because in such a task we would not be able to dissociate endorsement of principles from the choices people make. That is, if instrumental harm and impartial beneficence had strongly predicted choice patterns in the moral decision-making task, neural patterns associated with endorsement of those principles would be confounded with neural patterns associated with choices themselves. However, because we did not observe a very strong relationship between endorsement of the principles and choices in the task, we can probe how differential endorsement of the principles predicts dissimilar neural representations of the moral decision parameters, over and above the decisions people actually make. To this end, we examined multivariate neural activity patterns by running dyadic regression models across voxels, where for every pair of participants in our data, neural distance in encoding pain and profit was respectively regressed onto dissimilarity in instrumental harm and impartial beneficence, controlling for dissimilarity in κ_{other} (Figure 5).

We implemented IS-RSA with dyadic regression approach for neuroimaging data. (A) First, distances between every possible pair of individuals in the variables of interest (β estimates, endorsement of impartial beneficence and instrumental harm, as well as other controlling variables) were obtained. (B) We ran a GLM and obtained the β estimates of parametric BOLD responses to consequences in our moral decision-making task (pain and profit). Next, with those distance measures, we ran a dyadic regression model for each brain voxel, where distance in endorsement of utilitarian principles predicted the neural distance in parametric responses to pain and profit, while controlling for the distance in choice preference expressed as harm aversion parameter, distance in trait empathy, trait psychopathy, and trait self-control.

Figure 3

Dissimilarity in Endorsement of Utilitarian Principles Predicting Dissimilarity in Moral Judgment and Decision Making

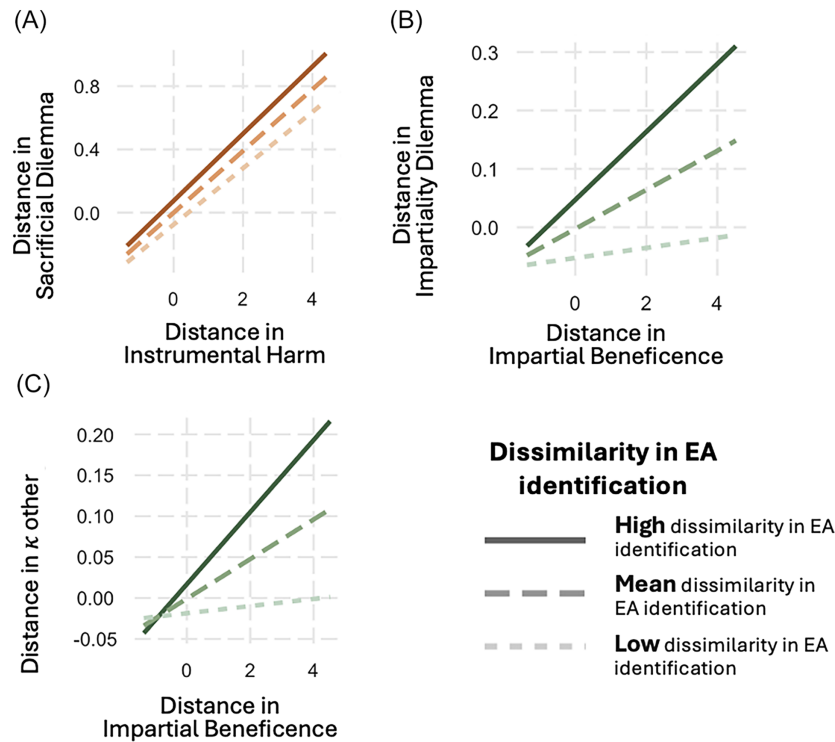


Note. Panel A (left): Dissimilarity in instrumental harm consistently predicted dissimilarity in responses to sacrificial dilemmas but not impartiality dilemmas across studies. Panel A (right): Dissimilarity in impartial beneficence consistently predicted dissimilarity in responses to impartiality dilemmas but not sacrificial dilemmas across studies. Panel B (left): Dissimilarity in instrumental harm did not consistently predict dissimilarity in moral decision making. Panel B (right): Dissimilarity in impartial beneficence consistently predicted dissimilarity in relative κ (relative orientation as cosine dissimilarity between κ_{other} and κ_{self}) across studies. See the online article for the color version of this figure.

We first examined how endorsement of utilitarian principles related to neural representations of pain during moral decision making. Differential endorsement of instrumental harm was associated with dissimilar neural representations of pain in the caudate and insula (Figure 6A; Supplemental Table S5). Meanwhile, differential

endorsement of impartial beneficence was associated with dissimilar neural representations of pain in the middle temporal gyrus, dorsal medial frontal gyrus, and precuneus, ventromedial prefrontal cortex /orbitofrontal cortex, dorsomedial prefrontal cortex, anterior cingulate cortex, amygdala, inferior frontal gyrus, and

Figure 4
Group Identification With Effective Altruism Amplifying the Relationship Between Utilitarian Principles and Moral Cognition



Note. Dissimilarity in identification with the EA community moderated the relationship between (Panel A) dissimilarity in instrumental harm and dissimilarity in sacrificial dilemmas; (Panel B) dissimilarity in impartial beneficence and dissimilarity in impartiality dilemmas; and (Panel C) dissimilarity in impartial beneficence and dissimilarity in κ_{other} . EA = Effective Altruism. See the online article for the color version of this figure.

caudate (Figure 6A; Supplemental Table S3). These sets of regions were largely nonoverlapping.

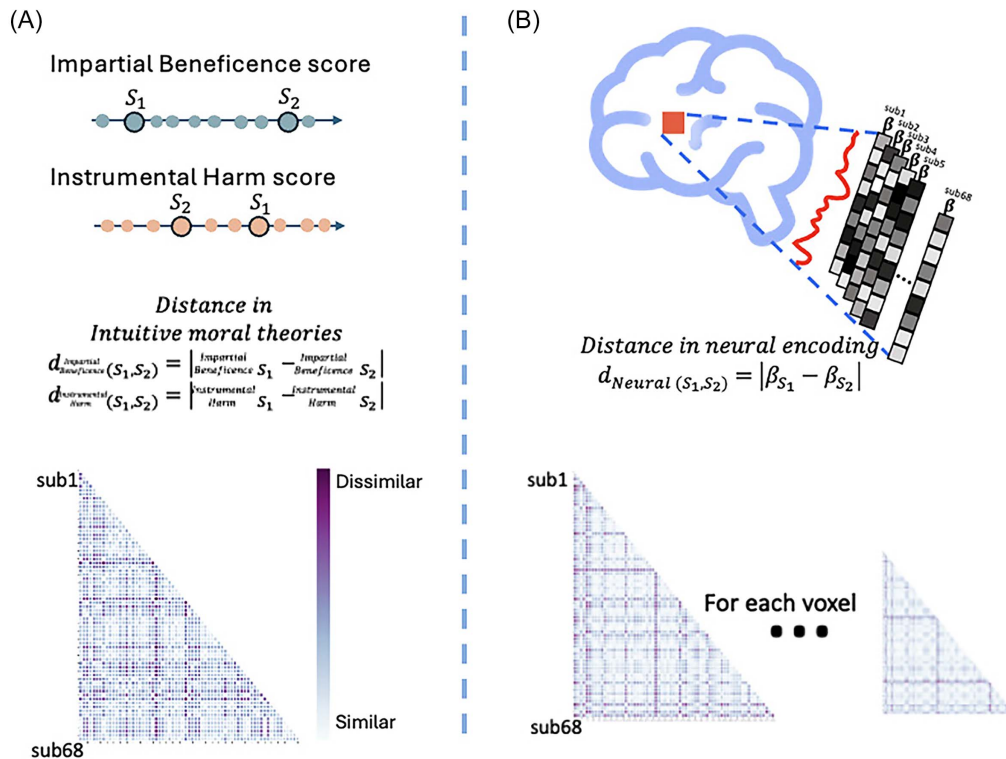
Next, we examined neural representations of profit. Differential endorsement of instrumental harm was associated with dissimilar neural representations of profit in the middle temporal gyrus, orbitofrontal cortex, precuneus, inferior frontal gyrus, thalamus, and precuneus (Figure 6B; Supplemental Table S4). Meanwhile, differential endorsement of impartial beneficence was associated with dissimilar neural representations of profit in the insula, thalamus, and cingulate gyrus (Figure 6B; Supplemental Table S2). Again, these sets of regions were largely nonoverlapping.

In summary, we found that even when pairs of participants make similar choices in the moral decision-making task, their brains represent those choices differently if they differentially endorse instrumental harm and impartial beneficence. Conversely, even when pairs of participants make different choices in the task, their brains represent those choices similarly if they similarly endorse instrumental harm and impartial beneficence. These two dimensions of utilitarianism predicted neural dissimilarity patterns during moral decision making in largely nonoverlapping brain networks, suggesting they make distinctive contributions to framing moral decision problems.

Discussion

In the present studies, we examined how two dimensions of utilitarian psychology—instrumental harm and impartial beneficence—predict distinctive patterns of moral judgment and decision making. We approached this question by considering instrumental harm and impartial beneficence as distinctive intuitive moral theories (Crockett et al., 2024; Dweck et al., 1995; Gerstenberg & Tenenbaum, 2017; Gottlieb & Lombrozo, 2018; Keil, 2024; Mahr & Csibra, 2021). We predicted that similar endorsement of instrumental harm or impartial beneficence should predict shared processing of moral situations, while divergent endorsement should predict distinctive representations of those situations. IS-RSA, applied to both behavioral and neural data, revealed that (a) divergent endorsement of instrumental harm and impartial beneficence predicted dissimilar patterns of moral judgment and, to a lesser extent, moral decision making and (b) divergent endorsement of instrumental harm and impartial beneficence predicted dissimilar multivariate neural responses in largely nonoverlapping brain networks. These findings lend further support to the two-dimensional model of utilitarianism, demonstrating that utilitarian inclinations do not represent a unitary psychological or neural profile (Everett & Kahane, 2020; Kahane et al., 2018). They also suggest these two dimensions in utilitarian tendencies operate as

Figure 5
Schematic Representation of Intersubject Representational Similarity Analysis Method



Note. The brain shape icon in was downloaded from the Noun Project. It is covered under a royalty-free license through subscription plan. sub = subject. See the online article for the color version of this figure.

distinct intuitive theories, each uniquely shaping patterns of moral cognition and behavior.

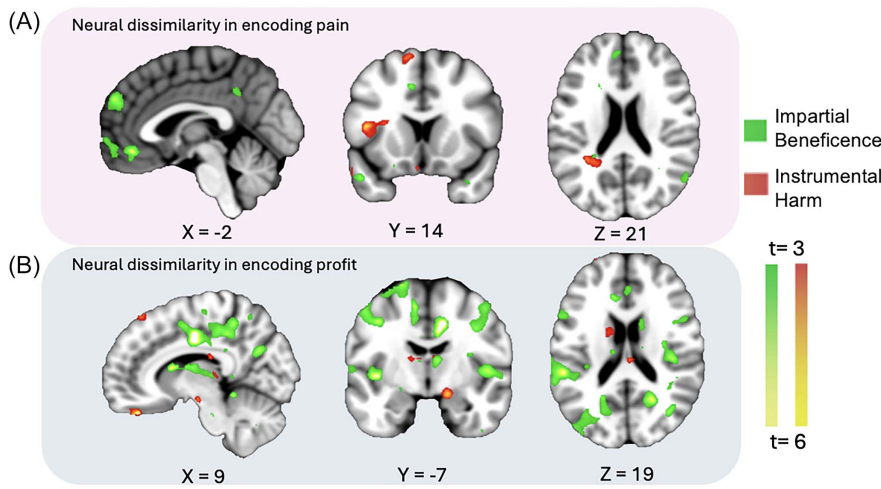
All our findings accounted for other sources of individual differences, including trait empathy, trait psychopathy, and trait self-control, by controlling for these variables within the same model—traits that previous research has linked to utilitarian tendencies and broader patterns of moral judgment (Cameron et al., 2022; Gleichgerrcht & Young, 2013; Glenn et al., 2010; Hofmann et al., 2018; Koenigs et al., 2012; Patil & Silani, 2014). The divergent patterns explained by the two dimensions of utilitarian tendencies remained robust even after accounting for these factors. This provides further evidence that the patterns we identified are not merely a byproduct of dispositional traits we controlled but are instead shaped by the intuitive moral theories individuals endorse.

However, we acknowledge an alternative interpretation of our findings: that endorsement of impartial beneficence and instrumental harm may reflect basic personality traits or preferences and that the observed neural dissimilarity could be driven by such individual differences rather than the application of shared moral principles functioning as an intuitive theory. This view is consistent with the primary goal of OUS, which aims to capture stable individual differences in moral orientation rather than the cognitive processes involved in moral decisions (Kahane et al., 2018). From this perspective, the neural dissimilarity patterns we observe might reflect trait-level tendencies rather than structured belief-generating systems in moral cognition.

While our findings do not conclusively dissociate between whether impartial beneficence and instrumental harm reflect simple trait-level dispositions or a more structured representation of moral beliefs, we do not view these interpretations as mutually exclusive. It is plausible that the endorsement of a coherent moral framework gives rise to both trait-level scores on instruments like the OUS and structured patterns of neural activity. While we do not claim to isolate pro-utilitarian cognitive system per se or to rule out trait-based explanations, the fact that neural similarity emerges among individuals with similar levels of impartial beneficence and instrumental harm endorsement—even in contexts where these principles are not explicitly engaged—may suggest the presence of structured neural representations aligned with each dimension of utilitarian tendencies. Future work could help disentangle the contributions of stable dispositions versus structured cognitive systems to the shared neural patterns.

Our unique sample recruited from the EA community allowed us to explore the possibility that socially identifying with utilitarian principles amplifies the extent to which impartial beneficence and instrumental harm frame moral judgment and decision making. These results present an intriguing opportunity to investigate the previously unexamined connection between social identity and intuitive moral theories. Previous research on moral identity has primarily focused on the extent to which individuals care about broad moral evaluations of their behavior (Aquino & Reed, 2002; Boegershausen et al., 2015;

Figure 6
Divergent Endorsement of Utilitarian Principles Predicting Dissimilar Neural Correlates of Moral Decisions



Note. Panel A: Dissimilarity in instrumental harm and impartial beneficence predicts dissimilar neural representations of pain in distinct brain networks. Panel B: Dissimilarity in instrumental harm and impartial beneficence predicts dissimilar neural representations of profit in distinct brain networks. See the online article for the color version of this figure.

Reynolds & Ceranic, 2007; Skitka et al., 2021). In contrast, our finding expands the investigation of the moral identity by exploring its specific connection with the particular moral principles an individual endorses. For future studies, it would be valuable to further explore the extent to which individuals integrate these theories into their sense of identity, as well as the mechanisms by which this integration regulates the degree to which these theories shape their understanding and interpretation of the moral world.

Our work builds on prior research using multivariate analysis of neural data, which demonstrated that neural responses of conservatives and liberals diverge when viewing the same political content, reflecting different interpretations (de Bruin et al., 2023; Leong et al., 2020). Using a similar approach, we found that differential endorsement of utilitarian principles predicts dissimilar neural representations of moral decisions—even when individuals make the same choices. Multivariate analyses capture a broader range of underlying effects, including both distributed neural patterns and mean differences in signal intensity. While there is ambiguity in interpreting such multivariate similarity—specifically, whether it reflects univariate differences, distributed patterns, or a combination of both—we highlight that our approach addresses a theoretically meaningful and broader question that univariate analysis alone cannot resolve: whether individuals with similar moral endorsements exhibit more similar neural response patterns, regardless of whether there are differences in the direction or magnitude of regional activation.

However, we note that the observed patterns of neural dissimilarity may arise from multiple levels of cognitive processing. At a lower level, they could reflect perceptual or attentional tracking of task features. Alternatively, they may reflect higher order processes such as evaluating moral consequences or affective responses like guilt or moral conflict. Prior neuroimaging studies support both possibilities, showing that neural dissimilarity can reflect both shared processing of

low-level stimulus features and abstract representations/integrative processing (Dmochowski et al., 2014; Hasson et al., 2008; Lahnakoski et al., 2014; Nummenmaa et al., 2012; Song et al., 2021). While we posit that intuitive theories—with their structured cognitive architecture—are likely to shape not just the low-level processing but ultimately more higher level representations of moral decision problems, our current analyses do not directly distinguish between these possibilities.

In line with the possibility of higher level processing, we consider that the dissimilar neural representations associated with individuals who subscribe to different intuitive moral theories may arise from the distinct meanings they assign to moral decisions. For example, individuals who strongly endorse impartial beneficence may view inflicting harm on others for personal gain as inherently immoral, given their commitment to equal concern for all individuals. By contrast, those who strongly endorse instrumental harm may regard such harm as a justifiable means to an end if the monetary gain could serve other purposes. In this way, we speculate that it is the subjective meaning attributed to moral choices, shaped by one's intuitive moral theories, that gives rise to divergent patterns of neural representation across individuals. However, future research will be needed to clarify the specific cognitive processes through which shared intuitive theories give rise to the observed neural dissimilarity.

Also, future work might explore whether, for other moral principles—such as those emphasizing the action itself (e.g., deontology; Fried, 1978; Kant, 1797/2002; Ross, 2002) or agreement among individuals (e.g., contractualism; Parfit, 1984; Rawls, 1971; Scanlon, 1998)—moral disagreement similarly arises from varying degrees of endorsement and whether such disagreement is systematically organized according to the specific ethical emphasis of each principle. If this is the case, it would provide deeper insights into the nature of moral disagreement, helping to clarify how differing

commitments to distinct ethical theories collectively shape patterns of divergence in moral cognition.

Constraints on Generality

We acknowledge that the artificial nature of the moral decision-making task we used in our study imposes limitations on the generalizability of our findings. Specifically, the decision to trade off one's monetary profit against another's pain represents a moral scenario that individuals do not commonly encounter in everyday life. Therefore, further investigation is required to determine the extent to which these findings can be generalized beyond the laboratory setting.

Another important limitation on the generalizability of our findings is that our data were collected in 2017–2018 from convenience samples in the United States and United Kingdom, with a majority of participants being White, young, and highly educated (Rad et al., 2018). Moral principles vary significantly across cultures, with different values and levels of importance assigned to these principles in diverse cultural contexts (Chiu et al., 1997; Shweder et al., 1987, 1997). Therefore, it is possible that our findings will not generalize beyond these populations. We hope that the methods and approach employed and developed in our work open avenues for understanding how diverse moral principles, which may vary across cultural contexts, are represented and function as intuitive theories shaping moral cognition.

References

- Amorino, P., Ploe, M. L., & Marsh, A. A. (2022). Moral foundations, values, and judgments in extraordinary altruists. *Scientific Reports*, *12*(1), Article 22111. <https://doi.org/10.1038/s41598-022-26418-1>
- Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. <https://doi.org/10.1037/0022-3514.83.6.1423>
- Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381–417. <https://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*(1), 154–161. <https://doi.org/10.1016/j.cognition.2011.05.010>
- Bentham, J. (1983). *The collected works of Jeremy Bentham: Deontology together with a table of the springs of action and article on utilitarianism*. Oxford University Press. <https://doi.org/10.1093/actrade/9780198226093.book.1>
- Boegershausen, J., Aquino, K., & Reed, I. I. A., II. (2015). Moral identity. *Current Opinion in Psychology*, *6*, 162–166. <https://doi.org/10.1016/j.copsyc.2015.07.017>
- Brethel-Haurwitz, K. M., Cardinale, E. M., Vekaria, K. M., Robertson, E. L., Walitt, B., VanMeter, J. W., & Marsh, A. A. (2018). Extraordinary altruists exhibit enhanced self–other overlap in neural responses to distress. *Psychological Science*, *29*(10), 1631–1641. <https://doi.org/10.1177/0956797618779590>
- Cameron, C. D., Conway, P., & Scheffer, J. A. (2022). Empathy regulation, prosociality, and moral judgment. *Current Opinion in Psychology*, *44*, 188–195. <https://doi.org/10.1016/j.copsyc.2021.09.011>
- Capraro, V., Everett, J. A., & Earp, B. D. (2019). Priming intuition disfavors instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology*, *83*, 142–149. <https://doi.org/10.1016/j.jesp.2019.04.006>
- Chen, P. A., Jolly, E., Cheong, J. H., & Chang, L. J. (2020). Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *NeuroImage*, *216*, Article 116851. <https://doi.org/10.1016/j.neuroimage.2020.116851>
- Chen, P. A., & Qu, Y. (2021). Taking a computational cultural neuroscience approach to study parent–child similarities in diverse cultural contexts. *Frontiers in Human Neuroscience*, *15*, Article 703999. <https://doi.org/10.3389/fnhum.2021.703999>
- Chiu, C.-Y., Dweck, C. S., Tong, J. Y.-Y., & Fu, J. H.-Y. (1997). Implicit theories and conceptions of morality. *Journal of Personality and Social Psychology*, *73*(5), 923–940. <https://doi.org/10.1037/0022-3514.73.5.923>
- Contreras-Huerta, L. S., Lockwood, P. L., Bird, G., Apps, M. A. J., & Crockett, M. J. (2022). Prosocial behavior is associated with transdiagnostic markers of affective sensitivity in multiple domains. *Emotion*, *22*(5), 820–835. <https://doi.org/10.1037/emo0000813>
- Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216–235. <https://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. <https://doi.org/10.1016/j.cognition.2018.04.018>
- Crockett, M. J., Kim, J. S., & Shin, Y. S. (2024). Intuitive theories and the cultural evolution of morality. *Current Directions in Psychological Science*, *33*(4), 211–219. <https://doi.org/10.1177/09637214241245412>
- Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(48), 17320–17325. <https://doi.org/10.1073/pnas.1408988111>
- Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. <https://doi.org/10.1038/nn.4557>
- de Bruin, D., van Baar, J. M., Rodríguez, P. L., & FeldmanHall, O. (2023). Shared neural representations and temporal segmentation of political content predict ideological similarity. *Science Advances*, *9*(5), Article eabq5920. <https://doi.org/10.1126/sciadv.abq5920>
- de Lazari-Radek, K., & Singer, P. (2017). *Utilitarianism: A very short introduction*. Oxford University Press. <https://doi.org/10.1093/actrade/9780198728795.001.0001>
- Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., & Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications*, *5*(1), 4567. <https://doi.org/10.1038/ncomms5567>
- Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, *6*(4), 267–285. https://doi.org/10.1207/s15327965pli0604_1
- Earp, B. D., McLoughlin, K., Caraccio, M., Calcott, R., Rottman, J., Clark, M. S., & Crockett, M. (2024). *Impartial beneficence predicts greater and more uniform concern for others across social relationships*. <https://doi.org/10.31234/osf.io/jazbn>
- Effective Altruism. (n.d.). *Effective Altruism*. <https://www.effectivealtruism.org/>
- Ellemers, N., Pagliaro, S., & Barreto, M. (2014). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, *24*(1), 160–193. <https://doi.org/10.1080/10463283.2013.841490>
- Everett, J. A. C., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences*, *24*(2), 124–134. <https://doi.org/10.1016/j.tics.2019.11.012>
- Finn, E. S., Glerean, E., Khojandi, A. Y., Nielson, D., Molfese, P. J., Handwerker, D. A., & Bandettini, P. A. (2020). Idiosyncrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, *215*, Article 116828. <https://doi.org/10.1016/j.neuroimage.2020.116828>

- Fowler, Z., Law, K. F., & Gaesser, B. (2021). Against empathy bias: The moral value of equitable empathy. *Psychological Science*, 32(5), 766–779. <https://doi.org/10.1177/0956797620979965>
- Fried, C. (1978). *Right and wrong*. Harvard University Press. <https://doi.org/10.4159/harvard.9780674332508>
- Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.
- Gleichgericht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLOS ONE*, 8(4), Article e60418. <https://doi.org/10.1371/journal.pone.0060418>
- Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. *Judgment and Decision Making*, 5(7), 497–505. <https://doi.org/10.1017/S1930297500001662>
- Gottlieb, S., & Lombrozo, T. (2018). Folk theories in the moral domain. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 320–331). Guilford Press.
- Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144–1154. <https://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(2), 389–400. <https://doi.org/10.1016/j.neuron.2004.09.027>
- Greene, J. D. (2009). The cognitive neuroscience of moral judgment. In M. S. Gazzaniga, E. Bizzi, L. M. Chalupa, S. T. Grafton, T. F. Heatherton, C. Koch, J. E. LeDoux, S. J. Luck, G. R. Mangun, J. A. Movshon, H. Neville, E. A. Phelps, P. Rakic, D. L. Schacter, M. Sur, & B. A. Wandell (Eds.), *The cognitive neurosciences* (4th ed., pp. 987–999). Massachusetts Institute of Technology.
- Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron*, 57(3), 452–462. <https://doi.org/10.1016/j.neuron.2007.12.009>
- Hofmann, W., Meindl, P., Mooijman, M., & Graham, J. (2018). Morality and self-control: How they are intertwined and where they differ. *Current Directions in Psychological Science*, 27(4), 286–291. <https://doi.org/10.1177/0963721418759317>
- 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(2), 131–164. <https://doi.org/10.1037/rev0000093>
- Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). “Utilitarian” judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, 134, 193–209. <https://doi.org/10.1016/j.cognition.2014.10.005>
- Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*, 7(4), 393–402. <https://doi.org/10.1093/scan/nsr005>
- Kant, I. (2002). *Groundwork for the metaphysics of morals*. Yale University Press. (Original work published 1797)
- Keil, F. (2024). Intuitive theories. In M. C. Frank & A. Majid (Eds.), *Open encyclopedia of cognitive science*. MIT Press. <https://doi.org/10.21428/e2759450.9666c9f2>
- Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, 7(6), 708–714. <https://doi.org/10.1093/scan/nsr048>
- Kuo, Y. A., Chou, F. B., & Chen, P. A. (2023). *Social isolation strengthens the association between intersubject similarity in fantasy and similarity in affective appraisal*. <https://doi.org/10.31234/osf.io/64bjc>
- Lahnakoski, J. M., Glerean, E., Jääskeläinen, I. P., Hyönä, J., Hari, R., Sams, M., & Nummenmaa, L. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, 100, 316–324. <https://doi.org/10.1016/j.neuroimage.2014.06.022>
- Leong, Y. C., Chen, J., Willer, R., & Zaki, J. (2020). Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences of the United States of America*, 117(44), 27731–27739. <https://doi.org/10.1073/pnas.2008530117>
- Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, 33(2), 273–286. <https://doi.org/10.1111/j.1551-6709.2009.01013.x>
- MacAskill, W. (2019). The definition of effective altruism. In H. Greaves & T. Pummer (Eds.), *Effective altruism: Philosophical issues* (pp. 10–28). Oxford University Press.
- Mahr, J. B., & Csibra, G. (2021). A short history of theories of intuitive theories. In J. Gervain, G. Csibra, & K. Kovács (Eds.), *A life in cognition: Studies in cognitive science in honor of Csaba Pléh* (pp. 219–232). Springer International Publishing.
- Mill, J. S. (1969). Utilitarianism. In J. M. Robson (Ed.), *Essays on ethics, religion, and society. Vol. 10 of Collected Works of John Stuart Mill* (pp. 203–259). University of Toronto Press.
- 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(10), 707–718. <https://doi.org/10.1111/spc3.12066>
- Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, 14(6), 667–685. <https://doi.org/10.1093/scan/nsz037>
- Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24), 9599–9604. <https://doi.org/10.1073/pnas.1206095109>
- Parfit, D. (1984). *Reasons and persons*. Oxford University Press.
- Paruzel-Czachura, M., & Charzyńska, E. (2022). Investigating the relationship between centrality of religiosity, instrumental harm, and impartial beneficence through the lens of moral foundations. *Religions*, 13(12), Article 1215. <https://doi.org/10.3390/rel13121215>
- Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology*, 5, Article 501. <https://doi.org/10.3389/fpsyg.2014.00501>
- Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*, 120(2), 443–460. <https://doi.org/10.1037/pspp0000281>
- Paulhus, D. L., Neumann, C. S., & Hare, R. D. (2016). *Manual for the Hare Self-Report Psychopathy Scale*. Multi-Health Systems.
- Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of *Homo sapiens*: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 115(45), 11401–11405. <https://doi.org/10.1073/pnas.1721165115>
- Rawls, J. (1971). *A theory of justice*. Belknap Press of Harvard University Press. <https://doi.org/10.4159/9780674042605>
- Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, 93(1), 84–95. <https://doi.org/10.1080/00223891.2010.528484>
- Reynolds, S. J., & Ceranic, T. L. (2007). The effects of moral judgment and moral identity on moral behavior: An empirical examination of the moral individual. *Journal of Applied Psychology*, 92(6), 1610–1624. <https://doi.org/10.1037/0021-9010.92.6.1610>
- Rhoads, S. A., O’Connell, K., Berluti, K., Ploe, M. L., Elizabeth, H. S., Amormino, P., Li, J. L., Dutton, M. A., VanMeter, A. S., & Marsh, A. A.

- (2023). Neural responses underlying extraordinary altruists' generosity for socially distant others. *PNAS Nexus*, 2(7), Article pgsad199. <https://doi.org/10.1093/pnasnexus/pgad199>
- Ross, W. D. (2002). *The right and the good*. Oxford University Press. <https://doi.org/10.1093/0199252653.001.0001>
- Scanlon, T. M. (1998). *What we owe to each other* (Vol. 66). Belknap Press of Harvard University Press.
- Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 1–83). University of Chicago Press.
- Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The “Big Three” of morality (autonomy, community, and divinity), and the “Big Three” explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). Routledge.
- Sidgwick, H. (1981). *The methods of ethics* (7th ed.). Hackett Publishing Company.
- Singer, P. (2011). *Practical ethics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511975950>
- Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. *Annual Review of Psychology*, 72(1), 347–366. <https://doi.org/10.1146/annurev-psych-063020-030612>
- Song, H., Finn, E. S., & Rosenberg, M. D. (2021). Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proceedings of the National Academy of Sciences*, 118(33), e2021905118. <https://doi.org/10.1073/pnas.2021905118>
- Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America*, 107(32), 14425–14430. <https://doi.org/10.1073/pnas.1008662107>
- Syropoulos, S., Law, K. F., Kraft-Todd, G., & Young, L. (2023). *Impartial intergenerational beneficence: The psychology of feeling equal concern for all future generations*. PsyArXiv. <https://doi.org/10.31234/osf.io/e34kv>
- Takamatsu, R. (2018). Turning off the empathy switch: Lower empathic concern for the victim leads to utilitarian choices of action. *PLOS ONE*, 13(9), Article e0203826. <https://doi.org/10.1371/journal.pone.0203826>
- Tangney, J. P., Boone, A. L., & Baumeister, R. F. (2018). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In R. F. Baumeister (Ed.), *Self-regulation and self-control* (pp. 173–212). Routledge.
- 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, Article 229. <https://doi.org/10.3389/fnhum.2013.00229>
- Tuen, Y. J., Bulley, A., Palombo, D. J., & O'Connor, B. B. (2023). Social value at a distance: Higher identification with all of humanity is associated with reduced social discounting. *Cognition*, 230, Article 105283. <https://doi.org/10.1016/j.cognition.2022.105283>
- van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, 10(1), Article 1483. <https://doi.org/10.1038/s41467-019-09161-6>
- van Baar, J. M., Halpern, D. J., & FeldmanHall, O. (2021). Intolerance of uncertainty modulates brain-to-brain synchrony during politically polarized perception. *Proceedings of the National Academy of Sciences of the United States of America*, 118(20), Article e2022491118. <https://doi.org/10.1073/pnas.2022491118>
- Van Lange, P. A. M., De Bruin, E. M. N., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, 73(4), 733–746. <https://doi.org/10.1037/0022-3514.73.4.733>
- Vekaria, K. M., Brethel-Haurwitz, K. M., Cardinale, E. M., Stoycos, S. A., & Marsh, A. A. (2017). Social discounting and distance perceptions in costly altruism. *Nature Human Behaviour*, 1(5), Article 0100. <https://doi.org/10.1038/s41562-017-0100>
- Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., & Hasson, U. (2017). Same story, different story: The neural representation of interpretive frameworks. *Psychological Science*, 28(3), 307–319. <https://doi.org/10.1177/0956797616682029>
- Yu, H., Contreras-Huerta, L. S., Prosser, A. M. B., Apps, M. A. J., Hofmann, W., Sinnott-Armstrong, W., & Crockett, M. J. (2022). Neural and cognitive signatures of guilt predict hypocritical blame. *Psychological Science*, 33(11), 1909–1927. <https://doi.org/10.1177/09567976221122765>

Received March 15, 2024

Revision received May 27, 2025

Accepted June 19, 2025 ■