

## Opinion

## Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology

Jim A.C. Everett <sup>1,2,4,\*</sup> and Guy Kahane<sup>2,3,4</sup>

**Sacrificial moral dilemmas are widely used to investigate when, how, and why people make judgments that are consistent with utilitarianism. However, to what extent can responses to sacrificial dilemmas shed light on utilitarian decision making? We consider two key questions. First, how meaningful is the relationship between responses to sacrificial dilemmas, and what is distinctive about a utilitarian approach to morality? Second, to what extent do findings about sacrificial dilemmas generalize to other moral contexts where there is tension between utilitarianism and common-sense intuitions? We argue that sacrificial dilemmas only capture one point of conflict between utilitarianism and common-sense morality, and new paradigms will be necessary to investigate other key aspects of utilitarianism, such as its radical impartiality.**

**Utilitarianism, Trolley Dilemmas, and Moral Psychology**

Moral philosophers aim to develop systematic normative theories of right and wrong. **Utilitarianism** (see Glossary) is a famous if controversial example of such a theory, and posits that the whole of morality can be deduced from a single general principle: always act in a way that impartially maximizes aggregate well-being [1–5]. Outside the philosophy seminar room, however, most people typically make moral judgments not by applying a theory or explicit principles but by following highly specific norms and intuitions (e.g., [6]). Philosophers often term this pre-philosophical sensibility **common-sense morality** (CSM), and a great deal of research into moral psychology involves mapping out CSM: uncovering its structure, psychological underpinning, and its developmental, social, and evolutionary origins. This research program has revealed that, in many moral contexts, most people reject choices that maximize utility if doing so violates certain moral rules or is perceived as compromising 'sacred' values. For example, people typically give greater moral weight to acts over omissions [7–9], depart from utilitarian analysis in charitable giving [10], and regard punishment as being deserved independently of any consequentialist deterrent effect [11]. In this body of research, it is usually assumed that such rejections of a utility-maximizing goal are driven by different cognitive biases in different contexts, and that these rejections and their corresponding biases therefore need to be studied piecemeal (e.g., [12–14]).

At the turn of the new millennium a different approach emerged that largely focuses on responses to so-called sacrificial dilemmas such as the trolley scenarios that were first introduced by philosophers [15,16]. In these scenarios, participants are asked whether it is morally acceptable to sacrifice one or more individuals to save the lives of a greater number. Whereas utilitarianism tells us to always save the greatest number, a large majority of people reject this **pro-sacrificial** choice in scenarios involving directly harming the victim, for example pushing someone off a footbridge to block a runaway train (note that throughout we use 'pro-sacrificial' as a purely descriptive label that simply refers to approving the sacrifice of some to save a greater number). The sacrificial dilemmas paradigm has since come to dominate the study of utilitarian and non-utilitarian (or 'deontological') approaches to moral decision-making (e.g., [17–25]), and, indeed,

## Highlights

Nearly two decades of research have used sacrificial dilemmas to shed light on utilitarian decision making.

This paradigm has conceptual and empirical limitations which suggest that it is a mistake to treat utilitarian decision making as a single, unitary psychological phenomenon.

There are at least two key ways that utilitarianism departs from common-sense morality, and according to the 2D model these are distinct not only conceptually but also psychologically, and thus have different psychological correlates and likely different underlying psychological processes.

A wider range of ways in which people can depart from common-sense morality in a utilitarian direction needs to be studied. Doing so will help researchers come to a more complete picture of both common-sense and utilitarian moral thinking.

<sup>1</sup>School of Psychology, University of Kent, Canterbury CT2 7NP, UK

<sup>2</sup>Uehiro Centre for Practical Ethics, University of Oxford, Oxford OX1 1PT, UK

<sup>3</sup>Faculty of Philosophy, University of Oxford, Oxford OX2 6GG, UK

<sup>4</sup>Equal contributions

\*Correspondence: [j.a.c.everett@kent.ac.uk](mailto:j.a.c.everett@kent.ac.uk) (J.A.C. Everett).

has become a 'standard methodology for research on moral judgment' [26]. Although some critics have highlighted the highly artificial character of sacrificial dilemmas [27], sacrificial dilemmas mirror difficult decisions that can arise in military and medical contexts. They are therefore a powerful tool for studying the cognitive and neural mechanisms underlying judgments about what we term **instrumental harm** (IH) – the moral permissibility of harming some for a greater good.

However, part of the reason the sacrificial dilemmas paradigm has been so influential is because it is taken to teach us general lessons about moral psychology. A prominent example is the **dual-process model** (DPM) of moral psychology (e.g., [18,19,21,28]). According to the DPM, a refusal to sacrifice individuals for the greater good, and thus to maximize utility, is based in immediate intuition and emotional gut-reactions. By contrast, the DPM claims that, when people make pro-sacrificial choices – often called utilitarian judgments, they employ deliberative processing to repress such intuitive aversion to harming, allowing them to resolve the dilemma by using utilitarian cost–benefit analysis. In its most ambitious form, the DPM draws on the unusual context of sacrificial dilemmas to make general claims about two opposing modes of moral decision making that echo explicit philosophical theories, suggesting that "the terms 'deontology' and 'consequentialism' refer to psychological natural kinds", and are 'philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking' [28]. Correspondingly, it has been argued that the DPM sheds light on the psychological sources of utilitarian ethics, and even supports it as a normative view (e.g., [18,19,21,28,64]). Importantly, however, even researchers who do not operate within the DPM framework routinely present sacrificial dilemmas research as capturing the core contrast between utilitarian and non-utilitarian ethical approaches, and often make seemingly general claims about the psychological factors and processes that drive 'utilitarian judgment' [31,32], and attribute utilitarian tendencies to individuals [24,33] and specific populations [29,30].

### Using Sacrificial Dilemmas to Understand Utilitarian Moral Psychology

Our aim here is to clarify the relationship between the sacrificial dilemma paradigm, utilitarian ethical theory, and lay moral psychology. This relationship has two aspects. The first, which we shall largely focus on, is whether utilitarianism – a normative ethical theory – provides a fruitful framework for interpreting the responses of lay persons to sacrificial dilemmas and, indeed, to other moral contexts. The second is whether empirical research using sacrificial dilemmas can shed light on the 'cognitive building blocks of utilitarian philosophy' [5] and even, as a potential further step, support (or undermine) normative ethical theories. The two are related: if responses to sacrificial dilemmas and the processes underlying them bear a sufficiently meaningful connection to utilitarianism, then it is more likely that these processes are a psychological source of this normative theory. However, this will require that the answer to the first question is sufficiently substantive: if by 'utilitarian' we mean something generic, non-distinctive, and only loosely linked to the ethical theory, then the psychological processes one identifies are unlikely to tell us much about utilitarianism.

In what follows, we review the debate about the relationship between pro-sacrificial judgments and utilitarianism. We highlight important conceptual and methodological advances that clarify different senses in which pro-sacrificial judgment might be usefully termed 'utilitarian'. However, we will also argue that sacrificial dilemmas can shed light only on some ways in which lay moral psychology echoes utilitarianism. What is needed is a multidimensional approach (e.g., [34]) that can incorporate insights from research into sacrificial dilemmas while also directing attention to hitherto neglected ways in which utilitarianism can inform the study of lay moral psychology.

### Glossary

**Common-sense morality (CSM):** a term that moral philosophers use to describe the pre-philosophical moral intuitions that humans typically share – what psychologists might refer to as 'lay morality'. Most people, for example, object to gratuitous cruelty, distinguish between acts and omissions, and think that we have special obligations to our family.

**2D model:** the model proposed by Kahane, Everett, and colleagues. According to the 2D model, proto-utilitarian decision making in the lay population involves two largely independent dimensions: instrumental harm and impartial beneficence. These have different psychological correlates and are likely to rely on different processes.

**Dual-process model (DPM):** the model proposed by Greene and colleagues [18,19,21,28,64]. According to the DPM, non-utilitarian (often referred to as deontological) aspects of CSM (e.g., refusing to sacrifice the one) are based on immediate intuitions and emotional responses, whereas 'utilitarian' judgments (e.g., sacrificing one to save a greater number) are uniquely attributable to effortful moral reasoning.

**Impartial beneficence (IB):** utilitarianism requires us to impartially maximize the well-being of all sentient beings on the planet, not privileging compatriots, family members, or ourselves over strangers. This is the 'positive' dimension of utilitarianism that we term impartial beneficence.

**Instrumental harm (IH):** one way that utilitarianism departs from common-sense morality is that utilitarianism permits, or even requires, many acts that CSM strictly forbids. This is a 'negative' dimension of utilitarianism that we term instrumental harm because, according to utilitarianism, we should instrumentally use, severely harm, or even kill innocent people to promote the greater good.

**Oxford utilitarianism scale (OUS):** this scale was developed by Kahane, Everett, *et al.* [34] as a brief measure of individual differences in proto-utilitarian moral tendencies. The scale consists of nine items in two subscales: instrumental harm (OUS-IH) and impartial beneficence (OUS-IB). The OUS-IB subscale consists of five items that measure endorsement of impartial

The discussion will be structured around two key questions about the relationship between pro-sacrificial judgments and utilitarianism. The first is the internal content question: is there a sufficient resemblance between pro-sacrificial judgments – and the processes generating them – with what is distinctive about utilitarian decision making? The second is the generality question: even if people do engage in something resembling a utilitarian decision-making procedure in the specific context of sacrificial dilemmas, do the associated psychological processes also drive other utilitarian departures from CSM? Answers to these questions can help to clarify what can be learned from the sacrificial dilemmas paradigm while also highlighting its limits and the need for new research paradigms.

### The Internal Content Question

One major challenge has centered on the persistent association of pro-sacrificial judgments with markedly antisocial personality traits and beliefs [35], leading researchers to suggest that characterizing pro-sacrificial judgments as utilitarian is misleading (e.g., [35–38]). It has been found, for example, that pro-sacrificial judgments are associated with reduced aversion to harm [39], psychopathy at both a clinical [29] and subclinical level (e.g., [35,36]), and even with endorsement of rational and ethical egoism: the idea that, in contrast to the utilitarian focus on impartial welfare maximization, an action is rational or moral only if it maximizes one's own self-interest [36]. This association raised the worry that many pro-sacrificial choices merely reflect a weaker aversion to harming others, regardless of the benefit, rather than a greater concern about consequences (e.g., [36–38]). If so, then there may be only a superficial overlap between the pro-sacrificial judgments and the prescriptions of utilitarianism. Consequently, studying sacrificial dilemmas will tell us little about why and how utilitarians depart from CSM. We term this the internal content question.

*Internal Content Question: Are Pro-Sacrificial Judgments the Result of Meaningfully Utilitarian Processes or Do They Reflect Only a Superficial Overlap in Judgments in a Highly Unusual Context?*

Note that what is at issue is not whether individuals make pro-sacrificial decisions because of conscious application of utilitarian principles (which few, if any, lay people are likely to do) but whether there is a sufficiently meaningful overlap between the moral reasons that justify the sacrifice from a utilitarian standpoint and what makes some ordinary lay people endorse the sacrifice [39].

Two developments have sought to address this challenge. The first is conceptual: conceding that many pro-sacrificial judgments may be 'utilitarian' only in the sense that they overlap with paradigmatic utilitarian prescriptions, while distinguishing a range of more meaningful ways in which judgments can echo genuine utilitarian decision making without involving the application of an explicit theory [21] (see Box 1 for discussion). This, however, leads to the question whether the pro-sacrificial judgments of at least some lay people do reflect such heightened concern for consequences instead of mere indifference to harm. The second advance aims to address this question via an important refinement of the sacrificial dilemma paradigm. As we have seen, conventional dilemma analyses fail to distinguish between the 'utilitarian' tendency to maximize good outcomes and the absence of 'deontological' concerns about causing harm. It has been argued that when these are teased apart using the technique of process dissociation (PD) (Box 2), we can identify a subset of pro-sacrificial judgments (the 'U-parameter') that reflect a genuine concern for the greater good and are therefore meaningfully utilitarian [20,21], although research so far suggests that indifference to harm is a stronger driver of pro-sacrificial choices [21]. More recent refinements of the paradigm

maximization of the greater good, even at great personal cost (e.g., 'It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal'). The OUS-IH subscale consists of four items relating to willingness to cause harm so as to bring about the greater good (e.g., 'It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people').

### Pro-sacrificial/pro-sacrificial

**judgments:** pro-sacrificial judgments, as we use the term, is a purely descriptive label that refers to morally approving the sacrifice of some to save a greater number. Crucially, this label is intended to describe such judgments without any presumption of motive, underlying process, or philosophical commitments, and is therefore, we suggest, preferable to describing such judgments as 'utilitarian'.

**Utilitarianism:** a normative ethical theory that is associated with philosophers such as Jeremy Bentham, John Stuart Mill, and Peter Singer, and posits that the whole of morality can be deduced from a single general principle: always act in the way that would impartially maximize aggregate well-being.

**Box 1. What Is a 'Utilitarian' Judgment?**

Researchers routinely describe pro-sacrificial judgments as 'utilitarian judgments', report factors associated with different rates of 'utilitarian' judgments, and describe populations as more or less utilitarian. We have argued that pro-sacrificial judgments often have little meaningful relationship to utilitarian ethics and that, although some pro-sacrificial judgments reflect aspects of utilitarian reasoning, even these are narrowly focused on the sacrificial context.

Any terminology can be valid if it is clearly defined and widely understood. However, some terminologies are perspicuous whereas others are imprecise, have irrelevant associations, or are potentially misleading. We propose that it is more helpful to refer to moral judgments favoring the sacrifice of some to save a greater number as 'pro-sacrificial': a purely descriptive label making no commitment to underlying motivations or processes. Because utilitarian reasoning has multiple dimensions, it is imprecise to categorically refer to judgments, processes, or individuals as 'utilitarian'. Instead, we should directly describe these in terms of the subcomponents that our 2D model highlights. For example, pro-sacrificial judgments may reflect endorsement of instrumental harm but not greater impartiality, whereas the reverse may be true of judgments endorsing sacrifices in aid of distant strangers.

A contrasting approach defines 'utilitarian' as any moral judgment that happens to be consistent with utilitarian theory, even if the reasons driving that judgment bear no relationship to utilitarianism [21]. Such judgments can be described as 'level 1' utilitarian, but should still be contrasted with ways in which judgments can genuinely echo utilitarian reasoning – by being driven by calculation of aggregate utility (level 2); genuine concern for the greater good in a specific context (level 3); a general concern for the greater good (level 4); and, finally, by applying an explicit utilitarian theory (level 5). By making clear that no meaningful relationship with utilitarianism is intended, such an approach presents an advance over the looser way in which the term 'utilitarian judgment' is often used.

However, it still seems unhelpful to describe behavior using theoretical labels that bear merely an accidental relationship to that behavior. Moreover, such a labeling system might misleadingly suggest that there is a common psychological phenomenon underlying a range of behaviors that are only arbitrarily grouped together. More importantly, even if researchers insist on using 'utilitarian' to refer to any pro-sacrificial judgment, there is no basis for reserving this label only for judgments about instrumental harm. Judgments endorsing self-sacrifice to aid distant strangers to make the world overall better are surely equally deserving of that label. This means that 'utilitarian' must always be explicitly relativized to a moral context, and we should not report without qualification that, for example, 'empathic concern is associated with reduced rates of utilitarian judgment' when this is the case only in the sacrificial dilemma context but not in others (e.g., that of impartial beneficence).

have sought to extract a third factor that shapes responses to sacrificial dilemmas: a preference for inaction over action [22].

These methodological advances try to address the internal content question by distinguishing pro-sacrificial judgments that merely superficially overlap with utilitarianism from those that involve genuine concern about outcomes, although at the cost of removing the simplicity that made sacrificial dilemmas so attractive as a general paradigm for studying moral decision making.

However, although concern about saving a greater number of lives bears an obvious resemblance to the utilitarian aim, several important gaps remain. First, PD measures a greater concern for better outcomes. However, nearly all ethical theories say that saving more lives is better. What is distinctive about utilitarianism in its classical form is that it says that we must save the most lives we can – that it is morally required (not merely permissible) to sacrifice one life to save two, or 50 to save 52. At present, there is no evidence from PD or traditional analyses suggesting that lay people make such judgments – and some evidence showing that they do not [40,41]. Second, utilitarianism instructs us to maximize utility in an uncompromisingly impartial way. Nevertheless, there is considerable evidence that pro-sacrificial judgments are strongly influenced by whether those sacrificed/saved belong to one's ingroup (e.g., [42,43]), and there is so far no evidence that the PD U-parameter is associated with greater impartiality. We will return to this issue below.

**Box 2. Process Dissociation**

Process dissociation is a data analytic approach that allows researchers to examine the contribution of two distinct processes to a given behavior by comparing the outcomes on trials in which the two processes should lead to the same outcome (congruent trials) versus those in which the two processes should lead to opposite outcomes (incongruent trials). Originally developed in cognitive psychology by Jacoby and colleagues [65], Conway and Gawronski [20] applied process dissociation to sacrificial dilemmas by studying responses in incongruent dilemmas in which harm maximizes outcomes, and congruent dilemmas in which harm does not maximize outcomes.

Incongruent dilemmas are typical sacrificial dilemmas where, for example, one must decide whether to administer a treatment that will prove fatal to some but will save the lives of many others. Congruent dilemmas, by contrast, are dilemmas where the harm is the same – administering a treatment that will prove fatal to some – but where doing so will not maximize outcomes (e.g., where it will shorten the duration of a non-fatal disease that most will recover from naturally).

Process dissociation then involves applying the responses of participants across these congruent and incongruent dilemmas to a decision-processing tree that allows researchers to calculate two parameters representing the influence of each tendency. The first parameter reflects those with relatively stronger harm-rejection tendencies (the 'D-parameter' that indicates 'deontological' inclinations to avoid causing harm), who consistently reject causing harm in the dilemmas, whether or not such harm would lead to overall positive consequences. The second parameter reflects those with outcome-maximization tendencies (the 'U-parameter', reflecting 'utilitarian' inclinations to optimize results) who tend to aim for the best possible consequences in the dilemma regardless of whether doing so requires causing harm or not (i.e., they endorse harm when it maximizes overall welfare but reject harm when it does not).

By calculating these parameters across the congruent and incongruent dilemmas, process dissociation is intended to allow researchers to distinguish between different patterns underlying the same conventional dilemma decision. Consider a person who tends to make pro-sacrificial decisions in the dilemmas – this response tendency could reflect a weak aversion to causing harm (i.e., low D-parameter) that is associated with antisocial personality traits, or could reflect increased concern to minimize overall harm. Moreover, PD can identify cases where people score high on both response tendencies, which then largely cancel out on conventional measures (i.e., a suppression effect). For example, people scoring higher on moral identity internalization tend to score high on both the D and U parameters, and these dueling positive effects cancel out to a null effect on conventional dilemma judgments that treat 'deontological' and 'utilitarian' responses as opposites [20,21].

Current evidence thus suggests that many instances of pro-sacrificial judgments are merely superficially consistent with utilitarianism, and describing such judgments as 'utilitarian' can be misleading. The PD approach offers an important tool for distinguishing such judgments that are merely driven by lack of aversion to harming from those reflecting genuine concern for saving more lives [20,21]. However, although the PD U-parameter bears similarity in content to utilitarianism, there is still a significant gap – meaning that even this subfactor of pro-sacrificial judgments can be described as utilitarian only in a qualified sense (see also [35–38,44,45]).

Similar issues can be raised about the content of the cognitive processes that drive pro-sacrificial judgments, or even specifically the U-parameter. Suppose that the DPM is correct in holding that deliberative processes are necessary to allow us to overcome common-sense intuitions against pro-sacrificial decisions – explaining why, for example, pro-sacrificial judgments have been found to be less frequent under cognitive load [46,47]. Even so, pro-sacrificial judgments do not seem to involve greater deliberative effort when they are 'impersonal' (e.g., diverting a trolley rather than pushing someone), and even in more emotive sacrificial dilemmas there is no effect of time pressure when making pro-sacrificial decisions with efficient kill–save ratios (e.g., sacrificing one to save 500 instead of five) [47]. Moreover, it is unlikely that deliberative effort is necessary to make the trivial 'cost–benefit analysis' that five lives are greater than one. This suggests that evidence for the role of deliberative processes in pro-sacrificial judgments may merely reflect the fact that any counter-intuitive moral judgment requires greater cognitive effort ([48,49], but see [50]). Such counterintuitive judgments

could be in line with utilitarianism, but could also be 'Kantian' [48] or even egoistic. Indeed, recent research supports a role of deliberative processes when overriding CSM in favor of self-interested choices [51,52]. To the extent that deliberation seems to underlie both egoistic and utilitarian departures from CSM, the contrast between deliberative versus intuitive processes is likely to be too generic to account for what is distinctive about proto-utilitarian forms of moral decision making [49].

### The Generality Question

Even if the antisocial pathway to pro-sacrificial judgments can be parceled out using process dissociation, and at least some people engage in something resembling genuine utilitarian decision-making within the specific context of sacrificial dilemma, there remains what we call the generality question.

*Generality Question: Is There a Meaningful Link between Pro-Sacrificial Judgments and Other Utilitarian Departures from CSM? If So, Does Investigating the Processes Driving Pro-Sacrificial Judgments Shed General Light on Why and How People Make Utilitarian Departures from CSM?* This would not be such a significant issue if sacrificial dilemmas captured the key way in which utilitarians reject CSM. However, although utilitarianism does notoriously instruct us to harm some to benefit a greater number, this endorsement of instrumental harm is only one of many ways in which utilitarianism departs from CSM, and arguably not the central one. A more fundamental, positive aspect of utilitarianism is what we term **impartial beneficence** (IB) – the injunction to act in ways that give equal moral weight to the interests of everyone on the planet. In practical terms, this can lead to demands for extreme self-sacrifice to benefit distant strangers [53,54]. In addition, utilitarians reject retributive justice, special moral obligations to those close to us, the act/omission distinction, the intrinsic significance of fairness or rights, and so forth [13,55,56]. The psychological underpinnings of these central utilitarian departures from CSM are of considerable theoretical and practical interest – and must be addressed by a comprehensive account of utilitarian psychology. What therefore, if anything, does the psychology of pro-sacrificial judgments tell us about the processes involved in other utilitarian departures from CSM?

At the level of individuals, if when individuals make pro-sacrificial judgments they are manifesting a generalizable proto-utilitarian approach to moral questions, we should expect them to also do so in at least some other contexts. There is considerable evidence, however, that pro-sacrificial judgments do not generalize in this way [36]. Even when controlling for the antisocial element in pro-sacrificial judgments, we found no association between 'utilitarian' judgments in sacrificial dilemmas and central utilitarian departures from CSM relating to impartial beneficence, such as assistance to distant people in need, self-sacrifice, and impartiality [36]. Moreover, even when the more 'utilitarian' U-parameter is extracted using PD, it is uncorrelated or even negatively correlated with such characteristic utilitarian prescriptions relating to impartial beneficence (e.g., thinking that the affluent should do more to help needy people in developing countries, or that we must tackle climate change to prevent harm to future generations) [21]. Such findings support a conceptualization of the U-parameter as 'tracking a commitment to the local minimization of harm rather than a global pursuit of the greater good that goes beyond conventional expectations' [21]. Nonetheless, such global pursuit of the greater good, well beyond conventional expectations, is in fact at the philosophical core of utilitarianism. Thus, even the U-parameter appears to reflect only a tendency to favor better consequences in a specific (and unusual) moral context rather than a more general proto-utilitarian approach to moral decision making (to our knowledge, there is so far no research investigating whether the U-factor is associated with greater impartiality within the context of sacrificial dilemmas, but this seems unlikely

given the evidence reviewed above). Moreover, given that a tendency to make pro-sacrificial judgments (or the U-parameter more specifically) does not generalize to other moral domains, we still lack an account of why such 'utilitarian decision making' is triggered in some contexts but not others.

In reply, it has been argued that sacrificial dilemmas are best seen as shedding light not on the proto-utilitarian tendencies of individuals but on the processes that underlie paradigmatic judgments consistent with utilitarianism [21,57]. However, what scant evidence there is suggests the cognitive processes that have been claimed to underlie pro-sacrificial decisions do not generalize to other types of characteristic utilitarian judgments. For instance, different psychological traits are associated with making judgments in line with utilitarianism in the context of sacrificial dilemmas and in that of impartial beneficence, thus providing indirect evidence that such judgments involve different psychological processes [34,36]. This is supported by a conceptual priming study that found that priming intuition reduces pro-sacrificial judgments – in line with the DPM – but there was no comparable effect on utilitarian judgments relating to self-sacrifice and impartial concern for others [58], contrary to predictions made by prominent proponents of the DPM [28]. Another recent study found that a tendency to morally prioritize humans over animals decreased when participants were primed to think emotionally as opposed to deliberately – in other words, greater deliberation was associated with reduced impartiality [59].

Research employing sacrificial dilemmas regularly reports findings about the processes driving or influencing 'utilitarian' judgment [17,20,31,32,60] as well as about the utilitarian tendencies of specific populations [29,30,33,61]. Such claims can suggest a generality, and in some cases – as in the initial statements of the DPM – are clearly intended to have a general scope ([28,62], although more recent formulations of the DPM are more qualified [21]). However, although further research into this issue is needed, the current evidence suggests that responses to sacrificial dilemmas do not generalize – at either the individual or the process levels – to other paradigmatic contexts where utilitarianism departs from CSM, such as that of impartial beneficence. Caution is therefore warranted when linking psychological factors, processes, populations, or individuals to 'utilitarian' judgment simply on the basis of sacrificial dilemmas research. For example, several studies have tied empathic concern to a reduced tendency to make 'utilitarian' judgments. However, this is only in the context of sacrificial dilemmas – we found that empathic concern is also associated with a greater tendency to make 'utilitarian' judgments in the context of impartial beneficence [34,36]. It would therefore be more precise, we suggest, to use the purely descriptive term 'pro-sacrificial judgments' (as we do here) or at least to explicitly contextualize by referring to utilitarian judgments in the specific context of sacrificial dilemmas.

### **Moving Forward: A Multidimensional Approach to Utilitarian Psychology**

We have argued that sacrificial dilemmas have limitations as a general tool for studying utilitarian decision making. They can shed light on one important way in which utilitarianism departs from CSM intuitions. However, utilitarianism also departs from CSM in other equally if not more important ways – most notably by demanding a radical form of impartiality. On both conceptual and empirical grounds, we should not assume that these departures all reflect a single, unitary cognitive phenomenon. Instead, the available evidence suggests that moral judgments in other paradigmatic contexts in which utilitarianism departs from CSM are likely to be driven by, and involve, different psychological factors and processes from those that drive pro-sacrificial judgments. If so, then sacrificial dilemmas cannot be used to draw general lessons about utilitarian judgment or decision making.

Although utilitarianism does provide distinctive answers to dilemmas involving runaway trolleys (or ticking bomb scenarios), it also provides distinctive answers to questions about, for example, our obligations to the poor of the world, or the treatment of animals. Correspondingly, we need to incorporate the insights of sacrificial dilemmas research while considering this much broader range of moral contexts. This will provide a fuller picture of the psychological sources of proto-utilitarian forms of moral decision making while also shedding light on counterintuitive moral views that are of considerable independent theoretical and practical interest.

These ideas form the basis for the **2D model** of proto-utilitarian psychology [34]. The 2D model (Box 3 and Figure 1, Key Figure) is inspired by the recognition that, conceptually, there are at least two primary ways in which utilitarianism departs from our common-sense moral intuitions: it permits harming innocent individuals when this maximizes aggregate utility (instrumental harm), and it tells us to treat the interests of all individuals our acts (or omissions) can affect as equally morally important, without giving priority to oneself or to those to whom one is especially close (impartial beneficence). As well as being conceptually distinct, these aspects of utilitarianism appear to be psychologically distinct in the lay population (although not in trained philosophers; [34] for further discussion). We have already reviewed research showing how judgments in sacrificial dilemmas have little correlation with judgments relating to donating money to assist distant needy strangers (an example of impartial beneficence) [36], and that cognitive priming manipulations influence instrumental harm but not impartial beneficence [58]. Most recently, in work developing the **Oxford utilitarianism scale** (OUS), large-scale factor analysis found that endorsement of the moral claims that are distinctive of utilitarianism clusters into two independently important factors, aligning closely with instrumental harm and impartial beneficence, each of which has very different psychological correlates (note that, although the 2D model emphasizes these two factors, which emerged from factor analysis of a wide list of items that covered other ways in which utilitarianism clashes with CSM, further utilitarian departures from CSM also merit close study).

### Box 3. The 2D Model of Proto-Utilitarian Psychology

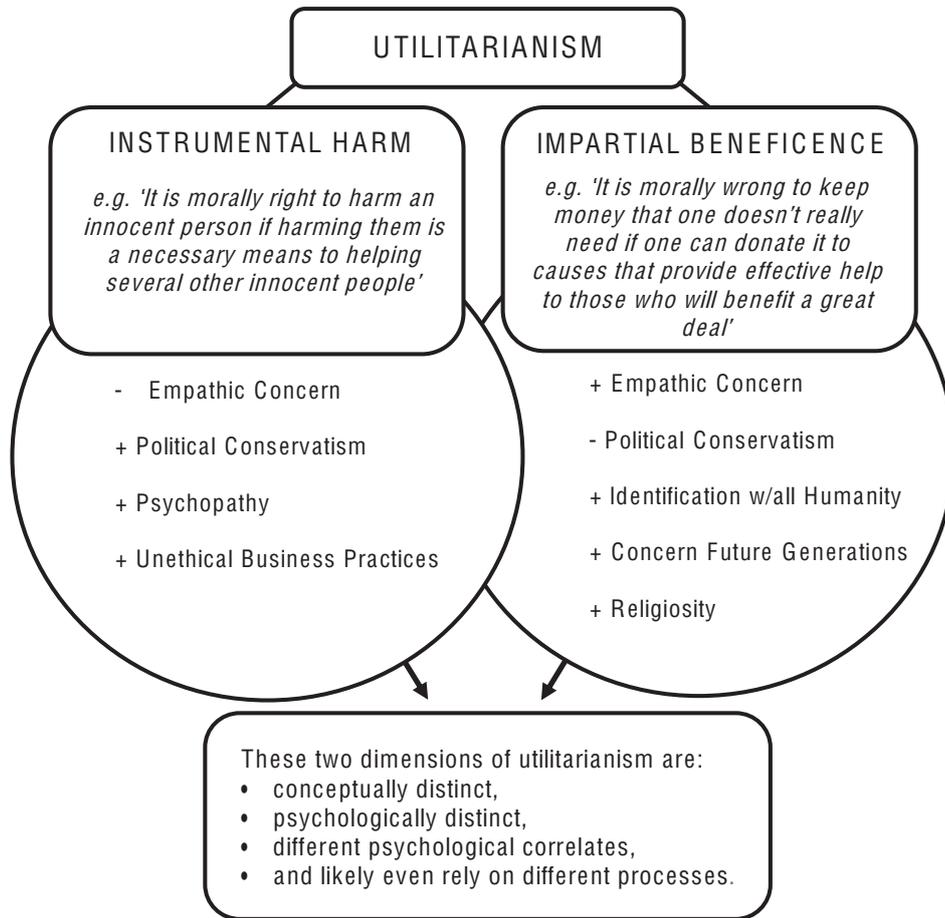
According to this model, utilitarianism has two main psychological dimensions. First, impartial beneficence (IB) reflects the extent to which individuals endorse impartial promotion of the welfare of everyone. Second, instrumental harm (IH) reflects the extent to which people endorse harm that brings about a greater good. By dissociating these two factors of utilitarianism, one can reach a more nuanced picture of proto-utilitarian tendencies in the lay population.

Instrumental harm can be measured using sacrificial dilemmas, reflecting one key way that utilitarianism departs from CSM: it permits, or even requires, many acts that CSM forbids. Whereas CSM tends to reject instrumental harm (except when the benefits are very large), according to utilitarianism we should always use, harm, or even kill innocent people if this leads to a greater good. Much sacrificial dilemma research has investigated this dimension, and the 2D model can incorporate many of these insights – while also recognizing that instrumental harm is not the only (or even most important) way in which utilitarianism departs from CSM.

Impartial beneficence reflects a second and more fundamental way that utilitarianism departs from CSM: utilitarianism requires us to impartially maximize the well-being of all sentient beings in such a way that '[e]ach is to count for one and none for more than one' [1], meaning that it requires altruist sacrifices that CSM at best sees as permissible or supererogatory. Whereas IH is an important implication of utilitarian principles, IB is directly written into the utilitarian ideal. It is for this reason that IB, and not IH, is the central utilitarian aim. Peter Singer – the most prominent living utilitarian – may have defended infanticide in some contexts, an example of IH, but his core moral aims are those relating to IB – for example making great sacrifices to prevent the suffering of the world's poor or of animals. Utilitarianism demands much more than CSM, both in how much we should sacrifice and for whose sake, and IB represents this radically impartial and demanding core of utilitarianism, beyond more familiar forms of altruism and pro-sociality. For example, whereas CSM encourages modest acts of charity if the sacrifice is not too great (with anything beyond being supererogatory), utilitarianism demands that we do the most good we can [54], forgoing luxuries and undergoing relative financial hardship to help those who are much worse off. Similarly, CSM typically endorses helping those who are near-and-dear to us, whereas utilitarianism requires that we make significant sacrifices to aid complete strangers in distant countries.

## Key Figure

## The 2D Model of Utilitarianism



Trends in Cognitive Sciences

Figure 1.

A multidimensional framework can address the issues outlined above. The internal content question asks whether lay moral judgments reflect meaningfully utilitarian processes or are merely a superficial overlap in judgments. As a normative theory, utilitarianism claims that the morally right act is that which maximizes utility from an impartial standpoint, regardless of the means needed to achieve it. This is a simple principle, but it has several dimensions which can come apart in the moral thinking of lay people who do not arrive at moral decisions by applying such an explicit principle. A key idea of the 2D framework is that lay judgments can resemble utilitarian theory in different ways and different degrees. Instead of categorically describing lay judgments [17,28], processes [20–22], and the biases of individuals [61,63] as 'utilitarian', as is now common, on the 2D framework we should approach the subcomponents of utilitarianism independently, investigating the degrees to which lay moral thinking is impartial and focuses on outcomes rather than on means, etc. In this way, the 2D approach breaks down the issue of

meaningful relation to more fine-grained sub-questions that need to be investigated separately (see Outstanding Questions). For instance, further research will be necessary to investigate the processes that underlie different types of utilitarian departures from CSM intuitions – distinguishing the generic processes involved in overcoming strong intuitions of any sort from those that reflect what may be distinctive about opting for counterintuitive utilitarian modes of moral thinking. Another issue that requires further investigation is the degree to which judgments in line with impartial beneficence meaningfully echo utilitarian ideals. In the same way as some individuals endorse instrumental harm merely because they are indifferent to violence, some individuals may make impartial choices simply because they independently care less about the self, or have weaker attachments to family, place, or country: the process-dissociation approach could potentially be applied to this issue.

The generality question concerns whether the same psychological processes are involved when people depart from CSM to make a pro-sacrificial judgment as when they depart from CSM in other utilitarian ways. Existing research indicates that people who make pro-sacrificial judgments do not tend to also endorse self-sacrificial actions to help strangers in developing countries, and vice versa. This suggests that it is unlikely that the same processes will underlie both types of judgment, although this issue requires further investigation. For example, it is possible that, because radical impartiality is highly counterintuitive, it might also rely on the deliberative processes that are thought to underlie some pro-sacrificial decisions. Crucially, however, the 2D model does not assume that the same processes underlie different types of proto-utilitarian decisions. Sacrificial dilemmas are useful for studying instrumental harm, but dedicated dilemmas will be necessary to investigate the psychology underlying endorsement of impartial beneficence [37].

### Concluding Remarks

Nearly two decades of research have used sacrificial dilemmas to shed light on utilitarian decision making, but sacrificial dilemmas are only one instance of where there is tension between utilitarianism and common-sense moral views. We have argued that, to understand proto-utilitarian decision-making more generally, it is crucial to adopt a multidimensional approach, looking at both instrumental harm and impartial beneficence. Previous research employing sacrificial dilemmas has yielded important insights into our understanding of instrumental harm, but has told only half of the story about the psychology of utilitarian decision making.

### References

- Bentham, J. (1983) *The Collected Works of Jeremy Bentham: Deontology, together with a Table of the Springs of Action; and the Article on Utilitarianism*, Oxford University Press
- Mill, J.S. (1863) *Utilitarianism*, Parker, Son, and Bourne
- Sidgwick, H. (1907) *The Methods of Ethics*, Hackett Publishing
- Singer, P. (1993) *Practical Ethics*, Cambridge University Press
- de Lazari-Radek, K. and Singer, P. (2017) *Utilitarianism: A Very Short Introduction*, Oxford University Press
- Haidt, J. (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814–834
- Baron, J. and Ritov, I. (1994) Reference points and omission bias. *Organ. Behav. Hum. Decis. Process.* 59, 475–498
- Ritov, I. and Baron, J. (1990) Reluctance to vaccinate: omission bias and ambiguity. *J. Behav. Decis. Mak.* 3, 263–277
- Spranca, M. et al. (1991) Omission and commission in judgment and choice. *J. Exp. Soc. Psychol.* 27, 76–105
- Baron, J. and Szymanska, E. (2011) Heuristics and biases in charity. In *The Science of Giving: Experimental Approaches to the Study of Charity* (Oppenheimer, D.M. and Olivola, C.Y., eds), pp. 215–230, Psychology Press
- Baron, J. et al. (1993) Attitudes toward managing hazardous waste: what should be cleaned up and who should pay for it? *Risk Anal.* 13, 183–192
- Baron, J. (1993) Heuristics and biases in equity judgments: a utilitarian approach. In *Psychological Perspectives on Justice: Theory and Applications* (Mellers, B.A. and Baron, J., eds), pp. 109–137, Cambridge University Press
- Baron, J. (1994) Nonconsequentialist decisions. *Behav. Brain Sci.* 17, 1–10
- Sunstein, C.R. (2005) Moral heuristics. *Behav. Brain Sci.* 28, 531–541
- Foot, P. (1967) The problem of abortion and the doctrine of double effect. *Oxf. Rev.* 5, 5–15
- Thomson, J.J. (1985) The trolley problem. *Yale Law J.* 94, 1395–1415
- Greene, J.D. et al. (2008) Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107, 1144–1154
- Greene, J.D. (2007) Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends Cogn. Sci.* 11, 322–323
- Greene, J.D. et al. (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108

### Outstanding Questions

Do similar processes underlie utilitarian departures from CSM in the domain of instrumental harm and impartial beneficence? For example (i) is making the 'utilitarian' choice in both dimensions associated with more deliberative and controlled cognitive processes? (ii) Do time-pressure and cognitive load affect the endorsement of instrumental harm and impartial beneficence in similar ways? (iii) Do individual difference tendencies towards reflective thinking correlate with making 'utilitarian' decisions in both dimensions?

Why do people depart from CSM in a utilitarian direction in some moral contexts but not in others? If these departures involve deliberative processing, why are deliberative processes triggered in some contexts but not in others, and what takes place during this deliberation? For example, to what extent are people engaging in cost–benefit analysis versus also considering whether the initial intuition deserves any weight, reflecting on the specific features of the scenario, weighing opposing claims against each other, applying a general theory to the situation at hand, and so on?

Do utilitarian counterintuitive judgments only rely on deliberation, or are deliberative processes also necessary when making other apparently counterintuitive moral judgments – for example, endorsing absolutist Kantian views?

Is the U-parameter (Box 2) associated with greater impartiality within the context of sacrificial dilemmas (e.g., would people higher in the U-parameter be equally likely to sacrifice their family or compatriots as they would a stranger)?

To what extent can all judgments in line with impartial beneficence be construed as meaningfully utilitarian? How much do they reflect an impartial desire to maximize the greater good, as opposed to decreased self-concern or weak attachments to family and community?

How stable are people's judgments relating to instrumental harm and impartial beneficence over time?

20. Conway, P. and Gawronski, B. (2013) Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *J. Pers. Soc. Psychol.* 104, 216
21. Conway, P. et al. (2018) Sacrificial utilitarian judgments do reflect concern for the greater good: clarification via process dissociation and the judgments of philosophers. *Cognition* 179, 241–265
22. Gawronski, B. et al. (2017) Consequences, norms, and generalized inaction in moral dilemmas: the CNI model of moral decision-making. *J. Pers. Soc. Psychol.* 113, 343–376
23. Duke, A.A. and Bègue, L. (2015) The drunk utilitarian: blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition* 134, 121–127
24. Choe, S.Y. and Min, K.-H. (2011) Who makes utilitarian judgments? The influences of emotions on utilitarian judgments. *Judgm. Decis. Mak.* 6, 580–592
25. Côté, S. et al. (2013) For whom do the ends justify the means? Social class and utilitarian moral judgment. *J. Pers. Soc. Psychol.* 104, 490–503
26. Christensen, J.F. et al. (2014) Moral judgment reloaded: a moral dilemma validation study. *Front. Psychol.* 5, 607
27. Bauman, C.W. et al. (2014) Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology: external validity in moral psychology. *Soc. Personal. Psychol. Compass* 8, 536–554
28. Greene, J.D. (2008) The secret joke of Kant's soul. In *Moral Psychology (Vol. 3): The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (Sinnott-Armstrong, W., ed.), pp. 35–80, MIT Press
29. Koenigs, M. et al. (2012) Utilitarian moral judgment in psychopathy. *Soc. Cogn. Affect. Neurosci.* 7, 708–714
30. Piazza, J. and Sousa, P. (2014) Religiosity, political orientation, and consequentialist moral thinking. *Soc. Psychol. Personal. Sci.* 5, 334–342
31. Bialek, M. and De Neys, W. (2017) Dual processes and moral conflict: evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgm. Decis. Mak.* 12, 148–167
32. Rosas, A. et al. (2019) Hot utilitarianism and cold deontology: insights from a response patterns approach to sacrificial and real world dilemmas. *Soc. Neurosci.* 14, 125–135
33. Helzer, E.G. et al. (2017) Once a utilitarian, consistently a utilitarian? Examining principledness in moral judgment via the robustness of individual differences: consistency of moral judgment. *J. Pers.* 85, 505–517
34. Kahane, G. et al. (2018) Beyond sacrificial harm: a two-dimensional model of utilitarian psychology. *Psychol. Rev.* 125, 131–164
35. Bartels, D.M. and Pizarro, D.A. (2011) The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 121, 154–161
36. Kahane, G. et al. (2015) 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition* 134, 193–209
37. Wiech, K. et al. (2013) Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition* 126, 364–372
38. Gawronski, B. and Beer, J.S. (2017) What makes moral dilemma judgments 'utilitarian' or 'deontological'? *Soc. Neurosci.* 12, 626–632
39. Cushman, F. et al. (2012) Simulating murder: the aversion to harmful action. *Emotion* 12, 2–7
40. Sheskin, M. and Baumard, N. (2016) Switching away from utilitarianism: the limited role of utility calculations in moral judgment. *PLoS One* 11, e0160084
41. Royzman, E.B. et al. (2015) Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cogn. Sci.* 39, 325–352
42. Cikara, M. et al. (2010) On the wrong side of the trolley track: neural correlates of relative social valuation. *Soc. Cogn. Affect. Neurosci.* 5, 404–413
43. Uhlmann, E.L. et al. (2009) The motivated use of moral principles. *Judgm. Decis. Mak.* 4, 476–491
44. Kahane, G. and Shackel, N. (2010) Methodological issues in the neuroscience of moral judgement. *Mind Lang.* 25, 561–582
45. Kahane, G. (2015) Sidetracked by trolleys: why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Soc. Neurosci.* 10, 551–560
46. Suter, R.S. and Hertwig, R. (2011) Time and moral judgment. *Cognition* 119, 454–458
47. Trémolère, B. and Bonnefon, J.-F. (2014) Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personal. Soc. Psychol. Bull.* 40, 923–930
48. Kahane, G. et al. (2012) The neural basis of intuitive and counterintuitive moral judgment. *Soc. Cogn. Affect. Neurosci.* 7, 393–402
49. Kahane, G. et al. (2014) Intuitive and counterintuitive morality. In *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*, pp. 9–39, Oxford University Press
50. Paxton, J.M. et al. (2014) Are 'counter-intuitive' deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Soc. Cogn. Affect. Neurosci.* 9, 1368–1371
51. Rand, D.G. (2016) Cooperation, fast and slow: meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychol. Sci.* 27, 1192–1206
52. Rand, D.G. et al. (2012) Spontaneous giving and calculated greed. *Nature* 489, 427–430
53. Singer, P. (1972) Famine, affluence, and morality. *Philos Public Aff* 1, 229–243
54. Singer, P. (2015) *The Most Good You Can Do: How Effective Altruism Is Changing Ideas about Living Ethically*, Yale University Press
55. Kagan, S. (1997) *Normative Ethics*, Routledge
56. Bykvist, K. (2009) *Utilitarianism: A Guide for the Perplexed*, Bloomsbury Publishing
57. Plunkett, D. and Greene, J.D. (2019) Overlooked evidence and a misunderstanding of what trolley dilemmas do best: commentary on Bostyn, Severhant, and Roets (2018). *Psychol. Sci.* 30, 1389–1391
58. Capraro, V. et al. (2019) Priming intuition decreases instrumental harm but not impartial beneficence. *J. Exp. Soc. Psychol.* 83, 142–149
59. Caviola, L. and Capraro, V. (2019) Liking but devaluing animals: emotional and deliberative paths to speciesism. *Soc. Psychol. Personal. Sci.* (in press), and PsyArXiv. Published online October 27, 2019. <http://dx.doi.org/10.31234/osf.io/sx5uw>
60. Baron, J. et al. (2015) Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *J. Appl. Res. Mem. Cogn.* 4, 265–284
61. Koenigs, M. et al. (2007) Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446, 908–911
62. Greene, J.D. et al. (2004) The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 389–400
63. Li, T. et al. (2019) Who is more utilitarian? Negative affect mediates the relation between control deprivation and moral judgment. *Curr. Psychol.* Published online July 13, 2019. <https://doi.org/10.1007/s12144-019-00301-1>
64. Greene, J.D. (2014) *Moral Tribes: Emotion, Reason and the Gap between Us and Them*, Atlantic Books
65. Jacoby, L.L. (1991) A process dissociation framework: separating automatic from intentional uses of memory. *J. Mem. Lang.* 30, 513–541

Even if the two dimensions rely on different processes ordinarily, can endorsement of impartial beneficence lead to higher endorsement of instrumental harm when the theoretical connection between the two is more salient?