**Everett, Jim A.C. and Kahane, Guy (2020)** *Switching Tracks? Towards a Multi-Dimensional Model of Utilitarian Psychology.* Trends in Cognitive Sciences, 23 (2). pp. 124-134. ISSN 1364-6613.

# Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology

Jim A.C. Everett₁* & Guy Kahane₂*

₁ School of Psychology, University of Kent
₂ Department of Philosophy, University of Oxford

*Both authors contributed equally to this piece.*

## Abstract

Sacrificial moral dilemmas are widely used to investigate when, how, and why people make judgments that are consistent with utilitarianism. But 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 of a utilitarian approach to morality? Second, to what extent do findings about sacrificial dilemmas generalise 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 are needed to investigate other key aspects of utilitarianism, such as its radical impartiality.

## Keywords

Utilitarianism; moral psychology; sacrificial dilemmas; harm; 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, positing that the whole of morality can be deduced from a single general principle: always act in the way that would impartially maximise aggregate well-being [1–5]. Out of 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 call 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 developmental, social and evolutionary origins. This research program has revealed that in many moral contexts, most people reject choices that maximise 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 deserved independently of any consequentialist deterrent effect [11]. In this body of research, it is usually assumed that such rejections of a utility-maximising 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].

A different approach emerged at the turn of the new millennium, one largely focusing on responses to so-called sacrificial dilemmas like the trolley scenarios first introduced by philosophers [15,16]. In these, 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, e.g. pushing someone off a footbridge to block a runaway train (note that throughout we use 'pro-sacrificial' as a purely descriptive label referring simply 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, 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 are a powerful tool for studying the cognitive and neural mechanisms underlying judgments about what we call **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. 19,28,18,21]. According to the DPM, a refusal to sacrifice individuals for the greater good, and thus to maximise utility, is based in immediate intuition and emotional gut-reactions. In 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 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. 18,21,33,34). 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' [30,31], and attribute utilitarian tendencies to individuals [24,32] and specific populations [33,34].

## 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. But this will require that the answer to the first question be substantive enough: 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 called '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 multi-dimensional approach [e.g. 35] 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.

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 of 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 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 centred around the persistent association of pro-sacrificial judgments with markedly anti-social personality traits and beliefs [36], leading researchers to suggest that characterising pro-sacrificial judgments as utilitarian is misleading [e.g. 36–39]. It has been found, for example, that pro-sacrificial judgments are associated with reduced aversion to harm [40], psychopathy at both a clinical [33] and sub-clinical level [e.g. 36,37], and even with endorsement of rational and ethical egoism: the idea that, contra the utilitarian focus on impartial welfare maximisation, an action is rational or moral only if it maximizes one's own self-interest [37]. 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. 37–39]. 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. Call this the Internal Content Question:

> *Internal Content Question: Are pro-sacrificial judgments the result of meaningfully utilitarian processes or is this just a superficial overlap in judgments in a highly unusual context?*

Note that what is at issue is not whether individuals make pro-sacrificial decisions due to a conscious application of utilitarian principles (which few, if any, laypeople 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 (40).

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 rather than 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 a "utilitarian" tendency to maximise good outcomes from 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) (see Box 2), we can identify a subset of pro-sacrificial judgments (the 'U-parameter') that does reflect a genuine concern for the greater good and is 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 have sought to extract a third factor shaping 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, while a concern about saving a greater number of lives does bear 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 1 life to save 2, 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 they don't [41,42]. Second, utilitarianism instructs us to maximise utility in an uncompromisingly *impartial* way. Yet there is considerable evidence that pro-sacrificial judgments are strongly influenced by whether those sacrificed/saved belong to one's ingroup [e.g. 43,44], and as of yet no evidence that the PD's U-parameter is associated with greater impartiality. We will return to this issue below.

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, while PD's U-parameter bears similarity in content

to utilitarianism, there is still a significant gap that means that even this sub-factor of pro-sacrificial judgments could be described as utilitarian only in a qualified sense [see also 36,39,37,45,46,38].

Similar issues can be raised about the content of the *cognitive processes* that drive pro-sacrificial judgments, or even specifically the U-parameter. Suppose the DPM is correct in holding that deliberative processes are required 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 [47,48]. Still, 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 1 to save 500 rather than 5) [48]. Moreover, it is unlikely that deliberative effort is needed to make the trivial 'cost-benefit analysis' that 5 lives is greater than 1. 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 [49,50,but see 51]. Such counterintuitive judgments could be in line with utilitarianism, but could also be 'Kantian' [49] or even egoistic. Indeed, recent research supports a role of deliberative processes when overriding CSM in favour of self-interested choices [52,53]. To the extent that deliberation seems to underlie both egoistic and utilitarian departures from CSM, the contrast between deliberative vs. intuitive processes is likely to be too generic to account for what is distinctive about proto-utilitarian forms of moral decision-making [50].

### The Generality Question

Even if the anti-social pathway to pro-sacrificial judgments can be parcelled 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? And 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 rejects CSM. But while utilitarianism does notoriously instruct us to harm some to benefit a greater number, this endorsement of instrumental harm is just 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 call **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 [54,55]. In addition, utilitarians reject retributive justice, special moral obligations to those close to us, the act/omission distinction, intrinsic significance to fairness or rights, and so forth [13,56,57]. 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. So what, 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 generalise in this way [37]. Even when controlling for the anti-social 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 [37] 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]. Yet 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 favour 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 as of yet 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 generalise 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 [21,58] . However, what scant evidence there is suggests the cognitive processes that have been claimed to underlie pro-sacrificial decisions do not generalize to other kinds 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, providing indirect evidence that such judgments involve different psychological processes [35,37]. 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 [59], contrary to predictions made by prominent proponents of the DPM [28]. Another recent study found that a tendency to morally prioritise humans over animals decreased when participants were primed to think emotionally as opposed to deliberately—i.e. greater deliberation was associated with *reduced* impartiality [60].

Research employing sacrificial dilemmas regularly report findings about the processes driving or influencing 'utilitarian' judgment [17,20,30,31,61] as well as about the utilitarian tendencies of certain populations [32–34,62]. Such claims can suggest a generality and in some cases—as in the initial statements of the DPM—are clearly intended to have such general scope [28,63; though more recent formulations of the DPM are more qualified; see 21]. However, while 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 level—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, a number of studies have tied empathic concern to a reduced tendency to make 'utilitarian' judgments. But this is so only in the context of sacrificial

dilemmas—we found that empathic concern is also associated with greater tendency to make 'utilitarian' judgments in the context of impartial beneficence [35,37]. 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 contextualise by referring to utilitarian judgments *in the specific context of sacrificial dilemmas.*

## Moving Forward: A Multi-Dimensional 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. But 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 than those that drive pro-sacrificial judgments. If so, then sacrificial dilemmas cannot be used to draw general lessons about utilitarian judgment or decision-making.

While 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 world's poor, 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 **two-dimensional (2D) model** of proto-utilitarian psychology [35]. The 2D model (see 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 maximises 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 those to whom one is especially close (impartial beneficence). As well as being conceptually distinct, these aspects of utilitarianism appear psychologically distinct in the lay population (though not in trained philosophers; see [35] 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) [37], and that cognitive priming manipulations influence instrumental harm but not impartial beneficence [59]. Most recently in work developing the **Oxford Utilitarianism Scale (OUS)**, a large-scale factor analysis found that the endorsement of moral claims distinctive of utilitarianism cluster into two independently important factors, aligning closely with instrumental harm and impartial beneficence, each of which have very different psychological correlates. (Note that while the 2D model emphasises 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).

A multi-dimensional framework can address the issues we 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 maximises 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 [62,64] 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, 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 is needed to investigate the processes that underlie different kinds 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 of 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. Just 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 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 kinds of judgments, though this issue requires further investigation. It is possible, for example, that since radical impartiality is highly counterintuitive, it might too rely on the deliberative processes thought to underlie some pro-sacrificial decisions. Critically, though, the 2D model does not assume that the same processes underlie different kinds of proto-utilitarian decisions. Sacrificial dilemmas are useful for studying instrumental harm, but dedicated dilemmas are needed to investigate the psychology underlying endorsement of impartial beneficence [see 37].

## Concluding Remarks

Nearly two decades of research have used sacrificial dilemmas to shed light on utilitarian decision-making, but sacrificial dilemmas are just one instance 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 critical to adopt a multi-dimensional 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.

## Glossary

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

**Dual Process Model (DPM):** The model proposed by Greene and colleagues. According to the DPM, non-utilitarian (often referred to as *deontological*) aspects of CSM (e.g. refusing

to sacrifice the one) are based in immediate intuitions and emotional responses, while "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 call 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 call instrumental harm because according to utilitarianism we should instrumentally use, severely harm, or even kill innocent people to promote the greater good.
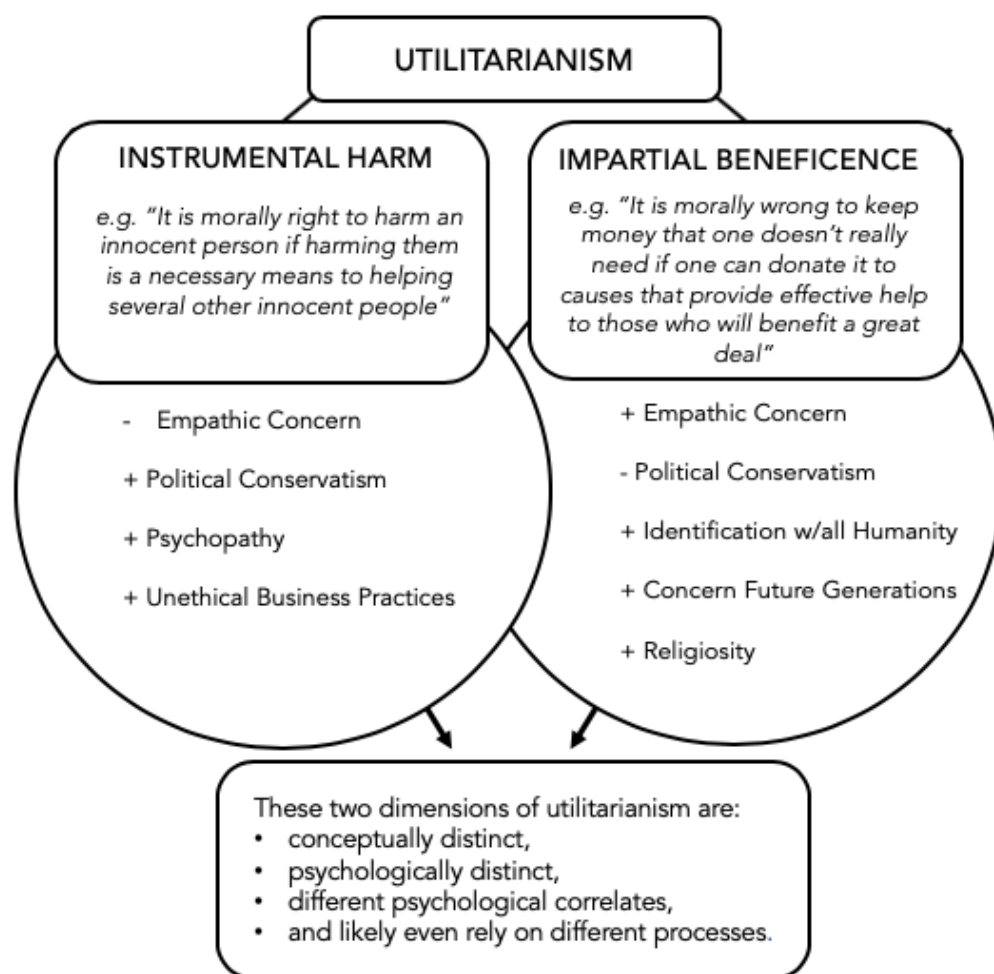
**Oxford Utilitarianism Scale (OUS):** The Oxford Utilitarianism Scale was developed by Kahane & Everett et al. [35] as a brief measure of individual differences in proto-utilitarian moral tendencies. The scale consists of 9 items in two sub-scales: *Instrumental Harm* (OUS-IH) and *Impartial Beneficence* (OUS-IB). The OUS-IB subscale consists of 5 items that tapping endorsement of the impartial maximization of the greater good, even at the 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 4 items tapping a willingness to cause harm in order 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 judgements**: 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. Critically, this label is meant 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".

**Two-Dimensional (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.

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

**Key Figure: The 2D Model of Utilitarianism**



UTILITARIANISM

**INSTRUMENTAL HARM**

e.g. *"It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people"*

- Empathic Concern
+ Political Conservatism
+ Psychopathy
+ Unethical Business Practices

**IMPARTIAL BENEFICENCE**

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"*

+ Empathic Concern
- Political Conservatism
+ Identification w/all Humanity
+ Concern Future Generations
+ Religiosity

These two dimensions of utilitarianism are:
• conceptually distinct,
• psychologically distinct,
• different psychological correlates,
• and likely even rely on different processes.

**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 while some pro-sacrificial judgments reflect aspects of utilitarian reasoning, even these are narrowly focused on the sacrificial context.

Any terminology can be valid if clearly defined and widely understood. However, some terminologies are perspicuous while others are imprecise, have irrelevant associations or are potentially misleading. We propose that it is more helpful to refer to moral judgments favouring the sacrifice of some to save a greater number as 'pro-sacrificial': a purely descriptive label making no commitment to underlying motivations or processes. Since 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 sub-components 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, 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 behaviour using theoretical labels that bear merely an accidental relationship to that behaviour. Moreover, such a labelling system might misleadingly suggest that there is a common psychological phenomenon underlying a range of behaviours 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 to judgments about instrumental harm. Judgments endorsing self-sacrifice to aid distant strangers to make the world overall better are surely as 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, e.g., '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. of impartial beneficence).

## 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 behaviour 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 maximises outcomes and *congruent* dilemmas in which harm does not maximise 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, in contrast, are dilemmas where the harm is the same – administering a treatment that will prove fatal to some - but where doing so will not maximise outcomes (for example, where it will shorten the duration of a non-fatal disease that most will recover from naturally).

Process dissociation then involves applying participants' responses 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" indicating '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-maximisation 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 maximises overall welfare but reject harm when it doesn't).

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 against causing harm (i.e., low D-parameter) associated with anti-social personality traits, or could reflect an increased concern to minimise 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 duelling positive effects cancel out to a null effect on conventional dilemma judgments that treat 'deontological' and 'utilitarian' responses as opposites [20,21].

**Box 3: The Two-Dimensional Model of Proto-Utilitarian Psychology**

According to the two-dimensional model of proto-utilitarian psychology, there are two main psychological dimensions of utilitarianism. First, Impartial Beneficence (IB) reflects the extent to which individuals endorse the impartial promotion of everyone's welfare. 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. While 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 recognising instrumental harm is not the only (or even most important) way in which utilitarianism departs from CSM.

Impartial beneficence reflects a second, 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 sees at best as permissible or supererogatory. While IH is an important implication of utilitarian principles, IB is directly written into the utilitarian ideal. It is for this reason that IB, 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 the ones relating to IB – e.g. making great sacrifices to prevent the suffering of the world's poor or of animals. Utilitarianism demands much more than CSM, both 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, while 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 [55], forgoing luxuries and undergoing relative financial hardship to help those who are much worse off. Similarly, CSM typically endorses the helping of those near-and-dear to us, while utilitarianism requires that we make significant sacrifices to aid complete strangers in distant countries.

**References**

1    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.

2    Mill, J.S. (1863) *Utilitarianism*, Parker, Son, and Bourne.

3    Sidgwick, H. (1907) *The methods of ethics*, Hackett Publishing.

4    Singer, P. (1993) *Practical Ethics*, Cambridge University Press.

5    Lazari-Radek, K. de and Singer, P. (2017) *Utilitarianism: A Very Short Introduction*, Oxford University Press.

6    Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814–834

7    Baron, J. and Ritov, I. (1994) Reference Points and Omission Bias. *Organ. Behav. Hum. Decis. Process.* 59, 475–498

8    Ritov, I. and Baron, J. (1990) Reluctance to vaccinate: Omission bias and ambiguity. *J. Behav. Decis. Mak.* 3, 263–277

9    Spranca, M. *et al.* (1991) Omission and commission in judgment and choice. *J. Exp. Soc. Psychol.* 27, 76–105

10   Baron, J. and Szymanska, E. (2011) Heuristics and biases in charity. *Sci. Giv. Exp. Approaches Study Charity*

11   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

12   Baron, J. (1993) Heuristics and biases in equity judgments: A utilitarian approach. In *Psychological perspectives on justice: Theory and applications* pp. 109–137, Cambridge University Press

13   Baron, J. (1994) Nonconsequentialist decisions. *Behav. Brain Sci.* 17, 1–10

14   Sunstein, C.R. (2005) Moral heuristics. *Behav. Brain Sci.* 28, 531–541

15   Foot, P. (1967) The problem of abortion and the doctrine of double effect. *Oxf. Rev.* 5, 5–15

16   Thomson, J.J. (1985) The trolley problem. *Yale Law J.* 94, 1395–1415

17   Greene, J.D. *et al.* (2008) Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107, 1144–1154

18   Greene, J.D. (2007) Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends Cogn. Sci.* 11, 322–323

19   Greene, J.D. *et al.* (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108

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,

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* pp. 35–80, MIT Press

29  Greene, J.D. (2014) *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*, Atlantic Books Ltd.

30  Białek, M. and De Neys, W. (2017) Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgm. Decis. Mak.* 12, 148–167

31  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

32  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

33  Koenigs, M. *et al.* (2012) Utilitarian moral judgment in psychopathy. *Soc. Cogn. Affect. Neurosci.* 7, 708–714

34  Piazza, J. and Sousa, P. (2014) Religiosity, political orientation, and consequentialist moral thinking. *Soc. Psychol. Personal. Sci.* 5, 334–342

35  Kahane, G. *et al.* (2018) Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychol. Rev.* 125, 131–164

36  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

37  Kahane, G. *et al.* (2015) 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition* 134, 193–209

38  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

39  Gawronski, B. and Beer, J.S. (2017) What makes moral dilemma judgments "utilitarian" or "deontological"? *Soc. Neurosci.* 12, 626–632

40  Cushman, F. *et al.* (2012) Simulating murder: the aversion to harmful action. *Emotion* 12, 2–7

41  Sheskin, M. and Baumard, N. (2016) Switching away from utilitarianism: the limited role of utility calculations in moral judgment. *PLOS ONE* 11, e0160084

42  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

43  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

44  Uhlmann, E.L. *et al.* (2009) The motivated use of moral principles. *Judgm. Decis. Mak.* 4, 476–491

45  Kahane, G. and Shackel, N. (2010) Methodological Issues in the Neuroscience of Moral Judgement. *Mind Lang.* 25, 561–582

46  Kahane, G. (2015) Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Soc. Neurosci.* 0, 1–10

47  Suter, R.S. and Hertwig, R. (2011) Time and moral judgment. *Cognition* 119, 454–458

48  Trémolière, B. and Bonnefon, J.-F. (2014) Efficient Kill–Save Ratios Ease Up the Cognitive Demands on Counterintuitive Moral Utilitarianism. *Pers. Soc. Psychol. Bull.* 40, 923–930

49  Kahane, G. *et al.* (2012) The neural basis of intuitive and counterintuitive moral judgment. *Soc. Cogn. Affect. Neurosci.* 7, 393–402

50  Kahane, G. (2014) Intuitive and Counterintuitive Morality. In *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (D'Arms, J. and Jacobson, D., eds), pp. 9–39, Oxford University Press

51  Paxton, J.M. *et al.* (2014) Are 'counter-intuitive' deontological judgments really counter-intuitive? An empirical reply to. *Soc. Cogn. Affect. Neurosci.* 9, 1368–1371

52  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

53  Rand, D.G. *et al.* (2012) Spontaneous giving and calculated greed. *Nature* 489, 427–430

54  Singer, P. (1972) Famine, affluence, and morality. *Philos. Public Aff.* 1, 229–243

55  Singer, P. (2015) *The most good you can do: How effective altruism is changing ideas about living ethically*, Yale University Press.

56  Kagan, S. (1997) *Normative Ethics*, Routledge.

57  Bykvist, K. (2009) *Utilitarianism: A Guide for the Perplexed*, Bloomsbury Publishing USA.

58  Plunkett, D. and Greene, J.D. (2019) Overlooked Evidence and a Misunderstanding of What Trolley Dilemmas Do Best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychol. Sci.* 30, 1389–1391

59  Capraro, V. *et al.* (In Press) Priming intuition decreases instrumental harm but not impartial beneficence.

60  Caviola, L. and Capraro, V. (In Press) Liking but Devaluing Animals: Emotional and Deliberative Paths to Speciesism. *Soc. Psychol. Personal. Sci.*

61  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

62  Koenigs, M. *et al.* (2007) Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446, 908–911

63  Greene, J.D. *et al.* (2004) The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron* 44, 389–400

64  Li, T. *et al.* (2019) Who is more utilitarian? Negative affect mediates the relation between control deprivation and moral judgment. *Curr. Psychol.* DOI: 10.1007/s12144-019-00301-1

65  Jacoby, L.L. (1991) A process dissociation framework: Separating automatic from intentional uses of memory. *J. Mem. Lang.* 30, 513–541