# Beyond Sacrificial Harm: A Two Dimensional Model of Utilitarian Decision-Making

Guy Kahane* [1], Jim A.C. Everett* [1, 2], Brian D. Earp [1], Lucius Caviola [2], Nadira Faber [1, 2], Molly J. Crockett [2], Julian Savulescu [1]

[1] *Uehiro Centre for Practical Ethics, University of Oxford*

[2] *Department of Experimental Psychology, University of Oxford*

Author Note

*Indicates joint contributions / shared first authorship.

Corresponding Author: Jim A.C. Everett, Department of Experimental Psychology, South Parks Road, Oxford, United Kingdom, OX1 3UD, Jim.ac.everett@gmail.com

**Abstract**

Recent research has relied on ‘trolley’ type sacrificial moral dilemmas to study ‘utilitarian’ vs. ‘non-utilitarian’ modes of moral decision-making. This approach has generated important insights into people’s willingness to endorse instrumental harm in certain circumstances—typically, sacrificing one individual in order to save a greater number—but also has serious limitations. Most notably, it ignores the positive, altruistic core of utilitarianism, which is characterized by impartial concern for the well-being of everyone, whether near or far. Here, we develop, refine, and validate a new scale – the Oxford Utilitarianism Scale – to dissociate individual differences in the ‘negative’ (willingness to cause instrumental harm) and ‘positive’ (impartial concern about the greater good) dimensions of utilitarian thinking as manifested in the general population. We show that these are two independent dimensions of proto-utilitarian tendencies in the lay population, each exhibiting a distinct psychological profile. Empathic concern, identification with the whole of humanity, and concern for future generations were positively associated with Impartial Beneficence but negatively associated with Instrumental Harm, and while Instrumental Harm was associated with sub-clinical psychopathy, Impartial Beneficence was associated with higher religiosity. Importantly, while these two dimensions were independent in the lay population they were closely associated in a sample of moral philosophers. Acknowledging this dissociation between the Instrumental Harm and Impartial Beneficence components of utilitarian thinking in ordinary people can clarify existing debates about the nature of moral psychology and its relation to moral philosophy as well as generate fruitful avenues for further research.

*Keywords*: moral psychology, utilitarianism, moral dilemmas; empathy; impartiality

## Disclosures and Acknowledgments

1. This research was supported by a research grant from the Uehiro Foundation on Ethics and Education.
2. All authors have contributed to this manuscript and have read and approved this submission.
3. The authors have no conflicts of interest to declare.
4. The authors gratefully acknowledge the assistance of our expert panel of philosophers who helped review items.

**Beyond Sacrificial Harm: A Two Dimensional Model of Utilitarian Decision-Making**

According to classical utilitarianism, we should always act in the way that would maximize aggregate welfare. Since its introduction in the 18th century by the philosopher Jeremy Bentham, this simple idea has been massively influential—and massively controversial. Modern secular morality can be seen as the gradual expansion of our circle of moral concern from those who are emotionally close, physically near, or similar to us, to cover the whole of humanity, and even all sentient life (Singer, 1981; see also Pinker, 2011). Utilitarians like Bentham, J. S. Mill, and, in our time, Peter Singer, have played a pivotal role in this process, and in progressive causes more generally. They have been leading figures in the fights against sexism, racism, and 'speciesism', influential supporters of political and sexual liberty, and key figures in efforts to eradicate poverty in developing countries as well as encouraging more permissive attitudes to pre-natal screening, abortion and euthanasia within our own societies (Bentham, 1789/1983; Mill, 1863; Singer, 2011). Yet utilitarians have never constituted more than a tiny minority, and utilitarianism has always faced fierce resistance. Pope John Paul II famously wrote: "Utilitarianism is a civilization of production and of use, a civilization of 'things' and not of 'persons,' a civilization in which persons are used in the same way as things are used." (John Paul II, 1995). And while it is not surprising that defenders of traditional morality reject utilitarianism, prominent progressive thinkers have criticized utilitarianism in similar terms (Rawls, 1971; Williams, 1973), and many continue to protest in anger what they regard as the dangerous views of utilitarians such as Singer (Schaler, 2009). Clearly, utilitarianism is a distinctive, influential, and controversial ethical view.

Given the influential but controversial reach of utilitarianism in ethics and society, questions about the psychological basis of utilitarian moral thinking—and why some people are so attracted to it while others are so repelled—have been of considerable interest to philosophers and psychologists alike. Utilitarians have often answered such questions by appealing to a contrast between cool logic and misguided intuitions and emotions. They argue that common moral views have their source in gut reactions and intuitions shaped by discredited religious views or evolutionary pressures, and that careful reflection should lead us to abandon these views and endorse utilitarianism, a more logical view based in rational reflection (Singer, 2005). But this notion is, in part, a testable hypothesis about human moral psychology, and in recent years, an influential body of research has seemed to confirm it. By studying responses to 'sacrificial' moral dilemmas (such as the famous 'trolley' scenario and its various permutations; see Foot, 1967) which present a choice between sacrificing one innocent person to save a greater number of people, or doing nothing and letting them die, researchers have tried to uncover the psychological and neural underpinnings of the dispute between utilitarians and its opponents, such as defenders of deontological, right-based views of the kind associated with Immanuel Kant. Dual process models conceptualize cognition as resulting from the competition between quick, intuitive, and automatic processes, and slow, deliberative, and controlled processes (e.g. Bargh & Chartrand, 1999; Chaiken & Trope, 1999; Shiffrin & Schneider, 1977). Running with this idea, influential research by Greene and colleagues has applied a dual process lens to our moral judgments to suggest that while deontological judgments (refusing to sacrifice the one) are based in immediate intuition and emotional gut-reactions, utilitarian judgments (sacrificing one to save a greater number) are uniquely due to effortful reasoning (Greene, 2007; Paxton, Bruni,

& Greene, 2014). It has also been suggested that these opposing utilitarian and deontological forms of decision-making are based in distinct neural systems (Greene, Nystrom, Engell, Darley, & Cohen, 2004).

More recent research, however, has yielded findings that are difficult to square with this flattering picture of utilitarian thinking. For example, multiple studies have reported an association between 'utilitarian' responses to sacrificial dilemmas and psychopathy and, more generally, with aggressive and antisocial tendencies and reduced concern about harm to others (Bartels & Pizarro, 2011; Glenn, Koleva, Iyer, Graham, & Ditto, 2010; Kahane, Everett, Earp, Farias, & Savulescu, 2015; Wiech et al., 2013). This is puzzling. Utilitarians are supposed to care about the good of all sentient beings; psychopaths notoriously care only about their own good. So why is psychopathy one of the traits most consistently associated with what are supposed to be paradigm cases of utilitarian judgment?

In March 2017, disability activists outraged by Peter Singer's support for the infanticide of severely disabled babies prevented him from speaking (via an internet link) at an event organized by the Effective Altruism Club of Victoria University in Canada—a club whose founding, in turn, was inspired by Singer's advocacy of self-sacrifice in the name of charity (Singer, 2015). This incident—and the two 'sides' of Singer's views attracting both censure and praise—offers the beginnings of an answer to our question by illustrating two distinct ways in which utilitarianism radically departs from commonsense morality.

The first way utilitarianism departs from commonsense ethical views is that it places no constraints whatsoever on the maximization of aggregate well-being. If killing a severely disabled child would lead to more good overall—as Singer believes is at least sometimes the

case—then utilitarianism, in stark contrast to commonsense morality, requires that the child be killed. This explains the angry protests at Singer's talk. But this is just one aspect of utilitarianism, and specifically it is the *negative* dimension according to which we are permitted (and even required) to instrumentally use, severely harm, and even kill innocent people to promote the greater good. We call this 'instrumental harm'.

There is also a positive dimension to utilitarianism, and this dimension, too, departs from common sense morality. Recall that utilitarianism requires us to maximize, not our own preferences or well-being—not even that of those near or dear to us, or of our compatriots—but the well-being of all sentient beings on the planet, and to do so in such a way that "[e]ach is to count for one and none for more than one" (Bentham, 1789/1983). This dimension explains why utilitarianism is sometimes described as a form of universal or impartial beneficence (which is what we shall call this positive dimension). For people in affluent countries, the demand to impartially maximize welfare is likely to require significant self-sacrifice—for example, by giving much of our income to charity. And although many find this level of sacrifice far too demanding (Cullity, 2004), this impartial ideal has inspired a global movement of 'effective altruists', including those in attendance at Singer's event at Victoria University (MacAskill, 2015; Singer, 2015).

The sacrificial dilemmas paradigm has yielded seemingly puzzling findings about utilitarian decision-making because it focuses almost exclusively on the negative side. It thus ignores or downplays the positive, impartial and altruistic core of a utilitarian approach to ethics. Accordingly, over a decade of research employing sacrificial dilemmas to study 'utilitarian' thinking has shed light only or primarily on instrumental harm: the conditions under which

people find it acceptable to cause harm for a greater good. Such dilemmas, however, tell us little about the sources of impartial concern for the greater good, despite the fact that this positive, all-encompassing altruistic aim is at the very heart of a utilitarian approach

A focus on the negative explains puzzling results: that psychopaths may be more willing to push someone off a footbridge to save five, and are likely to be less shocked by support for infanticide or euthanasia should not surprise us. But it would be surprising if these same psychopaths signed up to join an Effective Altruism Club or showed care for the plight of strangers in the developing world. In other words, recent research has told only half of the story about the psychology of utilitarianism. And since impartial beneficence is the philosophical core of utilitarian thought—whereas acceptance of instrumental harm is one implication of that central core, when it is endorsed without qualification — it has arguably focused on the less important half.

Our paper has three aims. First, we will propose a new conceptual framework for thinking about the psychology of utilitarian tendencies in the lay population. We hope that this framework will also serve as a general model for thinking about the relationship between the explicit ethical theories debated by philosophers and the pre-theoretical moral decision-making of ordinary people. Second, using this framework, we will propose a new approach for studying individual differences in proto-utilitarian tendencies. We will introduce and validate a new scale—the Oxford Utilitarianism Scale (OUS)—that was designed to address important limitations of the sacrificial dilemmas paradigm. Third, we will propose a new theory of the psychological sources of proto-utilitarian modes of thinking, a theory that can explain recent puzzling findings about utilitarian judgment as well as generate new directions for research.

Current work in moral psychology has largely assumed that utilitarian decision-making is a unitary psychological phenomenon. By contrast, the *Two Dimensional (2D)* model of utilitarian thinking that we develop here highlights the distinct positive and negative components of utilitarian decision-making. Although these two dimensions overlap in explicit utilitarian theorizing, they often come apart in the moral thinking of lay persons and are indeed in some tension in that domain. Moreover, our findings suggest that both dimensions are largely driven by affective dispositions rather than by explicit reasoning. We will end by exploring the theoretical, methodological and practical implications of this overlooked division within utilitarian thinking. We will highlight, in particular, the way the 2D model casts doubt upon dominant philosophical and psychological accounts both of the psychological basis of utilitarianism and of the sources of continuing resistance to it.

## Utilitarianism and Moral Decision-Making in the Lay Population

### Ethical Theory and Moral Judgment in Non-Philosophers

Although there is a large and growing body of psychological research into utilitarian decision-making, this research has largely proceeded without a precise account of the sense in which the moral judgments of non-philosophers can be usefully described in terms drawn from explicit philosophical theories (though see Greene, 2008).

Moral philosophers develop, elaborate and debate explicit ethical theories. But while some ethical theories, or ideas derived from such theories (e.g. the Kantian concept of human dignity) occasionally become more widely known, there is little reason to think that lay people employ explicit ethical theories in forming their moral judgments—let alone the specific theories debated by academic philosophers. It is plausible, however, that such philosophical theories draw

upon pre-theoretical moral intuitions and tendencies. It is also likely that both attraction to, and rejection of, explicit ethical theories is driven, at least in part, by individual differences in such pre-theoretical moral tendencies. Such tendencies would involve being responsive to and emphasizing the factors that a given ethical theory regards as morally relevant, while also being reflected by patterns of moral judgments that at least partly mirror those supported by the theory. Less centrally, such tendencies can also involve forms of moral reasoning and deliberation that echo (or are the precursors of) those recommended by the theory.

Importantly, we shouldn't treat such pre-theoretical tendencies in an all or nothing matter; few if any non-philosophers are full-blown utilitarians. Such tendencies are rather a matter of degree: the moral thinking of ordinary individuals will approximate to a greater or lesser extent—as well as in some but not other respects—the patterns of judgments and response that characterize a given explicit ethical theory.

Even if non-philosophers do not form their moral judgments by applying an explicit ethical theory, why not simply ask them to what extent they endorse such a theory? Although tempting, this is not, we claim, a promising approach to measuring the moral views of ordinary people. To begin with, there is considerable evidence that people often do not have introspective access to the principles and factors to which their moral judgments are actually responsive. In addition, their judgments about concrete cases needn't reflect the relevant general principles that they would endorse (Cushman et al., 2006; Hauser et al., 2007; Lombrozo, 2009). To illustrate: the utilitarian idea that we should act in ways that promote everyone's happiness can sound very attractive in the abstract; but many reject utilitarianism when they realize the highly counterintuitive implications of treating this idea as the sole criterion of moral action—i.e. its

uncompromising demand for impartiality and self-sacrifice, on the one hand, and for the sacrifice of innocent others for the greater good, on the other (for example, by killing a patient and using his organs to save five others; see Horne, Powell and Hummel, 2015). Thus, even when people endorse the core utilitarian principle in the abstract, their actual moral judgments may still be guided by deontological considerations relating to rights, duties, or degrees of personal relationship. Therefore, to measure the extent to which people approximate an ethical theory such as utilitarianism, we need to approach things more indirectly, by examining a broader range of patterns of moral thought and judgment.

Finally, we need to strike a balance between philosophical accuracy and empirical plausibility. It is unlikely that the moral judgments of non-philosophers mirror the most intricate and subtle forms of the ethical theories developed by philosophers, nor should we expect the moral views of ordinary people to be fully consistent. At the same time, if we use terms such as ‘utilitarian’ too loosely, these terms will lack any interesting theoretical content and, indeed, mislead us into reading more into more mundane forms of ordinary moral judgment than is really there (Kahane, 2015). In the next section, we give specific examples to illustrate these considerations.

**Understanding Utilitarianism**

Utilitarianism involves more than the mundane ideas that we should aim to prevent suffering and promote happiness, or that it is morally better to save more rather than fewer lives. Utilitarians make a far more radical claim: that we should adopt a thoroughly impartial standpoint, aiming to maximize the well-being of all persons (or even all sentient beings), regardless of personal, emotional, spatial or even temporal distance (*positive dimension*); and

that this should be our one and only aim, unconstrained by any other moral rules, including rules forbidding us from intentionally harming innocent others (*negative dimension*). In line with framework set out above, it is worth pausing to spell out the distinctive patterns of moral thought and judgment involved in each of these dimensions of utilitarianism.

*The Positive Dimension of Utilitarianism: Impartiality*

The philosophical core of utilitarianism lies in the impartial maximization of the greater good.[1] To adopt a thoroughly impartial moral standpoint is to treat the well-being of every individual as equally important. No priority should be given to one's own good, nor to that of one's family, friends, compatriots or even fellow humans over non-human animals. Such a standpoint would normally imply highly demanding forms of self-sacrifice—whether by becoming vegetarian or vegan, giving much of one's money to effective charities aiming to relieve suffering in distant countries, or perhaps even donating one's own kidney. Indeed, utilitarianism instructs moral agents to sacrifice their own well-being to the extent that there is an even tiny increment in the well-being of others over what they have lost.

Notice, however, that such moral impartiality is not the same as altruism and self-sacrifice. Someone might not hesitate to risk their life to save a drowning child, while at the same time failing to conclude that they have any reason to give up an affluent life style. Ordinary, 'common-sense' morality encourages modest acts of altruism (e.g. helping a beggar or making an occasional donation to charity) and rewards heroism in the context of acute emergencies. But complete impartiality requires more – much more. The utilitarian Peter Singer, for example, is a leading proponent of effective altruism, a movement built around the idea of using reason and evidence to find the best ways to help others. Many effective altruists have

pledged to give at least 10% of their income to cost-effective charities (MacAskill, 2015; Singer, 2015). In the UK, the median amount given to charity per year is £168, and of the four most popular causes that people donate to (charities focused on children, medical research, animals, and hospices: Charities Aid Foundation, 2016), none are focused on the developing world, where arguably most good can be done. With a median salary of around £27,000, on utilitarian effective altruism principles we should donate at least £2,700—or 16 times the actual amount—and, moreover, to charities that would impartially do the most good. In fact, people typically do neither.

Utilitarianism diverges from ordinary morality not only with respect to how *much* we should sacrifice but also for *whose* sake. Some individuals engage in acts of extreme self-sacrifice—in some cases, sacrificing their lives to promote the good of their family, country or religious group. But such altruistic acts are hardly expressions of impartiality since they focus on one's friends, family or group. Utilitarianism instructs similar sacrifices for complete strangers or even one's enemies. Utilitarianism requires refusing to give priority to those close to you over others (saving the lives of compatriots, or even family, before those of distant strangers). Of course, how much one approximates this impartial ideal is a matter of degree—even avowed utilitarians admit that they fail to realize it without qualification (Singer, 2016). And there is more than one way of departing from this ideal—a religious fundamentalist may discount self interest in favour of her ingroup while an egoist might care only about himself while not differentiating much between strangers and his closest family members.

*The Negative Dimension of Utilitarianism: Harming and Breaking Rules*

A thoroughly impartial moral outlook is not, however, sufficient for utilitarianism. One can adopt such an outlook while still holding that the goal of maximizing everyone's well-being must only be pursued in line with various moral rules constraining us from certain ways of harming innocent people, lying, breaking promises and the like. In other words, even if one endorses this impartial moral *goal*, one may still think that we are forbidden from taking certain *means* to achieve it. The negative component of classical utilitarianism is the denial that there are any such constraints. We should of course still usually tell the truth, keep our promises, and refuse to harm innocent people—but only when (and because) these acts are likely to lead to a better impartial outcome. When they get in the way of achieving such an outcome, such familiar moral rules can be broken.

The most central of these rules relate to what we called instrumental harm—willingness to harm and even kill others when this is needed to achieve a better outcome. This willingness can be seen when, in the classical philosophical thought experiment, one pushes an innocent person off a footbridge to save a greater number of lives, but also, to turn to a more realistic example, when one holds that torture is morally acceptable if needed to reduce the risk of major terrorist attacks. This is why some utilitarians support the legalization of active euthanasia as well as, more controversially, the abortion and in some cases even infanticide of the severely disabled. Utilitarians, however, also reject the authority of many other putative moral rules—including, as mentioned, those relating to honesty and keeping promises, as well as to fairness, hierarchy, and 'purity' (i.e., a concern for strict sexual and other boundaries). It is for this reason

that utilitarians were amongst the earliest to support the legalization of homosexuality and, more generally, to defend a permissive attitude to sexuality (Bentham, 1785/1978).

Notice again, however, that one can reject many or even most of these rules and values without endorsing the impartial positive aim of utilitarianism. An avowed egoist, for example, might also regard constraints against lying or even killing in a purely instrumental way yet see no reason at all to care about the greater good.

**Degrees of Proto-Utilitarian Tendencies in Lay Moral Judgment**

Understood as an explicit ethical theory, classical utilitarianism is firmly committed both to unqualified impartiality and to the rejection of all inherent moral constraints on the maximization of aggregate well-being. Now, few if any non-philosophers are likely to consciously apply such an explicit theory. However, the moral thinking of ordinary people may approximate such an outlook to varying degrees. We propose that the closer a person approaches moral questions in ways that give weight to the concepts and considerations central to paradigmatic utilitarianism (i.e., the 'classic' view associated with Bentham and Mill), the stronger the utilitarian tendencies of that individual. Spelled out in the terms set out above, a person's moral thinking should count as more utilitarian, (1) the greater its focus on the impartial maximization of well-being across different moral contexts (*positive dimension*), and (2) the less space and weight it gives to values other than well-being, and to moral rules constraining the promotion of well-being (*negative dimension*).[2]

With respect to (1), we have seen that an individual can reject impartial morality both by privileging the self but also by privileging family, friends, or compatriots, and generally those who are spatially and temporally closer. With respect to (2), the other values and rules in

question could be both a matter of *number* (e.g. traditional morality accepts multiple moral rules that can constrain the promotion of aggregate well-being, such as rules relating to hierarchy, purity, and so on; see Haidt, 2012) and *strength* (e.g., libertarians typically accept far fewer moral rules than traditionalists, but the few rules they accept with respect to, e.g., property rights are extremely strong). In their strongest form, such competing moral rules state absolute prohibitions. But they needn't be as strong as that. Many non-utilitarians accept that we can break certain moral rules (e.g., relating to truth-telling or promises) when adhering to them would lead to significant harm (Holyoak & Powell, 2016), and plenty of non-utilitarians are willing to endorse causing severe harm to innocents in situations where the cost of refusing to do so are catastrophic (Fried, 1978). Stronger rules will have higher thresholds, and different individuals will draw these thresholds at different points (Tremoliere & Bonnefon, 2012).

Someone who thinks about morality in unqualifiedly impartial terms, privileging no one over another while rejecting any constraint whatsoever on the maximization of well-being, would count as fully utilitarian on the proposed construct. Such a person would closely conform, at least in their moral thinking, to classical act utilitarianism, and to the form of utilitarianism presently defended by philosophers such as Peter Singer (2011).

Utilitarianism can also take other forms. One is rule utilitarianism, which holds that the morally right action is the one that conforms to rules that, if widely adhered to, would maximize well-being. There are also non-utilitarian forms of consequentialism that recognize values beyond that of utility (e.g., the value of fairness) and even accommodate forms of partiality (Scheffler, 1982; Sen; 1982). Because they depart from classical utilitarianism in ways that bring

them closer to commonsense morality, adherents of such views would count as somewhat less utilitarian on the proposed construct.

It may be worth clarifying at the outset why we privilege classical act utilitarianism in this way. First, as a minor point, this is the form of utilitarianism assumed by most work in current moral psychology; our framework aims to improve on existing practice but also to be continuous with it rather than to change the subject. Second and more importantly, classical act utilitarianism is the original form of the view and remains the most famous, most influential, and most controversial. Third, some more recent developments of utilitarianism (e.g. 'motive' or 'global' utilitarianism, or defense of the distinction between criterion-of-rightness and decision-procedure; see Sinnott-Armstrong, 2015) are probably too subtle or complex to be reflected in the moral thinking of non-philosophers: it is not by accident that it took many decades of intense philosophical reflection to identify these variants. Fourth, most important deviations from classical act utilitarianism—rule utilitarianism, satisficing utilitarianism and non-utilitarian consequentialism being prime examples—are attempts to bring utilitarianism closer to commonsense morality and tone down its more radical and counterintuitive aspects. Since we are proposing a way to rank moral outlooks as *more* or *less* utilitarian, it is hard to see what could replace the most unqualified form of the view (which also happens to be the most paradigmatic) at the 'top' of the scale. Importantly, however, other forms of utilitarianism are not ignored by our framework—their adherents would be ranked as leaning strongly towards utilitarianism but somewhat less so than classical utilitarians. This seems exactly right. \

As we move further away from utilitarianism in its paradigmatic forms, we find approaches to morality that are increasingly partial, and that recognize a greater number of

increasingly stringent deontological constraints. Individuals who think of morality in these ways would count as low on utilitarian dispositions. Notice, though, that the construct we are developing is a measure of utilitarian tendencies, not a general taxonomy of possible moral views. There are multiple ways to reject utilitarianism—we already mentioned traditional morality, libertarianism, Kantian ethics and other rights-based approaches; there is also virtue ethics (Hursthouse, 1999) as well as others. These are very different non-utilitarian views, and the proposed construct is not intended to differentiate between them.

Finally, although we explained above what would count as a stronger or weaker utilitarian tendency by reference to a range of utilitarian and non-utilitarian theories, it bears emphasizing again that the construct we have in mind is a measure of broad tendencies in moral deliberation and judgment in the lay population. It is not likely that ordinary people apply anything resembling an explicit ethical theory (whether utilitarian or not), nor is it likely that their moral judgments are consistent across different moral contexts.

## Existing Measures of Utilitarian Decision-Making

### Sacrificial Dilemmas

Armed with this theoretical framework, we can return to the sacrificial dilemmas paradigm. Sacrificial dilemmas are by far the most dominant experimental paradigm in contemporary moral psychology (Christensen & Gomila, 2012), and are widely assumed to be a reliable measure of utilitarian decision-making. Pro-sacrifice responses to such dilemmas are routinely classified as ‘utilitarian judgments’, and the psychological processes and mechanisms implicated in them interpreted as reflecting general features of utilitarian decision-making

(Greene, 2008). Moreover, the number of pro-sacrifice responses to batteries of such dilemmas are widely used as measures of differences in utilitarian tendencies both within (Lombrozo, 2009) and between populations (Koenigs et al., 2007).

The work of Greene and colleagues on sacrificial dilemmas has been deeply influential to the field. They have spurred a decade of fascinating research in moral psychology, and made substantial advances to our understanding of instrumental harm. That said, despite – or perhaps because – of how popular sacrificial dilemmas have been in moral psychology, this approach has naturally invited some criticism, for example, because of the highly artificial character of the scenarios typically used (Bauman, McGraw, Bartels, & Warren, 2014). Our aim here is not to offer further criticism of this paradigm but to highlight its limits as a general measure of utilitarian decision-making.

Sacrificial dilemmas only directly measure what we call the negative dimension of utilitarianism. In fact, it measures only attitudes to instrumental harm - just one aspect of the negative dimension, albeit a very important one. To judge that, for example, we should push one innocent person off a footbridge is to reject (or at least discount) one possible deontological rule against directly harming others (as a means to preventing a greater harm to others). But one could reject this particular deontological rule while still accepting many other rules—for example, rules relating to fairness, honesty, or promise-keeping. And one can certainly reject this rule while remaining highly partial in one’s moral decision-making.

That someone makes a judgment that happens to be in line with utilitarianism in a specific context does not, of course, immediately show that their judgments stem from, or are responsive to, the considerations that lie at the heart of a utilitarian moral outlook. Nor does it

show that they will make judgments in a way that resembles such an outlook in other contexts. It is nevertheless an empirical possibility that pro-sacrifice responses to sacrificial dilemmas indicate a more general utilitarian outlook. That would be true if utilitarian decision-making is a *unitary* psychological phenomenon in the context of the lay population. This assumption is explicitly endorsed by Greene's pioneering Dual Process model (Greene, 2008) but also seems to underlie much other research employing such dilemmas whereby results are routinely stated as supporting general conclusions about the psychology of utilitarian decision-making (see e.g. Koenigs et al., 2007; Cote et al., 2013; Duke & Begue, 2015; Robinson et al., 2015).

And yet, the association we discussed earlier between 'utilitarian' judgments in sacrificial dilemmas and antisocial traits such as psychopathy in both clinical (Koenigs et al., 2012) and sub-clinical (Bartels & Pizarro, 2011; Glenn et al., 2010; Kahane et al., 2015; Wiech et al., 2013) populations casts doubt on the assumption. Still other studies have shown a relationship between pro-sacrifice judgments and libertarian political views (Iyer et al. 2012) as well as explicit endorsement of various forms of egoism (Kahane, Everett et al., 2015), both of which give heightened or even exclusive priority to one's own self-interest over the welfare of others. Psychopaths are obviously not paragons of impartial concern for the greater good, and egoists explicitly reject any such concern. These findings directly contradict the strict impartial concern for all people's interests that is at the core of utilitarian theory. Further departure from such impartiality is seen in research suggesting that pro-sacrifice judgments are more likely to be made when they are in the participants' self-interest (Moore, Clark, & Kane, 2008), and in research suggesting that rates of pro-sacrifice judgments are strongly influenced by in-group membership: whether the comparison is between foreigners vs. compatriots (Swann, Gomez,

Dovidio, Hart, & Jetten, 2010), strangers vs. family members (Petrinovich, O'Neill, & Jorgensen, 1993), or even animals vs. humans (Petrinovich et al., 1993), in-group members are more likely to be saved. Such findings suggest that pro-sacrifice judgments in these dilemmas are rarely based in the kind of impartial maximization of aggregate welfare that utilitarianism demands.

In Kahane, Everett et al. (2015) we directly investigated the relationship between a tendency to make pro-sacrifice judgments in sacrificial dilemmas and a wide range of measures of impartial moral concern for the greater good in other contexts. Such measures included: willingness to donate money to reduce the suffering of those in need in poor countries, rejecting favoritism toward one's compatriots over distant strangers, and identifying with the whole of humanity. We consistently found either no relation or a negative relation between pro-sacrifice judgments and such impartial concern for the greater good (Kahane, Everett et al., 2015). These findings suggest that an impartial moral view may actually be in tension with a permissive attitude to instrumental harm. This hypothesis gains some support from other recent research. While psychopathy has been associated with smaller amygdalae, reduced amygdala responses to fear-related stimuli, inferior ability to recognize fearful expressions, (Dawel, O'Kearney, McKone, & Palermo, 2012; Marsh & Blair, 2008) and reduced empathic concern (Decety, Lewis & Cowell, 2015), Recent studies of individuals who donated their kidney to a complete stranger—an extreme form of altruism which is strongly consistent with the positive core of utilitarianism—found that such individuals have large right amygdalae and superior ability to recognize fearful expressions, compared to control subjects (Marsh et al., 2014). Such extreme altruism was also found to be associated with higher empathic concern (Brethel-Haurwitz et al.,

2016) and reduced social discounting towards distant strangers (Vekaria et al., 2017), indicating greater impartiality.

Utilitarian judgments in sacrificial dilemmas clearly do not measure the positive dimension of utilitarianism. In fact, it is not clear that they measure its negative dimension more generally, beyond attitudes to instrumental harm. For example, Kahane et al. (2012) found no association between pro-sacrifice judgments in sacrificial dilemmas and greater willingness to lie when this has overall positive consequences (another paradigmatic utilitarian judgment). It thus seems that rejecting moral rules relating to instrumental harm does not predict the rejection of deontological rules in other domains (e.g. relating to honesty).

These results are not entirely surprising when one considers that while utilitarians and deontologists often do endorse contrasting responses to sacrificial dilemmas, these dilemmas were not designed by philosophers as a way of bringing out the core disagreement between utilitarianism and its opponents. In fact—contrary to what is commonly assumed by many researchers—the 'trolley' scenarios on which sacrificial dilemmas are based were actually introduced as a way of exploring certain issues *within* the deontological approach (see Foot, 1967; Thomson, 1985). Thus, while sacrificial dilemmas were an important first step in studying utilitarian decision-making, and have already yielded valuable findings about attitudes for and against instrumental harm, they need to be supplemented with further tools that allow us to study utilitarian decision-making along both of its dimensions, and that do not rest on the assumption that utilitarian decision-making refers to any unified psychological phenomenon in the everyday context—an assumption that is already put into question by the evidence we have reviewed.

The theoretical framework we propose here makes no such problematic assumption. Importantly, it avoids classifying the moral judgments of non-philosophers as ‘utilitarian’, and does not conceive of proto-utilitarian tendencies in ordinary people in terms of the frequency of such judgments – let alone their frequency in one highly specific moral domain. Instead, our proposed framework understands utilitarian tendencies in terms of the distance of broader patterns of moral dispositions from the paradigmatic concerns of an unqualified utilitarian outlook, leaving open the possibility that one can be more or less utilitarian in some respects yet not in others.

In an important contribution, Gawronski & Conway (2013) have used a process dissociation approach to disentangle two distinct factors that can drive pro-sacrifice judgments: reduced commitment to the deontological principle against directly harming others, and giving greater weight to saving the larger number. This is a significant advance, but it is worth explaining why it does not address the issues we have been raising. Although the two may sound similar, the distinction between reduced deontological commitment and increased weight to consequences in the context of sacrificial dilemmas does not correspond to our distinction between negative and positive dimensions of utilitarianism. This because to give greater weight to saving the five in an emergency situation (and thereby being more willing to sacrifice one) need not indicate any kind of greater impartial concern. Moreover, simply giving greater weight than others to saving a larger number of lives while *still* giving considerable weight to a deontological constraint against direct harm is still to be at some distance from the utilitarian view even within the narrow domain of instrumental harm.

**Existing Individual Difference Scales**

An adequate measure of proto-utilitarian psychological tendencies would need to collect responses that are not confined to a specific moral context. This would be hard to achieve effectively using detailed vignettes along the lines of conventional sacrificial dilemmas, since a great many such vignettes would be needed to cover a wide range of scenarios, and this would be unduly burdensome on both participants and researchers. To the extent that we regard utilitarian decision-making tendencies as a fairly stable feature of individuals, a preferable approach would be to administer a scale that could capture individual differences in such tendencies using shorter items. Several such scales purporting to measure utilitarian (or, more broadly, 'consequentialist') tendencies have already been developed – for example, the Consequentialist Thinking Scale (Piazza & Sousa, 2013), or Robinson's unpublished Consequentialist Scale (Robinson, 2012). However, these have broadly the same limitations as the sacrificial dilemmas we discussed above: as with sacrificial dilemmas, they typically only capture the negative dimension of utilitarianism while ignoring utilitarianism's positive impartial ideal and the practical contexts in which it may be manifested (e.g. situations that pit self-interest or concern for those close to us against the well-being of distant strangers). Moreover, formal scale development procedures such as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) have not been reported for some existing scales (e.g. Piazza & Sousa, 2013). Using empirical, data-driven, models for scale development is particularly important when trying to measure an abstract concept like utilitarianism because it cannot be assumed a priori that what is conceptually unitary in the philosophical context will also be so in the psychology of ordinary people (Kahane, 2015) (see Supplementary Materials for a more extended discussion of existing scales).

## A New Approach to Measuring Proto-Utilitarian Tendencies

Both conceptual considerations and considerable empirical evidence thus strongly suggest that sacrificial dilemmas are a limited basis for studying utilitarian tendencies, and an adequate scale has not yet been developed. To study utilitarian decision-making in the lay population, then, a new measure is required. The theoretical framework we outlined above suggests an alternative approach.

First and foremost, an adequate tool for investigating proto-utilitarian tendencies would need to draw on philosophical expertise to ensure that the psychological construct being measured correctly captures the relevant ethical concepts. But it also needs to make sure that the construct maps on to tendencies in general populations rather than the theoretical views of professional philosophers. Consequently, such a tool would need to be a measure of *degrees* of individual differences in utilitarian tendencies instead of an all or nothing construct. Few individuals in the lay population are likely to be full-fledged utilitarians, but some may be more utilitarian than others, or may be so only on certain dimensions.

Second, in contrast to both sacrificial dilemmas and existing individual differences questionnaires that purport to measure utilitarian or ‘consequentialist’ tendencies (but which similarly focus on willingness to cause harm), such a measure would need to cover both the ‘negative’ and ‘positive’ aspects of utilitarianism. That is, the scale should also cover the degree to which individuals think about morality and the well-being of others in impartial terms, giving no more (or less) moral priority to the self, as well as to those with whom one has close ties.

The present study aimed to develop a new measure of individual differences in utilitarian tendencies that meets the above desiderata. It also sought to use this new measure to study the relationship between utilitarian tendencies and other traits in order to advance our understanding of proto-utilitarian thinking. Since the desiderata include treating utilitarian tendencies as a matter of degree and assessing responses across a range of moral situations (i.e., not only in relation to willingness to violently sacrifice others), we developed a measure more along trait-level individual differences in utilitarian tendencies rather than a measure of an episodic psychological process—that is, a measure of whether or not a subject is engaged in 'utilitarian decision-making' at a given point in time. In other words, we are primarily interested in utilitarian tendencies as a kind of individual difference akin to empathic concern (Davies, 1980) or the degree of endorsement of core moral values measured by the Moral Foundations Questionnaire (Graham et al., 2011). This as opposed to thinking of such tendencies as a state (broadly, a specific reaction to a specific moral context). Given the need to measure responses across a range of moral situations and dimensions, a battery of detailed moral dilemmas would be long and cumbersome. We therefore opted for a short scale measuring responses to a list of brief items. Finally, in order to ensure that the proposed scale adequately reflects the relevant philosophical concepts and theories, this scale was developed by a joint team of psychologists and moral philosophers. We formulated the pool of items on which the scale is based by conducting a systematic review of the ethical literature; these items were then vetted by leading professional philosophers in the USA and UK, both utilitarian and non-utilitarian.

## Scale Development Procedure

### Overview

To develop and validate our scale, we set in advance and then followed a formal scale development procedure to ensure that our measure was both reliable and valid. First, we created an initial pool of items based on the existing literature on the target construct of utilitarianism. After paring the pool down by, e.g., eliminating redundancies or unclear items, we submitted the items to be reviewed by an expert panel of academic philosophers and then modified the items in response to their feedback. Third, we recruited a large sample of participants to complete the revised pool of items and then conducted a series of exploratory factor analyses (EFA) to obtain the best factor structure with the best items. Fourth, we conducted a series of confirmatory factor analyses (CFA) to evaluate and refine the best factor structure. Fifth, we recruited a new sample of participants to complete the items determined by the CFA and then we confirmed, using this new sample, that the measure had appropriate factor structure and psychometric properties. Sixth, we confirmed that the data fit this model and factor structure better than alternative models (e.g., that a multi-dimensional model obtained from the CFA accounted for the data better than a one-dimensional model, or vice versa). Seventh, we explored construct validity by testing how scores on the final scale obtained from the previous steps were connected to other established measures. Finally, based on helpful feedback from reviewers of a previous draft of this paper, we investigated external validity by administering the scale to an expert sample of graduate students and academics specializing in moral philosophy. A more extended account of the scale development process can be found in the Supplementary Materials; in the interests of brevity we report only essential information in the main paper.

**Item Generation**

An initial pool of items was generated through a comprehensive survey of the existing literature on utilitarianism. In creating the initial pool of items, several considerations were taken into account. First, we judged it necessary to include items that tapped into the abstract tenets of utilitarianism as well as items that bore on real-world moral judgments that track utilitarian thinking. Second, it was essential to include items that captured both the positive and negative components of utilitarianism: namely, that the right act is the one that impartially maximizes the greater good (positive component), and that this maximization is all there is to morality such that deontological rules and constraints must be rejected when they stand in the way of achieving this goal (negative component). Moreover, we hoped to have a range of items that included: (a) abstract statements of utilitarian belief (b) anti-utilitarian views, (c) items reflecting the application of utilitarianism to concrete contexts, and (d) items briefly stating seminal examples or illustrations used by both critics and defenders of utilitarianism. Finally, the items all involved moral judgments of various kinds. For simplicity of presentation, these were phrased using a range of explicit normative terms with synonymous or closely overlapping content such as what is ‘right’, ‘required’ or what we ‘should do’ or are ‘obliged to do’ (prior studies suggest that such minor variation in wording has little or no effect on responses; O’Hara, Sinnott-Armstrong, & Sinnott-Armstrong, 2010). When an item could also be interpreted in legal terms, we made it explicit that the question was concerned with what is morally right or wrong rather than legally right or wrong—i.e., with what we should do from a moral point of view.

We employed both a “bottom-up” and a “top-down” approach to identifying relevant items. The “bottom-up” approach involved compiling sources from the existing empirical

literature purporting to measure utilitarian judgments and extracting relevant citations from their references sections in a systematic fashion (e.g., Piazza & Sousa, 2013; Robinson, 2012; Greene et al., 2001; Moore, Clark, & Kane, 2008). This resulted in over 200 items, mostly in the form of vignettes or short statements describing moral dilemmas. An initial review of these items revealed considerable redundancy in terms of theoretical content, with the majority of cases clustering around variants of the well-known "push" and "switch" dilemmas (Foot, 1967; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Thomson, 1985). We therefore deemed it necessary to perform a "top-down" analysis, as well, to ensure a more robust theoretical foundation. For this analysis, we drew on the philosophical literature as well as the expertise of professional moral philosophers—including members of the present research team—and reviewed both classical and more recent discussions of utilitarianism. Classical statements of the theory included those by Bentham (1789/1983), Mill (1863), and Sidgwick (1901); important recent contributions included work by, for example, Smart and Williams (1973), along with critiques (e.g., Rawls, 1971) and defenses (e.g., Kagan, 1989) of utilitarianism or consequentialism more generally, as well as influential further developments of the theory (e.g., Parfit, 1984). We also took care to include works focusing on the practical implications of utilitarianism (e.g., Singer, 1993). Special emphasis was put on identifying key points of conflict between a strict utilitarian approach and commonsense morality as well as competing ethical theories—conflicts that includes willingness to cause harm, break moral rules, compromise virtue or integrity, limits to the demandingness of morality, and so forth. Finally, after filtering out major redundancies between items, irrelevant items, and poorly-worded or confusing items,

we were left with a smaller pool of 93 items that were then edited for theoretical clarity and ease of understanding.

**Expert Review**

Having generated this pool of 94 items based on initial assessments from within the research team, we then recruited an external panel of leading experts to review the items. Our panel consisted of 11 professional philosophers working in ethics or moral philosophy who had a diversity of viewpoints (including classical act utilitarians, consequentialists who depart from classical utilitarian views, and ethicists who reject consequentialism in all of its forms). Our expert panel included some of the most prominent living contributors to the philosophical literature on utilitarianism, e.g., Peter Singer, Shelly Kagan, John Broome, and Alistair Norcross. To ensure that our final set of items would be intelligible to non-philosophers, we asked our experts to evaluate the items in terms of their brevity, simplicity, and accessibility, while maintaining theoretical specificity. At the same time, we explained that the planned scale was intended to measure utilitarian tendencies in non-philosophers, such that highly subtle philosophical distinctions that were unlikely to be relevant in such a context should not be emphasized in the assessment of items (see Supplementary Materials for actual text and instructions). In an online survey, each expert was given a list of all 94 statements and asked to indicate "How good do you think this item is for discriminating utilitarian and non-utilitarian views?" (*1 = not at all; 5 = very much*)[3], with space for written comments. Our interest was primarily in the experts' qualitative feedback and comments on the items, but we also collected the numerical ratings as a complementary source of data. We then removed or modified items to incorporate the experts' philosophical insights on a case-by-case basis, carefully considering

their written comments and their corresponding numerical ratings. We opted for a relatively inclusive pool of items, only dropping items that the experts found particularly unhelpful to discriminating utilitarian and non-utilitarian views. The items that were retained had a mean rating of 3.52 and those that were dropped a mean rating of 2.96. In a substantial number of cases a lower rating reflected 'fixable' concerns about the wording of an item rather intractable concerns than about its underlying content; in those cases we opted to revise the items rather than to drop them. The specific suggestions of the experts were discussed within the research team until a consensus on exclusion or revised wording was achieved.

This resulted in a smaller pool of 77 items with which to conduct the next stage of the scale development: the exploratory factor analysis in a lay sample. It is worth noting that although the experts were given space to propose new items for inclusion in our scale, none was suggested, indicating that the 94 items provided good coverage of the relevant moral issues. And while the numerical data was not our focus, all of the original items that the experts rated and which were included in the final scale (including modification, if necessary) were rated above the mid-point of the scale, suggesting that they were good items for discriminating utilitarian and non-utilitarian views.

## Study 1

### Ethics Statement

Relevant ethical guidelines were followed and the research was approved through University of Oxford's Central University Research Ethics Committee, with the reference

number MSD-R50145/RE001. Written informed consent was obtained electronically from all participants.

**Participants and Procedure**

1009 participants completed the survey online using Amazon Mechanical Turk (MTurk). Participants were excluded from analysis if they completed the survey more than once (in which case only their first attempt was included), or if they failed one or more of five simple attention checks embedded amongst the items requiring them to "Please click scale point X to confirm you are paying attention". This left a final sample of 960 participants (489 female, $M_{age}$ = 35, *SD* = 12.11), of whom the majority of participants had attended college or higher education (81%). Participants were given the list of 77 items that were chosen after being subjected to expert review, and these were presented in a semi-randomized order. For each item, participants were asked to "indicate how much you agree or disagree with each of the following statements" (*1 = strongly disagree, 4 = neither agree nor disagree, 7 = strongly agree*).

Our final sample size of 960 was more than adequate. Compared to experimental designs and statistical techniques such as ANOVA where one can compute the required power given the effect size, determination of sample size for factor analysis is notoriously tricky (Mundfrom, Shaw, & Ke, 2005). One approach is to focus on the absolute sample size. While some have suggested a minimum sample size of 250 (Cattell, 1978) or even 100 (Gorsuch, 1983), more recent estimates have suggested that a good sample size is at least 300, and that 1,000 or more is excellent (Comrey & Lee, 1992). Another approach is to focus more on the subject-to-item ratio, or the number of participants for each item used. It is typically accepted that a subject-to-item ratio should be no lower than 5:1 (Gorsuch, 1983), and that 10:1 is appropriate (Everitt, 1975;

Nunnally, 1978). Based on these considerations, we therefore decided to recruit 1000 participants. This represented an excellent absolute sample size, and (with 77 items), gave us a good final subject-to-item ratio of 12:1.

**Exploratory Factor Analysis**

We first performed a series of exploratory factor analyses (EFA) on the full set of items (77 in total) to ascertain the underlying factor structure. For factor extraction, we used both Kaiser's criterion (eigenvalues set to 1) in conjunction with inspecting the scree plots. We did this because Kaiser's criterion tends to over-extract factors when the number of variables is large (Linn, 1968). For all EFAs, we used the maximum likelihood estimator with direct oblimin rotation. The first EFA, according to Kaiser's criterion, yielded 19 factors explaining 57% of the variance in the 77 items whereas the scree plot indicated a 6 factor solution. Inspecting the rotated factor matrix indicated that a number of items did not load onto any factor, as indicated by a factor loading falling under the .30 mark. We excluded these 13 items and reran the EFA. This second EFA indicated that a further 5 items did not load significantly onto any factor. We repeated this process a total of 10 times. The $10^{th}$ EFA yielded a four-factor solution that explained 43% of the variance in all the items. See Table 1 for the factor solution and for the reliabilities.

**Confirmatory Factor Analysis**

Next, using *R*, we performed a Confirmatory Factor Analysis (CFA) on the items obtained from the initial EFA structure (see Table 2 for items). For the CFA, we used the maximum likelihood estimator with robust standard errors and a combination of fit indices to adjudge model fit. The chi-squared test is the classic test used in factor analysis, and indicates the

difference between the observed and expected covariance matrices. Unfortunately, the chi-squared test is very sensitive to sample size, with problems of false-negatives in small sample sizes and false-positives in large sample sizes. Moreover, because there can be concerns with using only one indicator of model fit (if some indices are more favorable than others, then this could be the only one reported) we decided before data collection to focus our attention on three other indices that assess model fit and which improve on the chi-square test (Bentler, 2007; Hu & Bentler, 1999; Kenny, 2015) First, we used root mean square error of approximation (RMSEA). This an absolute measure of fit that again adjusts for sample size when chi-squared tests are used. The RMSEA yields values ranging from 0 to 1, with smaller values indicating better fit and $\leqslant$ .05 indicating good fit. While some have suggested a cut-off for poor models of $\leqslant$ .10, we took a more conservative cut-off of $\leqslant$ .08 to indicate minimally acceptable fit (Kenny, 2015). Second, we used the standardized root mean square residual (SRMR), which is again an absolute measure of fit and represents the square root of the discrepancy between the observed covariance matrix and the hypothesised covariance matrix. Values range from 0 to 1, with smaller values indicating better fit and a value of $\leqslant$ .08 indicating acceptable fit. Third, we used the comparative fit index (CFI): a relative fit index that examines the discrepancy between the actual data and the hypothesized model, making adjustment for the issues of sample size that can be problematic in the chi-squared test of model fit. Like RMSEA and SRMR, CFI yields values from 0 to 1, but unlike the others, for the CFI it is higher values that indicate better fit. Traditionally, CFI scores of $\geqslant$ .90 have been taken to indicate acceptable fit, but this has since been revised to suggest that this is too lenient and that for good fit, scores should be $\geqslant$.95 (Hu & Bentler, 1999) It is this more conservative level we settled upon to use. Note that all of these are

recommendations and an overall decision should be made through considering the results across the different indices, but we provisionally agreed upon the following cut-off criteria as indicative of adequate model fit: RMSEA $\leqslant$ .08, SRMR $\leqslant$ .08 and CFI $\geqslant$ .95 (Bentler, 2007; Hu & Bentler, 1999; Kenny, 2015).

We first performed the CFAs for the four above factors individually, and then, once all items were finalized, we entered the factors into a CFA simultaneously. We had two main aims in our CFA analyses: to find a factor structure that would give the best statistical fit, but also to produce a short scale.

**Individual Factor CFAs**

***Factor 1 ('Impartial Beneficence')***

Factor 1 contained 11 items that converged around the concept of impartial beneficence, consisting of items such as "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 CFA for Factor 1 with 11 items returned fit statistics that were less than adequate, $\chi^2$ (44) = 401.52, $p$ < .001, CFI = .86, RMSEA = .09 [.08, .10], SRMR = .06. In order to increase factor stability, we decided to exclude all items with factor loadings lower than .40. As a result, item 63 (standardized factor loading = .39) and 48 (standardized factor loading = .39) were dropped. Dropping these two items resulted in a model that did fit better, $\chi^2$ (36) = 258.33, $p$ < .001, CFI = .89, RMSEA = .09 [.08, .11], SRMR = .05, but which still did not meet our cut-off criteria. Looking over the model modification indices, there was evidence of correlated error variances between items 77 and 73. Both items 73 and 77 seemed to measure rejection of the Doctrine of Double Effect, but because item 77 was much longer and potentially confusing, we decided to

remove item 77. The resultant modification indices showed acceptable fit, $\chi^2$ (20) = 163.69, $p$ < .001, CFI = .92, RMSEA = .09 [.08, .10], SRMR = .05. We next removed items 14 and 15 on theoretical grounds: item 14 was similar to, but less precise than, item 16; and item 15 was not theoretically critical since questions about self-defense are not central to debates about utilitarianism. After deleting these two items, the model showed acceptable model fit, $\chi^2$ (9) = 83.69, $p$ < .001, CFI = .93, RMSEA = .09 [.08, .11], SRMR = .05. Items 61 and 62 both concerned impartiality in helping those close to us, but because item 61 had the weakest loading and item 62 was better as an abstract statement of impartiality, we next deleted item 61, which substantially improved model fit, $\chi^2$ (5) = 26.83, $p$ < .001, CFI = .97, RMSEA = .07 [.04, .09], SRMR = .03. Therefore, items 16, 17, 11, 73, and 62 were included in the final CFA for the first construct. Because these items seemed to tap into the aspect of utilitarianism that seeks to impartially maximise welfare and the greater good, even at expense to oneself, we labelled this factor *Impartial Beneficence*.

***Factor 2 ('Instrumental Harm')***

Next, we conducted a CFA on Factor 2. Factor 2 consisted of 9 items that seemed to tap the construct of willingness to endorse instrumental harm, including items such as "It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people". The model with all 9 items showed acceptable model fit, $\chi^2$ (27) = 194.79, $p$ < .001, CFI = .92, RMSEA = .08 [.07, .09], SRMR = .05. Because item 70 had fairly low loadings, we removed it from the next CFA, $\chi^2$ (20) = 181.74, $p$ < .001, CFI = .92, RMSEA = .09 [.08, .10], SRMR = .06. Next, we noted that the error variance for item 59 was correlated with the error variances of three other items (67, 20, and 52), indicating that there was another factor

that was not accounted for by our model which explained a significant amount of variance between these three items. In order to gain as clean a factor structure as possible, we next removed item 59 from the CFA. This resulted in an improvement in the model fit, $\chi^2$ (14) = 123.01, $p < .001$, CFI = .94, RMSEA = .09 [.08, .11], SRMR = .05. Items 43 and 20 still shared a significant amount of error variance and so once again, to obtain as clean a factor structure as possible, we deleted item 42, which was theoretically less fundamental because it referred to political restriction of freedom rather than to causing acute harm. The resultant six-item CFA showed good model fit, $\chi^2$ (9) =50.15, $p < .001$, CFI = .98, RMSEA = .07 [.05, .09], SRMR = .03. We remained concerned, however, that two items (68 and 69) in the model were too similar in both content and language and so we next excluded the weaker loading item. This five item factor-solution showed good model fit: $\chi^2$ (5) =26.87, $p < .001$, CFI = .98, RMSEA = .07 [.04, .09], SRMR = .03. Finally, we were concerned that item 52 – the weakest loading – included elements relating both to instrumental harm and to impartiality and was therefore somewhat removed from the other items. Indeed, deleting this item resulted in a four-item solution with excellent fit, $\chi^2$ (2) =4.47, $p = .11$, CFI = .10, RMSEA = .04 [.00, .08], SRMR = .01. Therefore, items 20, 67, 68, and 72 were included in the final CFA for the second factor. Because the second factor tapped support of allowing harm in the service of the greater good, we labeled it *Instrumental Harm*.

***Factor 3 ('Anti-Traditional Morality')***

Factor 3 consisted of 10 items that were concerned with traditional morality—a set of deontological ideas associated with conservative thought, such as retributive punishment, sexual morality, human dignity, and an absolutist view of social and moral rules. An example of an item

from this factor is “Criminals should receive the punishment they deserve—even if this will not protect the public or deter crime in the future.” The CFA for the third factor with 10 items showed fairly poor model fit, $\chi^2$ (35) =417.18, $p < .001$, $\chi^2/df$ = 9.81, CFI = .83, RMSEA = .11 [.10, .12], SRMR = .07. Looking at the factor loadings, items 21 and 23 had very low factor loadings (both = .25), and so we removed them from the model. Removing these two items improved model fit, although the model fit was still sub-standard, $\chi^2$ (20) = 252.74, $p < .001$, CFI = .88, RMSEA = .11 [.10, .12], SRMR = .06. The modification indices highlighted a strong correlation between the error variances for items 27 and 26. We decided to remove the item with more error variance, which meant dropping item 26. Dropping this item improved the model fit, $\chi^2$ (14) = 92.07, $p < .001$, CFI = .94, RMSEA = .08 [.06, .09], SRMR = .04. This left us with seven items. Given that the previous two factors consisted of five and four items respectively, in order to more closely match the factors, we deleted the theoretically weakest item from the CFA for Factor 3. This resulted in removing item 28, which focused more on political authority than morality. The resulting model based on six items fit the data well, $\chi^2$ (9) = 49.62, $p < .001$, CFI = .96, RMSEA = .07 [.05, .09], SRMR = .03. This third factor tapped the rejection of traditional deontological moralities and so we labeled this *Anti-Traditional Morality*.

***Factor 4 (‘Truth-Telling and Promise-Keeping’)***

Finally, we looked at Factor 4, which consisted of 5 items concerning truth-telling and promise keeping, such as “It is morally wrong to break promises even if this would bring about good outcomes” and “It is morally permissible to lie if doing so would help others a great deal”. In addition to explaining the least variance, Factor 4 also represented a group of moral views that are not distinctive of utilitarianism given that few hold that it is never permitted to lie or break

promises, limiting its usefulness to the scale. Nonetheless, for completeness we again conducted a CFA. The first CFA with all 5 items showed weak model fit, $\chi^2$ (5) =181.94, $p$ < .001, CFI = .88, RMSEA = .19 [.17, .23], SRMR = .07. Because item 46 shared a lot of variance with items 45 and 47 we then excluded this item. This four-item solution showed better fit, but was still suboptimal, $\chi^2$ (2) =30.77, $p$ < .001, CFI = .98, RMSEA = .12 [.09, .16], SRMR = .03. Given both prior theoretical concerns (issues relating to honesty and promise-keeping are not central, or even particularly important, to utilitarianism) and the sub-optimal results from the CFA, we therefore decided to not include this factor in the final scale.

**Overall Factor Solutions**

We first tested a three-factor solution by conducting a CFA with all three factors and their corresponding manifest items entered into the model simultaneously (Factor 1: items 11,16,17,73,62; Factor 2: items 20,67,68,72; Factor 3: items 2,4,18,25,27,57) The resulting model fit was acceptable, but not ideal, $\chi^2$ (87) = 352.17, $p$ < .001, CFI = .91, RMSEA = .06 [.05, .06], SRMR = .06. Given this, we compared this factor model to one where all 15 items loaded onto a single factor. This model produced a significantly worse model fit, $\chi^2$ (90) = 2256.24, $p$ < .001, CFI = .29, RMSEA = .16 [.15, .16], SRMR = .16. This empirically supported our theoretical basis for believing that a multi-factor solution would best characterize utilitarian tendencies in the lay population. At the same time, these results suggested that a three-factor might not be the most appropriate.

Given the less than ideal model fit with a three-factor solution, we considered whether a two-factor solution might be more appropriate. We discuss the theoretical grounding and strength of our final scale below, but – briefly - recall that the key aims of our scale were for it to: a) be

theoretically grounded, b) be empirically validated, c) be short, and d) capture both the positive and negative components of utilitarianism. To this end, a shorter two-factor solution that taps these positive and negative components is just as appropriate as a three-factor solution that taps these plus something else. Also note that the first two subscales (Impartial Beneficence and Instrumental Harm) directly and non-controversially tap core aspects of utilitarianism – which is, of course, the central construct measured here – and moreover that these two factors closely match with the arguments and data advanced by Kahane, Everett et al. (2015). The theoretical basis of the third potential factor (Anti-Traditional Morality), in contrast, is more an indicator of 'what utilitarianism is *not'*, rather than 'what utilitarianism *is'*. Thus, the third factor might represent an opposition to utilitarianism from a particular historical or social context. Consider an analogy. Imagine that instead of measuring utilitarianism, we wished to develop a scale measuring support of communist ideology. We might obtain three factors: two subscales tapping what communism *is*, and then a third subscale that measured endorsement of capitalism – that is, what communism is *not*. While we could reverse-code this capitalism measure and it might relate to the communism subscales, it isn't clear that measuring lack of a support for capitalism directly taps support of communism. Similarly, consider that the anti-traditional morality factor largely taps the rejection of deontological absolutism and traditional moral rules (e.g. those relating to punishment or sexual purity). While such views are very distant from utilitarianism, this is not of itself diagnostic in distinguishing utilitarians from non-utilitarians because such extreme deontological views are also rejected by many non-utilitarian views, such as right-based liberalism and libertarianism. Indeed, that the three-factor solution showed poor fit indicates that

measuring absolutism as the key contrast with utilitarianism (as previous scales do) may not actually be useful for measuring genuine utilitarian tendencies.

There are strong theoretical grounds to prefer a two-factor utilitarianism scale over a three-factor one which included an additional factor assessing support for traditional moralities. Moreover, our analyses showed that there are strong empirical grounds as well. In contrast to the poor fit observed for the three-factor solution, our two-factor solution fit the data much better. As well as each factor individually having excellent fit, the overall two-factor solution had excellent fit: $\chi^2(26) = 84.91$, $p < .001$, CFI = .97, RMSEA = .05 [.04, .06], SRMR = .04 (see Table 3 for factor loadings and Table 4 for recommended and actual fit indices). We therefore decided to use this two-factor solution as a provisional final scale.

## Study 2: Scale Validation

In Study 1 we established a provisional 2-factor *Oxford Utilitarianism Scale* (OUS). In Study 2 we sought to confirm – and if needed, refine – this scale. To confirm the factor structure and establish contrast validity, we recruited a new set of participants to complete the scale and a number of theoretically related constructs. This allowed us to perform a second CFA using the new dataset, and to examine how well scores on the OUS related to existing measures of Utilitarianism.

### Participants and Procedure

300 participants were recruited online using Amazon Mechanical Turk. Eighteen participants were excluded from analysis for answering a simple attention check incorrectly or not completing the survey, leaving a final sample of 282 participants (178 female, $M_{age}$ = 39, *SD*

= 12.66). The majority of participants had attended college or higher education (80%). As before, participants rated how much they agreed or disagreed with the statements (*1 = strongly disagree*, *7 = strongly agree*). Participants completed the provisional OUS first (including the anti-traditional items, for completeness) and then moved on to complete a series of other measures. These measures were designed to assess both construct validity and to show how the OUS can shed light on previously found relationships between individual differences and (so-called) utilitarianism. In order to prevent order effects, these other measures were presented in a randomized order, with demographics and questions on political and religious belief at the end.

**Confirmatory Factor Analysis**

The two-factor CFA model showed excellent fit: $\chi^2$ (26) = 39.55, $p$ = .04, CFI = .98, RMSEA = .04 [.01, .07], SRMR = .04 (see Table 3 for factor loadings and Table 4 for recommended and actual fit indices). Looking again at the factors separately, the first factor showed very good model fit, $\chi^2$ (9) = 49.62, $p$ < .001, CFI = .96, RMSEA = .07 [.05, .09], SRMR = .03. While the first factor (unlike the first study) was just over the criterion we had initially set on the RMSEA, the other fit indices showed excellent fit, therefore confirming the validity of the factor. The second factor showed excellent fit, $\chi^2$ (2) = 2.56, $p$ = .17, CFI = .99, RMSEA = .05 [.00, .14], SRMR = .02. Thus, for theoretical and empirical reasons we settled on a two-factor Oxford Utilitarianism Scale (OUS) (see Table 5 for final items). The first subscale - *Impartial Beneficence* (OUS-IB) - consisted of 5 items that all reflected endorsement of impartially maximizing the good.[4] In contrast, the second subscale - *Instrumental Harm* (OUS-IH) - consisted of 4 items that all reflected a willingness to cause harm in order to bring about such maximization (see Table 6 for correlations between the subscales).

**The Oxford Utilitarianism Scale (OUS)**

Having assessed the scale on both theoretical and empirical grounds, we therefore arrive at a final scale consisting of 9 items in two subscales. The first subscale - *Impartial Beneficence* (OUS-IB) - consists of 5 items that all tap endorsement of the impartial maximization of the greater good, even at the cost of personal self-sacrifice:

1. If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.
2. From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.
3. From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.
4. It is just as wrong to fail to help someone as it is to actively harm them yourself.
5. 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 second subscale was labeled *Instrumental Harm* (OUS-IH). This subscale consists of 4 items that all tap a willingness to cause harm in order to bring about the greater good:

1. It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.
2. If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.
3. It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.
4. Sometimes it is morally necessary for innocent people to die as collateral damage—if more people are saved overall.

**Construct Validity**

For the final stage of our scale development procedure, we used the same dataset as used for the second CFA ($N$ = 282) to assess the construct validity of the OUS. Since 'proto-utilitarian tendencies' is a new construct and existing measures are flawed, there is no straightforward way to validate the scale. We decided to assess construct validity with three key tests assessing convergent validity (i.e. whether things that are theoretically connected are in fact connected in the data). Specifically, if our OUS scale is measuring what it should be measuring, (1) overall OUS scores should be associated with the extent of agreement with an explicit statement of a utilitarian approach to ethics; (2) OUS-IH scores should be associated with responses in sacrificial moral dilemmas; and (3) OUS-IB scores should be associated with responses in 'greater good' moral dilemmas that capture participants' endorsement of self-sacrifice and impartiality in morality. Note that our focus here is primarily on convergent rather than divergent

validity. We have of course developed an overall OUS scale and because scores on the two subscales are positively but weakly correlated, it would not be entirely unsurprising if - in addition to the primary relationship between the OUS-IB and the 'greater good' dilemmas - there is also a weak positive relationship between the OUS-IH and these dilemmas. What is critical for our 2D model is not that there is absolutely no relationship between instrumental harm and impartial beneficence, but that they are independent and dissociable factors.

We have focused our assessment of contrast validity by comparing the OUS with other measures that have been claimed, however accurately, to directly measure utilitarianism. That said, it is prudent to note that the distinction we draw below between measures directly assessing construct validity and those showing how it works in practice could be seen as somewhat arbitrary, given that nearly all previous psychological work on utilitarianism has ignored impartial beneficence. Both sets of measures are important and meaningful in confirming the scale's validity, but for the specific purpose of assessing construct validity we begin by looking at purported measures of utilitarianism rather than associated constructs.

**Explicit Utilitarianism**

Our first key test of construct validity was to ensure that scores on the OUS were associated with agreement of an explicitly utilitarian moral philosophy. Participants were asked to indicate how much they agreed or disagreed with the following statement:

> "The only thing that determines whether an act is morally right is whether, out of the available options, it is the act that would lead to the most happiness and the least suffering in the world, taking into account the welfare of all sentient beings, whether human or animal. An act that doesn't maximize welfare in this way is morally wrong. On

> this moral view, no one counts for more than anyone else: our own interests and needs, and the interests and needs of our family and friends, never count for more than the interests and needs of any other person, however distant from us. Finally, on this view the only thing that matters is how our actions affect the amount of happiness in the world. It is always morally right to break a rule or principle if doing so would lead to the better outcome."

Note that this explicit description of utilitarianism is well-grounded theoretically and the one that we took as our standard when developing this scale and thus, to the extent that scores on the OUS reflect endorsement of specifically utilitarian moral decisions, they should correlate positively with agreement of this statement of explicit utilitarianism. This was the case. Agreement with the statement of explicit utilitarianism was associated with higher scores on the OUS overall ($r$ = .35, $p$ < .001), and for each of the two subscales: the Impartial Beneficence subscale ($r$ = .37, $p$ < .001) and, more weakly, the Instrumental Harm subscale ($r$ = .16, $p$ < .001). A test of significance between these two correlations (Steiger, 1980) showed that these correlations were significantly different ($Z$ = 2.84, $p$ = .005), providing empirical support for our theoretically grounded claim that it is impartial beneficence rather than instrumental harm that is at the core of a distinctively utilitarian outlook[5].

**Sacrificial Moral Dilemmas**

Our first key test was to look at how overall OUS scores were associated with a statement of explicit utilitarianism. Our next two tests were focused more on providing convergent validity for the two subscales in isolation, starting with responses to 'trolley-style' sacrificial moral dilemmas. Although there are serious concerns about identifying utilitarian judgments with pro-

sacrifice judgments in such dilemmas (Kahane, 2015), such sacrificial dilemmas do track closely the endorsement of instrumental harm. To the extent that our Instrumental Harm subscale directly measures this negative component of utilitarianism - a willingness to cause harm for the greater good – there should be a significant, positive association between scores on the OUS-IH and the OUS overall with endorsement of the sacrificial action in these trolley-style sacrificial dilemmas.

To assess the convergent validity of the OUS through the association with sacrificial dilemmas, we used three dilemmas involving 'up-close-and personal' harm that were adapted from previous research (Moore et al., 2008). The dilemmas we used included, and were inspired by, the classic Footbridge case, in which one can save five people from a runaway trolley only by pushing another person onto the tracks, leading to their death (see Supplementary Materials for a full description). For each of the three dilemmas, participants were asked "How wrong would it be to [perform the 'utilitarian' act, e.g. push the stranger in the Footbridge case]?' (*1 = not at all wrong, 7 = extremely wrong*). These ratings across the three dilemmas were combined into a single reliable measure of participants' 'utilitarianism' (viz., endorsement of instrumental harm) in sacrificial dilemmas ($\alpha$ = .74), where lower scores indicated higher 'utilitarianism'. Confirming the validity of the OUS, participants who scored higher on the OUS-IH subscale ($r$ = -.32, $p < .001$) and the OUS overall ($r$ = -.34, $p < .001$) were more likely to endorse the sacrifice in these dilemmas.

Recall that our focus here was on the convergent validity of the OUS-IH subscale with responses in sacrificial dilemmas, because the OUS-IH is supposed to measure instrumental harm and this is also measured in sacrificial dilemmas. It was less important to determine, but

nonetheless interesting, if scores on the OUS-IB were associated (positively or negative) with responses to these dilemmas. In fact, as suggested by the correlation between the dilemmas and the overall OUS, participants who scored higher on the OUS-IB were more likely to endorse the sacrificial option in the dilemmas ($r$ = -.21, $p$ < .001), and the strength of correlations between the dilemmas and the two subscales were not significantly different ($Z$ = -1.48, $p$ = .14).

This significant association between pro-sacrifice judgments and the OUS-IB is surprising given that that in some of our previous work (Kahane, Everett et al., 2015) we found no relation between such judgments and a battery of 'greater good' dilemmas, a vignette-based measure of the 'positive' dimension of utilitarianism (see below). One possibility is that this is a result of our scale development procedure in which we sought to obtain a factor solution that had good fit for both the IB and IH items and which cohered into a reliable scale. This procedure will naturally have led to the items in both subscales being more compatible (or, at least, less directly in opposition) than they would have been had we been interested only in the two factors individually and not as an overall measure of utilitarianism. Another possibility is that the first-person presentation of the dilemmas in the previous (e.g. 'should you push the man?'), while for consistency all dilemmas here were presented in the third person (e.g., 'should Adam push the man'). Finally, since participants were asked to evaluate the wrongfulness of sacrificial acts rather than categorically endorse or reject them, our data do not distinguish between two possible explanation of this association: (i) rejection of an absolute deontological prohibition against such acts, and thereby regarding such sacrificial acts either as significantly but not completely wrong or as merely permissible and (ii) regarding such acts as not wrong in any way, or even as required. It is only the latter than we should expect to be specifically associated with

Instrumental Harm but not Impartial Beneficence. Further research is needed to test this hypothesis.

**‘Greater Good’ Moral Dilemmas**

Our final key test of construct validity was to establish the convergence between OUS-IB and overall OUS scores with responses to dilemmas tapping participants’ endorsement of self-sacrifice and impartiality in morality. To accomplish this, we used three ‘greater good’ dilemmas designed by Kahane, Everett et al. (2015) to directly pit an explicit utilitarian action promoting the greater good against a narrower, more partial moral view that allows us to give priority to self, family, and country. For each item participants were asked to rate how wrong it would be for someone to perform the *non*-utilitarian action (*1=not at all wrong; 7 = very wrong*), such that higher scores indicated greater utilitarianism (in contrast to the sacrificial dilemmas). Confirming our construct validity, endorsement of the utilitarian action in these Greater Good dilemmas was indeed significantly associated with scores on the OUS-IB subscale ($r$ = .50, $p$ < .001) and with OUS scores overall ($r$ = .40, $p$ < 001).

We also looked at the correlation between responses to the greater good dilemmas and the other subscale. Interestingly, there was no relationship between scores on the OUS-IH subscale ($r$ = .07, $p$ = .26) and responses to the greater good dilemmas, and the correlations between the OUS-IB and OUS-IH with the dilemmas were significantly different ($Z$ = 6.04, $p$ < .001). Thus, while sacrificial judgments were correlated with scores on the OUS-IB, more concrete cases of impartial beneficence were not associated with instrumental harm. These results again highlight that while there is a connection between the utilitarian components of instrumental harm and impartial beneficence, these often have relatively distinct psychological correlates.

Finally, it is important to note that of the 5 items in the OUS-IB subscale, three are concerned with self-sacrifice for the greater good (1, 2, 5), one with impartiality with respect to strangers vs. those close to you (3), and one with the act/omission distinction (4). The three greater good dilemmas we used here covered self-sacrifice to aid distant strangers, self-sacrifice to prevent harm to animals, and refusal to give priority to one's own country in the context of charitable donation. The OUS-IB subscale was significantly correlated with all 3 dilemmas (*r*s > .31, *p*s < .001), suggesting that it measures both the element of self-sacrifice and the rejection of special duties to those closely associated with you.

## Study 3: Expert Validation

In Study 2, we established construct validity by showing that both the OUS overall and its two subscales were strongly correlated with the degree to which participants endorsed an explicit statement of utilitarianism and that each of the subscales was associated in the appropriate direction with more specific measures of the positive and negative dimensions of utilitarianism (e.g. the OUS-IH with pro-sacrifice responses to sacrificial dilemmas and the OUS-IB with more impartial responses in the greater good dilemmas). Next, to provide further validation for the construct, we compared OUS scores with the self-reported moral views of experts in moral philosophy. To the extent that the OUS is a reliable measure of utilitarian tendencies, it should strongly correlate with the degree to which such experts self-identify as utilitarian. Notice, however, that the OUS was designed as a measure of the moral outlook of ordinary people, not as a tool enabling fine-grained classification of the ethical views of moral philosophers.

Furthermore, collecting responses from experts also allowed us to begin to investigate another question. While both the OUS-IB and OUS-IH were strongly correlated with explicit endorsement of utilitarianism ($r$ = .81 and $r$ = .70, respectively), they were only weakly associated with each other ($r$ = .14), supporting our hypothesis that these two dimensions of utilitarianism, while conceptually linked in the philosophical level, are nevertheless independent factors in the psychology of ordinary people. We predicted, however, that the two sub-scales would be more strongly correlated in an expert sample. This is because, among other factors, formal education in philosophy may create pressure for greater coherence in one's views (Holyoak & Powell, 2016), leading students of philosophy to revise their moral views over time to align more closely with dominant philosophical views (whether utilitarianism or other).

## Method

### Participants

Our sample of experts in moral philosophy were recruited through an email sent to the following mail-lists associated with ethics and moral philosophy: the Future of Humanity Institute, the Centre for Effective Altruism, Ethox Centre, and Uehiro Centre for Practical Ethics at the University of Oxford; the Anscombe Bioethics Centre for Medical Ethics; the Bioethics Centre at Monash University; the Hastings Center for Bioethics; and the Yeoh Tiong Lay Centre for Politics, Philosophy & Law at King's College London, and BioEdge.org. These groups were purposely selected to consist of people who work extensively on questions of ethics and moral philosophy, and care was taken to include both utilitarian-leaning groups (e.g. the Future of Humanity Institute and the Centre for Effective Altruism) and anti-utilitarian-leaning groups (e.g. the conservative Anscombe Bioethics Centre and Bioedge), as well as groups that focus more on

the applied aspects of moral philosophy (e.g. the Hastings Center and Yeoh Tiong Lay Centre). Participants completed the survey online via an electronic link sent in the email. Participants were not personally paid for taking part but instead had the option to enter an email address to enter a charity-raffle. For every person that completed the survey we put £2 (approx $2.50 USD) into a pot to be donated to charity, and at the end of data collection we randomly selected one response and asked that person which charity we should donate to from the following list: Against Malaria Foundation (AMF), Royal Society for the Prevention of Cruelty to Animals (RSPCA); Schistosomiasis Control Initiative (SCI); Cancer Research UK; British Heart Foundation; the Red Cross; and Doctors without Borders. The person chose X.

Eighty-six participants completed the survey in full, and five participants were excluded from data analysis because they were undergraduate students. Thus, our final sample consisted of 81 experts in moral philosophy (23 female, $M_{age}$ = 32, *SD* = 9.72). Around half of our sample were graduate students (56%), followed by post-doctoral researchers (17%), lecturers/associate professors (14%), and full professors (6%). Participants had spent an average of 8 years (including graduate studies) working in moral philosophy (*M* = 7.65, *SD* = 8.17).

**Measures**

In this study, participants completed the 9-item Oxford Utilitarianism Scale before rating their own self-reported utilitarianism and indicating their own ethical views (see below). Finally, participants indicated their gender and age, their status in the university (graduate student; post-doctoral researcher; lecturer/associate professor; full professor) and how many years they had spent in professional philosophy.

*Own Ethical View*. Participants were asked to indicate which of a list of 10 common ethical systems was closest to their own ethical view (act utilitarianism; rule utilitarianism; other form of utilitarianism; non-utilitarian consequentialism; Kantian ethics; other forms of deontology; virtue ethics; care ethics; religious ethics; commonsense morality; non-Western ethical view; or 'other'). We noted that perhaps none of these would perfectly describe participants' views, but asked them to indicate which was the most similar.

*Self-Report Utilitarianism*. To assess how utilitarian our expert participants judged themselves to be, we asked participants to read the following description and rate how utilitarian they were on a 1-10 scale:

> We would like you to tell us to what extent you consider your ethical views to be close to or far from utilitarianism. Since 'utilitarianism' can mean different things and since there are many ways in which one might reject utilitarianism, let us explain what exactly we mean by this question. By 'utilitarianism' we mean unadorned, classical act utilitarianism: roughly, the view that an act is right if and only if it maximizes aggregate welfare from a thoroughly impartial standpoint. So for our purposes views are more utilitarian the closer they are to this view, less utilitarian the further they are from this view. If you are an unqualified act utilitarian then you count as maximally utilitarian on this scale. If you are a rule utilitarian or a consequentialist whose axiology includes more than welfare then you are somewhat less utilitarian. Moving further away from the top end of the scale, the more a person thinks of morality in partial terms, and the more (and the stronger the) deontological constraints they accept, the less 'utilitarian' they count on our measure. If you are attracted to W. D. Ross's pluralist deontological theory you would rank low on this

scale. If you hold an absolutist Kantian theory which gives limited weight to consequences, or a highly traditional moral view, then you should rank at the very bottom of the scale. Please indicate where you yourself fall on this dimension, where 10 indicates that you fully endorse classical utilitarianism and 1 indicates that your view is as far as can be from such a utilitarian view (e.g. you hold very strong deontological views).

**Results**

The aim of this study was simple: to look at whether, for our expert participants who were well-versed in moral philosophy, scores on the OUS would be positively correlated with their own self-reported utilitarianism. Indeed, this is what we found: participants who rated themselves as being more utilitarian on the explicit self-report item also scored higher on the OUS. Participants' self-reported explicit utilitarianism was significantly correlated with scores on the OUS-IB ($r = .73$, $p < .001$), OUS-IH ($r = .69$, $p < .001$), and overall OUS scores ($r = .76$, $p < .001$). To probe and confirm this primary analysis, we subsequently looked at how participant's self-reported ethical view (e.g. act utilitarianism, Kantian ethics etc.) was associated with OUS scores, and conducted analyses controlling for participants' expertise level. Across all additional analyses the same pattern was observed: self-described utilitarian tendencies were significantly associated with scores on the OUS.

First, we looked at OUS scores as a function of the ethical view that our expert participants said characterized their own view. Participants who described themselves as believing most in "act utilitarianism" ($N = 11$) scored significantly higher on the OUS than those who described themselves as endorsing "other forms of deontology" ($N = 18$): on the OUS-IB

($t(14.93) = 6.06$, $p < .001$); the OUS-IH ($t(27) = 6.29$, $p < .001$); and overall ($t(14.09) = 6.56$, $p < .001$). Similarly, participants who described themselves as some kind of utilitarian (act, rule, or other: $N = 19$) scored higher than everyone else ($N = 58$): on the OUS-IB ($t(75) = 5.36$, $p < .001$); the OUS-IH ($t(75) = 5.30$, $p < .001$); and overall ($t(75) = 5.81$, $p < .001$). And looking just at those people who described themselves as some kind of consequentialist (utilitarian or non-utilitarian: $N = 36$), scores were higher than those endorsing deontological ethics (Kantian or other form of deontology: $N = 22$): on the OUS-IB ($t(55.68) = 5.83$, $p < .001$); the OUS-IH ($t(56) = 5.10$, $p < .001$); and overall ($t(55.98) = 6.29$, $p < .001$). Across the two measures, then, expert participants who reported themselves as being more utilitarian scored higher on the OUS. Such results provide strong face validity for our claims: while the OUS is not designed specifically for those with substantial experience with the theory of utilitarianism (one does not need a scale to measure such an expert's view - you can just ask them!), it was important that differences in the OUS mirrored self-described descriptions. Of course, if a Kantian and a utilitarian scored similarly on the OUS, it would be a rather useless tool with which to assess lay utilitarian tendencies. Fortunately, our results yielded strong differences across different measures: self-described utilitarians scored significantly higher on the OUS than those who didn't.

Second, we looked at whether results were robust to controlling for years spent in philosophy as a proxy for expertise-level within our expert panel. Partial correlations showed that self-reported explicit utilitarianism was significantly correlated with scores on the OUS when statistically controlling for years spent in philosophy: on the OUS-IB ($r = .72$, $p < .001$), OUS-IH ($r = .69$, $p < .001$), and overall ($r = .76$, $p < .001$). When looking only at graduate students ($N = 45$), self-reported explicit utilitarianism was again significantly correlated with

scores on the OUS-IB ($r$ = .70, $p$ < .001), OUS-IH ($r$ = .66, $p$ < .001), and overall OUS scores ($r$ = .73, $p$ < .001). When looking only at those at the post-doctoral level or higher, the same pattern was seen but with slightly higher correlations: on the OUS-IB ($r$ = .77, $p$ < .001), OUS-IH ($r$ = .73, $p$ < .001), and overall OUS scores ($r$ = .80, $p$ < .001). Similarly, when looking at graduates versus post-doctoral researchers or higher, an identical pattern of results was observed when looking at the effects of self-described ethical view on OUS scores.

Turning next to the question of the relation between OUS-IB and OUS-IH, our prediction that these would be more strongly correlated in an expert population was confirmed ($r$ = .75, $p$ < .001 in the expert sample compared to $r$ = .14, $p$ = .02 in the lay sample of Study 2, with these being significantly different, $Z$ = 6.5, $p$ < .001). This association was not lower in the graduate group (r = .74, p < .001) compared to the postgraduate one ($r$ = .78, p < .001; $Z$ = -.04, $p$ = .68). There was no relationship between years spent in philosophy and scores on the OUS overall ($r$ = .14, $p$ = .23), the OUS-IB ($r$ = .08, $p$ = .47), or the OUS-IH ($r$ = .17, $p$ = .13). Correspondingly, the poasitive relationship between OUS-IB and OUS-IH was robust to controlling for years spent in philosophy ($r$ = .75, $p$ < .001). This supports the notion that philosophical education may create pressure for greater coherence (Holyoak & Powell, 2016), leading students (and later professional philosophers) to revise their moral views to align more closely with dominant philosophical views, as noted above. An alternative explanation is that the shape of the debate within contemporary ethics attracts individual with an unusual psychological profile.

## The Oxford Utilitarianism Scale in Practice

Having established and validated our new Oxford Utilitarianism Scale, we next looked at how the OUS could be used in practice. Specifically, we sought to explore whether the OUS would shed further light on how utilitarian tendencies are associated with individual differences (e.g. psychopathy), moral attitudes (e.g. prosocial intentions), and ideological factors (e.g. religious belief). In doing so we show that the OUS will be a valuable tool for researchers working in moral psychology, demonstrating how previous work exploring utilitarian tendencies has been limited by failure to take into account the ways in which the two core aspects of utilitarianism – impartial beneficence and instrumental harm – have separate and divergent influences on moral cognition. By using the OUS we can clarify and extend the relationship between utilitarianism and other constructs.

To look at these issues, we used the same dataset as that used in Study 2 for assessing construct validity in a lay population. As noted above, this dataset consisted of 282 participants recruited from MTurk (178 female, $M_{age}$ = 39, *SD* = 12.66), of whom the majority of participants had attended college or higher education (80%). For all of the subsequent discussion, the reader can see Table 7 for correlations, Table 8 for full *M*s and *SD*s, and the Supplementary Materials for results with other non-central measures we have omitted here in the interests of space.

We wish again to highlight here at the outset that scores on the OUS-IB and OUS-IH were only weakly correlated (Study 1, $r$ = .04, $p$ = .22; Study 2, $r$ = .14, $p$ = .02). Moreover, relatively few people in our sample can be characterised as having a genuine overall utilitarian tendency. On the 1-7 scale, the mean overall OUS score was below the midpoint (Study 1, *M* = 3.58, *SD* = 0.86; Study 2, *M*= 3.50, *SD* = 0.92), with slighter higher average scores on the IB

subscale (Study 1, *M* = 3.75, *SD* = 1.15; Study 2, *M* = 3.65, *SD* =1.20) than the IH (Study 1, *M* = 3.37, *SD* = 1.24; Study 2, *M* = 3.31, *SD* = 1.22). Only a quarter of participants (Study 1, 28%; Study 2, 26%) scored above the midpoint of the scale (>4), and this was again higher for the IB (Study 1, 39%; Study 2, 35%) than the IH (Study 1, 27%; Study 2, 24%). This indicates that only a minority of lay people have significant proto-utilitarian leanings. And if we consider a more robust indicator of significant proto-utilitarian tendencies – people who scored at 5 or above on the scale – only 5% in Study 1 and 4% in Study 2 could be classified as significantly utilitarian overall, with this higher for the IB (Study 1 and Study 2, 16%) than the IH (Study 1, 9%; Study 2, 5%). While there is, then, a distribution of utilitarian tendencies among the population, the prevalence of people who can meaningfully be characterized as having a moderate to strong utilitarian tendency should not be overstated. Befitting its status as a radically demanding and counterintuitive moral philosophy, few laypeople even approximate the views of genuine utilitarians. This is an issue to which we return in our discussion.

## Personality and Individual Differences

### Psychopathy

A growing body of research has shown a relationship between sub-clinical psychopathy and utilitarian judgment in sacrificial dilemmas, whereby ‘utilitarian’ judgments are associated with higher levels of both clinical (Koenigs, Kruepke, Zeier, & Newman, 2012) and sub-clinical psychopathy and related anti-social traits (Bartels & Pizarro, 2011; Kahane, Everett et al., 2015; Koenigs et al., 2012; Wiech et al., 2013). Recent work suggests, however, that this may be an artifact of the use of sacrificial moral dilemmas as the measure of utilitarianism: while psychopathy is associated positively with a willingness to personally harm others for the greater

good, it is associated negatively with endorsement of moral self-sacrifice (Kahane, Everett et al., 2015). We predicted, therefore, that while overall OUS scores might be positively associated with psychopathy, this pattern would be driven by the OUS-IH subscale and that there would be either be no relationship, or a negative one, between psychopathy and the OUS-IB subscale. We measured psychopathy using Levenson's primary psychopathy sub-scale (Levenson, Kiehl, & Fitzpatrick, 1995), which consists of 16 items including 'Success is based on survival of the fittest; I am not concerned about the losers.' ($\alpha$ = .89). Supporting predictions, overall OUS scores were not significantly associated with primary psychopathy ($r$ = .11, $p$ = .07), and neither were scores on the OUS-IB subscale ($r$ = .09, $p$ = .13). Higher scores on the OUS-IH, however, were significantly associated with increased psychopathy ($r$ = .30, $p$ < .001), and the strength of correlation between psychopathy and the OUS-IB and OUS-IH was significantly different ($Z$ = -2.77, $p$ = .01).

**Empathic Concern**

What is the relationship between empathic concern and utilitarian tendencies? Some work has suggested that deontological – but not utilitarian – tendencies are related to increased empathic concern (Choe & Min, 2011; Conway & Gawronski, 2013; Gleichgerrcht & Young, 2013; Patil & Silani, 2014). Yet on theoretical grounds, this is more than a little surprising, given that empathy has been claimed to be a central psychological source of utilitarianism (Hare, 1981; Smart & Williams, 1973). However, once we take into account that these prior studies all used sacrificial dilemmas as a measure of utilitarianism, this discrepancy can be explained. Such dilemmas focus exclusively on the negative dimension of utilitarianism, and specifically, on willingness to cause harm for the greater good. If one considers killing one to save five others, a

stronger empathetic response is likely to be an obstacle to endorsing such instrumental harm. By contrast, the historical tie between utilitarianism and empathy is most plausibly related to utilitarianism's positive aspect, its impartial concern for everyone's welfare. It is therefore likely that the previously found negative relationship between empathic concern and utilitarianism may have been driven entirely instrumental harm, rather than impartial beneficence.

To explore this question we looked at the relationship between scores on the OUS and empathic concern. We used the Empathic Concern subscale of the Interpersonal Reactivity Index (Davies, 1980), which measures sympathy and concern for others and consists of 7 items including 'When I see someone being taken advantage of, I feel kind of protective towards them' ($\alpha$ = .90). Given the strong conceptual (and empirical) negative relationship between psychopathy and empathic concern (here: $r$ = -.59, $p$ < .001), our predictions mirrored those reported above for psychopathy. Specifically, we again predicted that while overall OUS scores might be negatively associated with empathic concern, this pattern would be driven by the OUS-IH sub-scale and that there would be a positive relationship between empathic and the OUS-IB subscale.

There was an overall positive relationship between OUS scores and empathic concern ($r$ = .14, $p$ = .02), and while both OUS subscales were significantly correlated with empathic concern, these were in opposite directions. While increased empathic concern was associated with scores on the OUS-IB subscale ($r$ = .33, $p$ < .001), decreased empathic concern was associated with the OUS-IH ($r$ = -.16, $p$ = .006); these correlations were significantly different ($Z$ = 2.28, $p$ = .03). Thus, while people who feel greater empathic concern care more about impartially maximizing welfare, they are also less likely to accept instrumental harm to achieve

those ends. In other words, as suggested by Kahane, Everett et al. (2015), the negative utilitarian tendency to endorse harm as a means to the greater good might stand in opposition to the positive utilitarian tendency to self-sacrifice and impartially maximize happiness. Again, at a philosophical level, the aspects of impartial beneficence and instrumental harm are consistent, but at a psychological level these factors may have distinct antecedents and consequences on cognition.

**Identification with all of Humanity**

We next looked at how OUS scores were associated with scores on the *Identification with All Humanity Scale* (IWAH): a measure of the extent to which people identify more with humanity as a whole as opposed to exhibit more parochial attachment to one's own community or country (McFarland, Webb, & Brown, 2012). The extent to which an individual identifies with the whole of humanity is best seen as an affective disposition rather than a moral view. However, such an all-encompassing impartial concern is often claimed to be a core feature of classical utilitarianism (Hare, 1981; Smart, 1956). For this reason, we predicted that IWAH would be positively related to the Impartial Beneficence sub-scale, but unrelated (or even negatively related) with the negative or anti-traditionalism subscales.

The IWAH Scale was taken from McFarland et al. (2012) and consisted of 9 questions, including requiring participants to rate, for people in their community, people in their country, and people all over the world, "How close do you feel to each of the following groups?" In analyzing results, the procedure advised by McFarland et al. was used, regressing the raw scores to give a more accurate representation of the variance in identification with all of humanity, whereby higher scores indicate greater identification with all of humanity.

As predicted, IWAH was significantly associated with OUS scores overall ($r$ = .13, $p$ = .03), but this was driven in opposite directions by the OUS-IB ($r$ = .33, $p$ < .001) and OUS-IH ($r$ = -.19, $p$ < .001), with these correlations being significantly different ($Z$ = 6.94, $p$ < .001). Therefore, greater impartial beneficence was associated with greater identification with all of humanity, but greater instrumental harm was associated with less identification with all of humanity.

**Need for Cognition**

It has been argued that utilitarian decision-making is uniquely based on effortful conscious reasoning (as opposed to immediate emotional responses) and that utilitarian inclinations are uniquely associated with a need for cognition (Conway & Gawronski, 2013); the motivational tendency to engage in effortful thinking (Cacioppo, Petty, & Kao, 1984). Indeed, the tendency for an individual to engage in and enjoy thinking seems intuitively related to utilitarianism to the extent that utilitarian decision-making involves the rational weighing of different consequences and the rejection of simpler intuitive solutions to moral problems. It has also often been claimed that greater reliance on reason was the source of the historical emergence of utilitarianism, and generally of more impartial and inclusive moral views (Pinker, 2011; Singer, 2011). The extent to which need for cognition might have divergent relations with the twin factors of impartial beneficence and instrumental harm, though, remains unclear. To explore this question with the OUS, we had participants complete the 18-item Need for Cognition scale (Cacioppo et al., 1984), in which participants rate how characteristic or uncharacteristic certain statements are of them, such as “I would prefer complex to simple problems” (*1 = extremely uncharacteristic of me; 7 = extremely characteristic of me*).

In contrast to some previous research, there was no relationship between need for cognition and OUS scores: not overall ($r$ =.02, $p$ = .73), for the OUS-IB ($r$ =.06, $p$ = .35) or for the OUS-IH separately ($r$ =-.03, $p$ = .57). And interestingly, there was no association between need for cognition and other relevant measures, such as Robinson's utilitarianism scale ($r$ = -.09, $p$ = .13) or sacrificial moral dilemmas ($r$ = -.04, $p$ = .46).

## Moral Attitudes

### Hypothetical Donation

Present-day utilitarianism is commonly associated with movements to encourage more charitable donations to strangers in need in developing countries (MacAskill, 2015; Singer, 2015). On utilitarian grounds, donating to effective charities like those in the developing world is not just morally praiseworthy, but actually morally required: we ought to give the money away, and it is wrong not to do so (Singer, 1972). We therefore next looked at how participant's self-reported tendency to engage in altruistic behavior would be associated with scores on the OUS. To the extent that the OUS measures utilitarianism, it should be positively correlated with a self-reported intention to perform effective charity. To do this we used a measure taken from Kahane, Everett et al. (2015) where participants were told to:

> "Imagine that you have just received a bonus at work of $100. Your company, however, has a policy for all staff members that they can either take this bonus money of $100 for themselves with no penalties, or choose to donate it to a respected charity that helps people in the Third World. You can choose to keep or donate as much of the bonus as you wish. If you choose to donate the bonus, the company will double what you put in: if you donate $50 then the company will donate $100, and if you donate $100 the

company will donate $200. This choice is made using an anonymized online system, so no one in the company will know what you decided to do."

Participants then rated how much of the bonus they would donate to charity on a scale of $0 to $100. Perhaps unsurprisingly given the importance of charity to contemporary utilitarian thought, overall OUS scores were positively associated with larger hypothetical donations ($r$ = .31, $p$ < .001). This overall correlation was, however, driven entirely by the Impartial Beneficence subscale ($r$ = .40, $p$ < .001), with no relationship between hypothetical donations and endorsement of Instrumental Harm ($r$ = .03, $p$ = .62). The strengths of these correlations were, again, significantly different ($Z$ = 5.00, $p$ < .001). It is worth noting that the measure here is strictly a hypothetical one, with no actual cost to participants for making the altruistic decision. And yet people who are higher in the endorsement of instrumental harm still don't even hypothetically donate more. This result would be surprising if utilitarianism reflected a unitary psychological construct, but is to be expected if – as we have argued – what we call utilitarianism actually is grounded in two relatively distinct psychological constructs.

**Environmental Protection**

The relationship between utilitarianism and attitudes towards the environment is complex. On the one hand, from a utilitarian perspective natural things such as rivers and rainforests are of only instrumental value, insofar as they bear on the utility of sentient beings. On the other hand, because climate change will likely cause severe harm to humans and other animals, there are strong utilitarian reasons to protect the environment now to prevent that future suffering. We were therefore interested in the relationship between proto-utilitarian tendencies and support for environmental protection. Such support reflects utilitarianism's positive

dimension given that relevant policies require significant self-sacrifice and that most of their beneficiaries will be physically and temporally distant people and, indeed, people who will exist only in the future. We thus predicted a positive relation between support for environmental protection and Impartial Beneficence; by contrast, there is no reason to expect such a relation with Instrumental Harm.

Participants were asked how much they agreed or disagreed that "It is important to protect the environment and the world's resources because of the negative consequences on future humans if we don't" (*1 = strongly disagree, 7 = strongly agree*). Results showed that there was no overall relationship between OUS scores and support for environmental protection based on utilitarian considerations ($r$ =.03, $p$ = .67), but this was due to the two subscales being significantly associated in different directions. As predicted, increased impartial beneficence was associated with greater endorsement of environmental protection ($r$ =.14, $p$ = .02). By contrast, instrumental harm was associated with less willingness to protect the environment ($r$ = -.21, $p$ < .001), and these were significantly different (Z = 4.55, p < .001). While instrumental harm measures willingness to harm people for the sake of the greater good, it is likely—as suggested by its association with psychopathy—that it also measures general indifference to harm and destruction, including harm to the environment.

## Ideology

### Religiosity

How would scores on the OUS relate to religious belief? Religious moral systems often emphasize the importance of rule-based moral decision making. Delve into the holy books of almost all major religions and you will find an extravagant number of rules dictating actions that

are either morally obligatory (e.g., observing the Sabbath), or morally impermissible (e.g., coveting thy neighbor's wife). Such actions are typically seen as morally required or forbidden by virtue of being what they are, rather than as a result of their consequences – God commands it, and so it must be. It is thus not surprising that utilitarianism as a historical movement often came into direct conflict with religious views (Mill, 1863) and the utilitarian approach to morality is widely denounced by religious thinkers (Anscombe, 1981; Paul II, 1995).

In line with the historical relationship between religion and utilitarianism, psychological research has reported that religious individuals are more likely to endorse a non-utilitarian – approach to morality (Piazza, 2012; Piazza & Sousa, 2013; Szekely, Opre, & Miu, 2015) and religious individuals show a tendency to prefer rule-based (deontological) over outcome-based (consequentialist) explanations for the wrongness of a moral action (Piazza, 2012). However the exact nature of the relationship between religious belief and utilitarianism might be less straightforward once we recognize that utilitarianism has both negative (Instrumental Harm) and positive (Impartial Beneficence) dimensions. The two-factor OUS therefore makes it possible to investigate the relation between utilitarianism and religious belief in a more fine-grained way than was possible before.

We measured religiosity with a 5-item Centrality of Religiosity Scale (CRS) (Huber & Huber, 2012). The CRS is a measure of the centrality, importance, and salience of religiousness in a person, and consists of five items each tapping one of the theoretically defined core dimensions of religiosity: public practice, private practice, religious experience, ideology, and intellect. Participants were asked to rate on a 5-item scale "How often do you think about religious issues?" (Intellect: *1 = never; 5 = very often*); "To what extent do you believe that God

or something divine exists?" (Ideology: *1 = not at all, 5 = very much so*); "How often do you take part in religious services?" (Public practice: *1 = never, 5 = very often*); "How often do you pray?" (Private practice: *1 = never, 5 = very often*); and "How often do you experience situations in which you have the feeling that God or something divine intervenes in your life?" (Experience: *1 = never, 5 = very often*). Scores were combined into a single reliable measure (α =.93) of the centrality of religiousness for each participant.

Overall OUS scores were associated with religiosity ($r$ = .15, $p$ = .01), which is perhaps surprising given that previous work has associated deontological thinking with religiosity (Piazza & Sousa, 2013; Szekely et al., 2015). But a positive relation between religiosity and Impartial Beneficence, i.e. the tendency to endorse the impartial promotion of everyone's welfare even at personal sacrifice to oneself, would be less surprising. Religious systems often include injunctions to "love thy neighbor" – even those from other communities, as exemplified in the parable of the Good Samaritan. Christians are told to "turn the other cheek" and "take up the cross" such that they humble themselves before God and forego material or social benefits in lieu of the divinely ordained moral good. Upon his appointment, Pope Francis said that "It hurts me when I see a priest or a nun with the latest model car, you can't do this… please, choose a more humble one. If you like the fancy one, just think about how many children are dying of hunger in the world" (Francis, 2013). Peter Singer has said almost identical things. Overall, the focus on self-sacrifice and love for one's neighbors, combined with the idea that all are equal under God's eyes, might suggest that scores on the OUS-IB would actually be positively associated with religiosity. Indeed, this was what we found ($r$ = .15, $p$ = .01): those people who were more religious were also those who showed a greater endorsement of the impartial maximization of the

greater good. The overall correlation between OUS scores and religiosity thus was driven by the OUS-IB and there was no relationship with religiosity and scores on the OUS-IH ($r$ = -.06, $p$ = .31); and again these were significantly different ($Z$ = 2.70, $p$ = .01). This pattern of results suggests that the negative relationship observed between utilitarian thinking and religiosity in previous research may be restricted to the component of utilitarianism that rejects traditional deontological moralities, rather than any reduced concern for impartially maximizing welfare *per se*. Indeed, this contention is supported when looking at the 6-item Anti-Traditional Morality measure that was not included in the final scale. As would be expected, there was a strong negative relationship such that people who were more religious were less willing to break with the rules of traditional deontological moralities ($r$ =-.48, $p$ < .001). Given that for our American participants traditional moralities are based heavily in the Judeo-Christian tradition, this result is not surprising.

The positive relation between religiosity and utilitarianism is not surprising when one considers the dissociation of impartial beneficence and instrumental harm in our 2D model. When Pope John Paul II criticized utilitarianism for treating people as means to an end, he primarily took issue with the instrumental harm aspect of the theory, and indeed it is this focus on instrumental harm that seems to have driven the negative relationship between religiosity and utilitarianism seen amongst laypeople in previous work. But the positive correlation of religiosity and impartial beneficence is not surprising, and it is this overlap that can be seen in the pronouncements of Pope Francis I on the importance of sacrificing material goods for oneself to help those in foreign countries. Indeed, from a historical perspective the relationship between utilitarianism and the Judeo-Christian tradition is complex: although utilitarianism has been

historically often in conflict with religious belief, some of the direct intellectual precursors of utilitarianism, such William Paley, were Christian believers (Schneewind, 1977), and the self-sacrificing impartial beneficence at the core of utilitarianism has been claimed to have its roots in Christian ideas (Gray, 2015). Indeed, Christian ethics involve a radical demand for self-sacrifice, impartiality, and universal love. It is really the willingness to harm and rejection of traditional moral rules that lie at the heart of the historical tension between utilitarianism and religion, and Christian thinkers have recently begun exploring the affinities between the two views (Camosy, 2012).

**Political Ideology**

In addition to associating deontological moral thinking with religiosity, some work has suggested that political conservatism exerts a significant effect on deontological moral reasoning – independent of religiosity (Piazza & Sousa, 2013). For this reason, we measured both economic and social political ideology (*1 = very liberal / left*, *7 = very conservative / right*), as well asking participants to indicate which political party (if any) they identified with.

Overall OUS scores were not associated with either self-reported economic ($r$ = .02, $p$ = .69) or social conservatism ($r$ = .06, $p$ = .28). But, again, more nuanced results were observed when looking at the subscales independently. Economic conservatism was significantly associated with *reduced* impartial beneficence ($r$ = -.12, $p$ = .054) but *increased* acceptance of instrumental harm ($r$ = .18, $p$ = .02), with these being significantly different (Z = -3.88, p < .001). Such results mesh well with previous work showing a relationship between libertarianism and pro-sacrifice judgments in sacrificial dilemmas (Iyer, Koleva, Graham, Ditto, & Haidt, 2012). Social conservatism was not associated with scores on the OUS-IB ($r$ = -.06, $p$ = .34), but, as

with economic issues, greater social conservatism was associated with increased endorsement of instrumental harm ($r$ = .18, $p$ = .003), and again these associations were significant different (Z = -3.09, p = .002).

A similar pattern of results was seen when looking at political party identification. There was no overall difference between Republicans and Democrats on overall OUS scores, $t$(194) = -0.35, $p$ = .73, but again, this was because party identification was associated with the two subscales in opposite ways. There was a significant difference on OUS-IB scores as a function of political party, $t$(194) = -2.38, $p$ = .02, whereby Democrats reported greater impartial beneficence ($M$=3.77, $SD$=1.22) than Republicans ($M$=3.32, $SD$=1.22). Similarly, there was a significant difference on OUS-IH scores as a function of political party, $t$(194) = 2.34, $p$ = .02, but this was such that Republicans reported greater endorsement of instrumental harm ($M$=3.70, $SD$=1.21) than Democrats ($M$=3.25, $SD$=1.26). In sum, Democrats were more likely to endorse impartial beneficence, and Republicans more likely to endorse instrumental harm. The distinct subcomponents of proto-utilitarian thinking, then, have distinct psychological correlates and even appear to be differently represented amongst different groups of people.

**Demographics**

Finally, we looked at the relationship between OUS scores and demographic measures of age, education level, and gender. There was no relationship between age and OUS scores: not overall ($r$ = -.06, $p$ = .30), for the OUS-IB ($r$ = -.08, $p$ = .19), or for the OUS-IH ($r$ = -.01, $p$ = .89). Similarly, there was no relationship between education level and OUS scores: not overall ($r$ = -.05, $p$ = .40), for the OUS-IB ($r$ = -.04, $p$ = .56), or for the OUS-IH ($r$ = -.04, $p$ = .47). There were no gender differences in OUS scores overall, $t$(279) =.04, $p$ = .97, nor for the OUS-IB,

*t*(279) =-1.40, *p* = .16. While the effect of gender on OUS-IH scores was not significant, *t*(279)= 1.79, *p* = .07, the mean difference was consistent with previous research such that men showed a slightly greater tendency to endorse instrumental harm (*M*=3.48, *SD*=1.33) than women (*M*=3.21, *SD*=1.13).

## General Discussion

Utilitarianism tells us to impartially maximize the aggregate well-being of everyone—and that we can severely harm or even kill innocent people if doing so is needed to achieve this overarching moral ideal. Most psychological research on utilitarian decision-making has so far focused on this last, negative dimension of utilitarianism, largely ignoring the more foundational impartial ideal that underlies it. In this paper, we have introduced a new theoretical framework for thinking about proto-utilitarian moral thinking in the lay population. Using this framework, we have developed, refined and validated a new approach to measuring such thinking that taps both positive and negative dimensions of proto-utilitarian tendencies in the lay population. The Oxford Utilitarianism Scale (OUS) is designed to address the limitations of prior measures, especially the currently standard method of measuring so-called 'utilitarian' judgments with sacrificial dilemmas. In creating the scale, we took care to ensure that it was conceptually accurate without inappropriately imposing abstract philosophical notions on the moral thinking of non-philosophers. Thus, while the pool of items we initially considered was based on a thorough analysis of the relevant literature in ethics and vetted by leading professional moral philosophers, the final scale is empirically driven and reflects clusters of moral evaluations that were statistically robust in large samples taken from the lay population. Most importantly, our

findings show that proto-utilitarian tendencies do not form a unitary psychological phenomenon, and in fact consist of two largely independent subcomponents. This division explains otherwise puzzling results in the existing literature, sheds new light on the psychological sources both of utilitarianism and common opposition to it, and opens up new directions for future research.

**The 2D Model of Utilitarian Decision-Making**

One of the central ways in which the OUS departs from previous work is by not assuming that utilitarian moral tendencies form a unitary psychological phenomenon, let alone one that can be understood by studying responses in a highly specific moral context. According to our 2D model, utilitarian decision-making involves two dissociable and independently important aspects: the first, Impartial Beneficence reflects the extent to which individuals endorse the impartial promotion of everyone's welfare, while the second, Instrumental Harm, reflects the extent to which people are willing to endorse instrumental harm to achieve the greater good. By dissociating these two independent factors of utilitarianism, it is possible to reach a more nuanced and accurate picture of how utilitarian tendencies are related to a host of individual difference measures and other moral attitudes.[6]

The initial results reported here already show how the 2D model of utilitarian decision-making allows us to make better sense of prior findings and associations that would otherwise seem puzzling. They also show how the model can be used to shed light on the psychological sources of the core 'positive' aspect of utilitarianism, its radically impartial character and the demands on the self that it consequently imposes. For example, multiple prior studies claimed to find that the moral-decision making of psychopaths is abnormally utilitarian—a surprising result given the antisocial character of this condition. But once we take into account the distinction

between the Instrumental Harm and Impartial Beneficence dimensions of utilitarian thinking, this result can easily be explained. As reported above, psychopathy in a non-clinical sample was associated with greater endorsement of Instrumental Harm but not with Impartial Beneficence, a far less surprising result. In a prior study we similarly found a positive relation between psychopathic tendencies and pro-sacrificial judgments in sacrificial dilemmas but not with the 'greater good' dilemmas that pit impartial vs. partial moral concerns (Kahane, Everett et al. 2015), a finding replicated here.

The 2D model also allows us to better clarify the relationship between utilitarian moral thinking and empathic concern. Again, the negative relationship between utilitarian decision-making and empathic concern suggested by multiple prior studies (Choe & Min, 2011; Conway & Gawronski, 2013; Gleichgerrcht & Young, 2013; Patil & Silani, 2014) is highly puzzling given the historical and, on some views, strong conceptual tie between classical utilitarianism and empathic concern (Hare, 1981; Singer, 1972; Smart, 1961) and evidence tying empathic concern to extreme altruism to distant strangers (Brethel-Haurwitz et al., 2016). But our results show that this association is driven exclusively by the Instrumental Harm dimension of utilitarianism and that empathic concern is at the same time positively associated with Impartial Beneficence, utilitarianism's positive core. Indeed, our results show that these two dimensions of utilitarianism are not only independent but are also inversely correlated with a psychological trait that is highly relevant for morality, further confirming our view that it is a mistake to treat utilitarian decision-making as a unitary psychological phenomenon outside of the philosophical context. In line with this, we also found that while Impartial Beneficence was associated with greater donation rates to a hypothetical charity helping people in the third world, greater support for welfare-based

concern about the environment, and higher levels of identification with the whole of humanity, these measures were either not, or in the case of environmental concern, negatively, associated with Instrumental Harm.

The 2D model similarly sheds new light on the relation between utilitarian moral thinking and religiosity. While the two have often been historically opposed, our results suggest that the core of that conflict may lie in utilitarianism's endorsement of instrumental harm, as well as in its rejection of traditional and absolutist moral rules. Yet religiosity was positively correlated with impartial beneficence, reflecting the important affinities between utilitarianism and Judeo-Christian moral ideals, affinities being explored in some recent work in ethics (Camosy, 2012).

Utilitarianism is typically viewed as a radically progressive moral view. But as the association between Impartial Beneficence and religiosity reveals, the relationship between utilitarianism and religious and political commitments might be a bit more complicated. This notion is further confirmed by the nuanced relationship between the OUS sub-scales and political ideologies. For example, economic conservatism was at once positively associated with Instrumental Harm and, perhaps unsurprisingly, negatively correlated with Impartial Beneficence. This finding is in line with previous studies tying libertarianism to pro-sacrifice judgments in sacrificial dilemmas (Iyer et al. 2012), again illustrating the point that willingness to harm others is compatible with having a radically partial and even self-centered view of morality (Kahane 2015). Political party affiliation was also associated with different dimensions of utilitarianism, with Democrats scoring higher on impartial benevolence and Republicans higher on instrumental harm.

Given that the OUS-IB and OUS-IH sub-scales measure independent psychological factors that are differentially or even inversely associated with a range of traits and measures, it may be questioned whether these two dimensions belong in a single scale. However, the psychological disunity of aspects of utilitarian decision-making in the lay population does not undermine the theoretical unity of the overall construct. The items on this scale were derived from an extensive review of the ethical literature and vetted by a panel of moral philosophers, and scores on both of the sub-scales were associated with endorsement of an explicit statement of utilitarianism. The two sub-scales were also strongly correlated in a sample of expert moral philosophers, further confirming their theoretical unity. And ordinary individuals who score highly on both sub-scales would be appropriately classified as highly utilitarian—and would rank as considerably more utilitarian than someone who merely has a permissive attitude to instrumental harm.

However, it is true that focusing on the overall OUS score risks overlooking or even distorting the relation between utilitarian tendencies and various psychological phenomena. At this early stage, therefore, we suggest that primacy should be given to the sub-scale scores. Such an approach is not uncommon with constructs that target conceptually related but psychologically distinct constructs. A prominent example is the Interpersonal Reactivity Index (Davies, 1980), a multi-dimensional measure of trait empathy that consists of four independent sub-scales, each measuring a theoretically related but psychologically distinct trait.

**Instrumental Harm: Implications for the Sacrificial Dilemmas Approach**

The sacrificial dilemmas method is currently the dominant approach to measuring utilitarian tendencies (Christensen & Gomila, 2012) and it has yielded important,

groundbreaking advances into our understanding of how people consider instrumental harm in moral decision making (Greene, 2014). At the same time, our studies support the argument that sacrificial dilemmas only tap a narrow dimension of utilitarian thinking and that it is unwarranted to infer general claims about the psychological processes underlying utilitarian decision-making only on the basis of the study of pro-sacrifice judgments in such dilemmas (Kahane, Everett et al., 2015; Kahane & Shackel, 2010; Kahane, 2015). Our present work confirms that instrumental harm is a significant dimension of proto-utilitarian tendencies in ordinary people, albeit one that is dissociable from the more positive dimension of impartial beneficence.

The OUS is a measure of individual differences in utilitarian tendencies rather than a tool for directly studying the psychological processes (such as aversion to instrumental harm) involved in specific episodes moral decision-making. While much work by Greene and colleagues has focused on how sacrificial dilemmas illuminate the cognitive architecture of moral decision making rather than utilitarian tendencies *per se*, sacrificial dilemmas are often used to measure individual differences in utilitarian tendencies within a population as well as differences in such tendencies between populations. We propose that the OUS should supersede sacrificial dilemmas as a method for measuring such individual differences, even in the domain of Instrumental Harm. Besides its brevity, the Instrumental Harm sub-scale of the OUS covers a broader range of considerations relating to instrumental harm than those captured by sacrificial dilemmas, and also avoids the far-fetched and often fantastic character of many sacrificial dilemmas (Bauman et al., 2014).

At the same time, sacrificial dilemmas remain an important tool for studying the psychological processes underlying support for instrumental harm in moral decision-making.

Indeed the OUS could be used to distinguish those pro-sacrifice judgments that reflect only greater acceptance of instrumental harm and those that may also reflect a broader impartial moral perspective; the findings reported here suggest that there may be such a subset of pro-sacrifice responses to sacrificial dilemma given that such responses were associated with Impartial Beneficence (albeit weakly).

Nevertheless, we believe that there is a strong case for abandoning the widespread but unhelpful terminological practice of classifying pro-sacrifice responses in sacrificial dilemmas as 'utilitarian judgments' (Kahane, Everett et al., 2015; Kahane, 2015). On that terminology, a psychopath who is concerned only about his self-interest and has no compunction about harming others would be described as 'strongly utilitarian' while an extreme altruist who is a moral vegetarian, gives most of her money to charities helping people in the developing world, and who donated her kidney to a stranger, yet who recoils from the idea of violently sacrificing an innocent person to save five others, would count as 'strongly deontological'. By contrast, the OUS would reveal that these two individuals are unusually high on one aspect of a utilitarian outlook but low on the other. Neither is usefully described as fully utilitarian, although it could be argued that the extreme altruist comes much closer.

Theoretical clarity and precision would be better served, for example, if instead of describing the pro-sacrificial tendencies of individuals with antisocial traits as a 'utilitarian bias', it was more consistently referred to as a 'consequentialist bias' (since consequentialism need not imply the positive, impartial aspect of utilitarianism), or even more precisely as a bias in favor of instrumental harm. Importantly, on our framework such individuals *are* more utilitarian than others in one important respect, even if not in others. Even if this bias is driven by reduced

aversion to harm, its overt manifestation seems to be genuinely focused on what we called instrumental harm—i.e. seeing harm as permitted only when it leads to a morally better consequence. This interpretation would be defeated only if such individuals had a morally permissive attitude toward harming irrespective of consequence. Yet this seems unlikely, at least in most individuals whose anti-social tendencies are sub-clinical.

**Impartial Beneficence: Preliminary Findings and Future Directions**

While the psychology of instrumental harm has received a great deal of attention in moral psychology, less work has so far been devoted to investigating the sources of radically impartial moral views. There is of course a considerable literature studying altruism (e.g. Batson, 1991; de Waal, 2007; Krebs, 1982) and charitable giving more specifically (e.g. Bekkers & Wiepking, 2011; Caviola, Faulmüller, Everett, Savulescu, & Kahane, 2014; Harbaugh, Mayr, & Burghart, 2007; Oppenheimer & Olivola, 2011; Van Lange, Bekkers, Schuyt, & Vugt, 2007), but as discussed earlier, Impartial Beneficence goes beyond a willingness to make sacrifices to help others, or even willingness to make sacrifices to help complete strangers who will not reciprocate (as studied, for example, using the Dictator Game; see e.g. Benenson, Pascoe, & Radmore, 2007; Brañas-Garza, 2006; Eckel & Grossman, 1996; Engel, 2011; Everett, Haque, & Rand, 2016). Individuals differ in the degree to which they are disposed to make sacrifices to help strangers, but such differences are compatible with commonsense morality since it regards such sacrifices as permissible and even praiseworthy.

Impartial Beneficence is the far more radical view that we are *required* to treat the well-being of *all* sentient beings *equally* — and that we should therefore give as much moral weight to distant strangers as to our closest relatives. While such Impartial Beneficence is theoretically

distinct from identification with the whole of humanity (McFarland et al., 2012), the two constructs are obviously closely related and as expected were strongly correlated.[7]

Impartial Beneficence was also associated with greater levels of empathic concern, suggesting an important role for emotion in the core positive dimension of utilitarianism. This finding is in line with the theorizing of some prominent utilitarians (Hare, 1981; Smart & Williams, 1973), but in tension with accounts of Impartial Beneficence that see it as based in cold reason (de Lazari-Radek & Singer, 2012; Sidgwick, 1901) and with some recent psychological theorizing (Greene, 2008). The link with empathic concern is also consonant with work suggesting that the psychological and neural profile of extreme altruists is the reverse of that of psychopaths (Marsh et al., 2014). To the extent that extreme altruism is driven by unusual levels of empathic concern (Brethel-Haurwitz, Stoycos, Cardinale, Huebner, & Marsh, 2016), the 2D model predicts that such extreme altruists would be high on Impartial Beneficence but low on Instrumental Harm. And although we did not find a negative association between psychopathy and Impartial Beneficence in the lay population, we had previously found that psychopathy was linked to reduced endorsement of self-sacrifice for the greater good in the greater good vignettes (Kahane, Everett et al. 2015)..

While Impartial Beneficence is not identical with self-sacrifice, even in its extreme forms, it is still the case that 3 of the 5 items in the OUS-IB focus on self-sacrifice. Further research is needed to clarify the relationship between endorsement of self-sacrifice and refusal to give priority to those who are emotionally, socially, spatially and temporally near—two aspects of impartiality that are potentially distinct. The ‘greater good’ vignettes introduced in Kahane,

Everett et al. (2015), which cover different aspects of partiality/impartiality, could be used for this purpose.

The association between Impartial Beneficence and empathic concern may seem inconsistent with recent criticisms of empathy that contrast it with a more impartial, consequentialist approach (Bloom, 2017). However, that criticism targets empathy understood as feeling what you believe others feel, and as contrasted with compassion or concern for the well-being of others (Jordan, Amir, & Bloom, 2016): it is largely the latter which is measured by the empathic concern sub-scale employed in the present study (Davies, 1980). Still, further research into the relationship between empathic concern and different aspects of moral impartiality is needed. Moreover, it should not be assumed that an attitude of impartial beneficence will be automatically translated into focused efforts to maximize utility of the kind advocated by the effective altruist movement (MacAskill, 2015; Singer, 2015). It may be that a tendency toward impartial beneficence makes up the affective background to such an explicit maximizing aim, but that further cognitive steps are needed to endorse it explicitly.

Recent work in moral psychology and its evolutionary origins has highlighted the ‘groupish’ character of human morality (Greene, 2014; Haidt, 2012) and the ways in which altruism is often constrained by self-interest (e.g. Bardsley, 2008; Dana, Cain, & Dawes, 2006; Yamagishi & Kiyonari, 2000) and social distance (Jones & Rachlin, 2006). Commonsense morality allows individuals to prioritize self over others, family and friends over strangers, and country and other ingroups over outgroup members. The British Prime Minister Theresa May recently stated that ‘If you believe you’re a citizen of the world, you’re a citizen of nowhere’. Further familiar injunctions include ‘Family comes first’ and ‘charity begins at home’. The

impartial cosmopolitan outlook of utilitarianism is a radical departure from such commonsense attitudes. But while the factors that underlie allegiance to various ingroups have been studied extensively, the psychological underpinnings of such an impartial moral view have received less attention.

Interestingly, recent work found that participants expressed negative attitudes towards actors who sacrificed one to save a larger number in sacrificial dilemmas (Everett, Pizarro, & Crockett, 2016) as well as towards actors who acted impartially by not giving priority to those to whom they have close personal relationships (Hughes, 2017). It is not surprising that both aspects of utilitarianism are regarded with suspicion—the utilitarian ideal of impartial beneficence was already ridiculed by Charles Dickens in *Bleak House* in his character of Mrs Jellyby, a 'telescopic philanthropist' who is obsessed with aiding a remote African tribe while showing little concern for her own family (Dickens, 1853). At the same time, such an impartial attitude is unlikely to arouse the aversive reaction most people feel towards the idea of violently sacrificing an innocent person—let alone towards the idea of infanticide. Moreover, Hughes (2017) found that impartial acts were regarded as less moral than partial ones because of inferences about lack of empathy and compassion; yet we found that Impartial Beneficence is closely tied to empathic concern. In addition, Hughes (2017) focused on cases were actors do not give priority to those close to them. Attitudes towards actors who engage in acts of extreme self-sacrifice are likely to be more ambivalent, and may be influenced by whether such acts are regarded by the actor is supererogatory (i.e. going beyond duty) or, in line with utilitarianism, as obligatory—and thus as issuing a similar demand to others.

The evidence currently suggests that the variance in moral views about instrumental harm is largely driven by differences in aversion to harming. Individuals who feel a strong aversion to harming tend to reject instrumental harm; those with less or no such aversion regard it more permissively. There is at present less evidence about the sources of differences in Impartial Beneficence. It is very likely that low Impartial Beneficence scores, indicating strong partiality, reflect either a strongly self-centered outlook and reluctance to make significant sacrifices to others, or a strong commitment to various narrow ingroups of the kind measured, for example, by the Loyalty sub-scaled of the Moral Foundations Questionnaire (Graham et al., 2011). But while factors driving strongly partial views are well-studied, the factors that drive above average impartial views, all the way to the radical impartiality of utilitarianism, are not as well understood. Our findings suggest that impartiality might be driven by high empathic concern, perhaps leading to greater identification with the whole of humanity (McFarland, Web and Brown, 2012). On this hypothesis, Impartial Beneficence is largely driven by an emotional disposition and an extension of one's ingroup to cover all human beings and perhaps even all sentient beings. This is in contrast to the view defended by prominent utilitarians which says that such impartiality is based on rational reflection on the arbitrariness of privileging some humans over others or of giving moral significance to mere distance (Sidgwick, 1907; de Lazari-Radek & Singer, 2012).

While Impartial Beneficence and Instrumental Harm are two independent dimensions of moral judgment, they can interact in important ways. Views on Impartial Beneficence determine which beings are taken to fall within the scope of morality and the extent to which they are given moral weight. These considerations will in turn affect judgments about whether certain people

(who may be family members, compatriots or strangers) should be sacrificed to save others (who again may be close or distant from the agent (see Petrinovich et al., 1993; Swann et al., 2010)). Conversely, views on instrumental harm will affect how one will approach promoting the greater good: are we permitted to promote this impartial aim by engaging in cold calculation of cost and benefit, sacrificing some to help a greater number, or must our efforts to help others also respect the rights of individuals? This two-way relationship between the two factors offers further justification to measuring both of them in a single scale.

**The Psychological Sources of Utilitarianism and Anti-Utilitarianism**

Utilitarianism is often portrayed by its proponents as the product of clear-headed rational reflection, and resistance to utilitarianism as largely due to the way it clashes with powerful intuitions and emotions (such as those elicited by the idea of violently sacrificing an innocent person in order to maximize utility; Singer, 1974, 2005). Greene (2008) has argued that rather than deontology and consequentialism being “philosophical inventions”, they are better understood as “philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking, that have been part of the human repertoire for thousands of years” (p. 360). We concur that support for utilitarian (consequentialist) or deontological ethical principles is almost certainly driven by psychological dispositions and processes that are part of our basic mental toolkit. Where we disagree is with the notion that utilitarian thinking is the result of a single, unitary psychological pattern. Our findings suggest that there are two dissociable psychological sources of a full-blown utilitarian approach to ethics: one relating to impartial beneficence and one to instrumental harm.

While a unitary model of utilitarianism-as-cognitive (and deontology-as-emotional) seemed to be supported by earlier studies using sacrificial dilemmas that tied pro-sacrifice judgments to effortful deliberation (Greene et al., 2004), more recent work has related such judgments to reduced aversion to harming (Kahane, Everett et al. 2015). This latter work suggests therefore that pro-sacrifice judgments are largely driven by reduced emotion. In line with this interpretation, in the present study we found an association between the Instrumental Harm sub-scale and psychopathy and reduced empathic concern. The positive dimension of utilitarianism has also often been claimed to be based in rational reflection (Sidgwick, 1907; de Lazari-Radek and Singer, 2012)—Peter Singer's (1974) famous argument that there is no moral difference between letting a drowning child die and refusing to donate money to prevent the deaths in developing countries is based on an appeal to consistency, not on pulling at our heartstrings. As we saw, however, the present study also ties Impartial Beneficence to an affective disposition—indeed, to one that is exactly the reverse of that associated with Instrumental Harm. At the same time, neither sub-scale was significantly associated with Need for Cognition, a trait measure of motivation to engage in effortful cognition. These results are consonant with recent studies that found that extreme altruists who donated their kidneys to strangers exhibit higher empathic concern (Brethel-Haurwitz et al., 2016).

Importantly, our 2D model suggests a hitherto overlooked source of opposition to utilitarianism. Utilitarianism notoriously has many implications that many find highly counterintuitive and even repugnant. Individuals low on both Instrumental Harm and on Impartial Beneficence are likely to be strongly resistant to utilitarianism. However, the psychological independence and degree of tension between these two dimensions of

utilitarianism suggests a further obstacle to its acceptance: psychological factors such as empathic concern that make individuals receptive to one dimension may also make them resistant to the other. There is a degree of psychological instability *internal* to utilitarianism.

Instrumental Harm and Impartial Beneficence are two aspects of a single coherent philosophical theory but they come apart in the psychology of ordinary people. It is striking that while the two were only weakly associated in the lay population, they were strongly correlated in a sample of expert moral philosophers. It is unclear what explains this sharp contrast. One natural explanation would be that this change may be due to philosophical education, including exposure to explicit views and arguments that tie the two moral dimensions together. On the 2D model, individuals are likely to arrive at such an overall utilitarian view by following two distinct psychological paths. Some individuals—perhaps driven by unusually high levels of empathic concern—begin by endorsing a radically impartial vision of moral concern and, in an attempt to turn this endorsement into a coherent theory, eventually come to endorse forms of instrumental harm as well, in order to promote such impartial goals. Other individuals may start with greater acceptance of instrumental harm—likely driven by low levels of empathic concern—and a general rejection of traditional moral rules, and, seeking to find a systematic moral framework to replace the commonsense morality that they reject, come to endorse a sweeping impartial view. In both cases, reasoning may serve not as the impetus to the embryonic utilitarian view, but rather as a means to integrate two aspects of utilitarianism that are often psychologically opposed. Utilitarianism may be the product, not of pure rational reflection and argument, but of an attempt to bring pre-theoretical tendencies and intuitions into a coherent equilibrium (along the lines suggested, in the deontological context, by Holyoak & Powell, 2016). Further research

could investigate (1) the extent to which explicit endorsement of utilitarian views involves such adjustment, (2) whether one of the two starting points is predominant, and (3) whether the initially dominant dimension predicts the degree to which the behavior of utilitarians mirrors their theoretical commitments. One can predict, for example, that individuals who start out high only on Instrumental Harm give less money to charity compared to those who start out high on Impartial Beneficence.

An alternative explanation of the association between the two sub-scales in the expert sample is that the structure of the debate in current moral philosophy attracts those individuals in whom the two dimensions are already aligned. Possible support for this speculative hypothesis is the fact that the proportion of the lay population who could be described as having strong overall utilitarian tendencies—around 26% scored above the midpoint of the OUS (>4) and only 4% scored 5 or more—is not substantially lower than the self-reported view of professional philosophers as reported in a recent survey. Out of 931 participants, 23.6% favored or leaned towards consequentialism over deontology or virtue ethics; of these, only 9.7% endorsed consequentialism outright (Bourget & Chalmers, 2014). Since consequentialism includes both utilitarianism in its different forms as well as theories that ascribe value to things other than utility (e.g. to justice), these results suggest that the percentage of professional philosophers who endorse utilitarianism without qualification may not be much higher than the percentage of participants in our study who consistently endorsed strongly utilitarian views. Coupled with our own findings, this at least raises the possibility that full-blown endorsement of utilitarianism partly reflects an unusual pre-philosophical psychological disposition in a minority of

individuals, rather the end point of a process of philosophical reasoning from a common psychological starting point shared by non-utilitarians.

**Limitations and Future Directions**

The OUS does have a number of limitations. Some of these have been mentioned in passing in the discussion above but it is worth making them explicit. First, the OUS is a measure of individual differences in moral outlook, not of the processes and mechanisms that underlie a particular episode of moral decision-making; it measures traits rather than states. Second, the items of the OUS-IB largely focus on self-sacrifice and less on impartiality with respect to others: a more fine-grained approach may be needed to investigate different strands of the psychology of moral partiality and impartiality.

A third potential limitation of the OUS is that it doesn't directly measure the calculating dimension of utilitarianism. Classical utilitarians not only claim that we should impartially consider the good of all in our moral decisions—they also hold that we are morally required to *maximize* that good, and it is this calculating and maximizing aspect of utilitarianism that some find problematic. Many items on both sub-scales of the OUS implicitly involve comparisons of overall utility (sacrificing one's leg to save another's life; killing some innocent people to save a greater number), and the OUS was positively associated with an explicit statement of utilitarianism that included this maximizing component. But the scale admittedly does not directly measure this dimension of utilitarianism. This is for two reasons. First, in order to directly measure such a maximizing tendency, we would need to see whether subjects would endorse, not only sacrificing 1 to save 5, but 1 to save 2, and 20 to save 21 (Kahane 2015). A short item scale is not the best tool for this achieving this end. Second, it is doubtful whether

such a maximizing tendency is a distinctive moral phenomenon rather than a general decision-making tendency—an egoist driven by pure self-interest can also engage in elaborate cost-benefit analysis when others may merely satisfice. We concede, however, that the question of whether such a maximizing tendency reflects a distinctive dimension of moral decision-making and, possibly, even a further dimension of proto-utilitarian tendencies, is worth investigating.

Finally, we wish to emphasize that the OUS is meant to be a measure of an overall pattern of moral views and judgments, not behaviour or intentions to act. Individuals may strongly endorse instrumental harm yet find it difficult to sacrifice someone to save a greater number if actually confronted with such a decision in real life. Similarly, someone may endorse a highly impartial vision of morality yet fail, for example, to donate much money to relevant effective charities. Indeed, several studies found that the moral behavior of moral philosophers does not significantly differ from that of others (Schwizgebel, 2009; Schwizgebel and Rust, 2009; Schwizgebel and Rust, 2014). We did find that impartial beneficence was positively associated with greater rates of hypothetical donation to charity but further research is needed to clarify the relationship between the sub-scales of the OUS and relevant forms of actual moral behavior.

**Conclusion**

The use of sacrificial dilemmas to study the contrast between utilitarian and deontological judgments has dominated research in moral psychology. But while the psychological sources of this major ethical dispute is of great interest, sacrificial dilemmas are a limited tool for studying proto-utilitarian tendencies in the lay population (Kahane & Shackel, 2011; Kahane, Everett et al., 2015; Kahane, 2015). In this paper we have introduced the 2D model of utilitarian decision-making, a new theoretical framework for studying such tendencies. The 2D model treats

utilitarian moral decision-making not as an all-or-nothing category but as a matter of degree, and as involving two independent ‘positive’ and ‘negative’ dimensions. On this basis, we developed the OUS, a new scale that is both philosophically rigorous and empirically driven, and which attempts to address concerns about sacrificial dilemmas and existing scales. Our preliminary application of the scale already demonstrates how the distinction between impartial beneficence and instrumental harm can help clarify the relationship between a utilitarian moral outlook and a range of other psychological constructs and measures while generating important new avenues for further research. Importantly, our results strongly suggest that, in the context of the lay population, utilitarian decision-making does not constitute a unitary psychological phenomenon.

The division between impartial beneficence and instrumental harm may also have important practical implications. Ethicists who wish to promote wider acceptance of utilitarian moral approaches in the general population may need to divide their efforts. Those individuals who are more likely to endorse instrumental harm, or generally be willing to dismiss or discount traditional moral rules, may at the same time be indifferent to—or even especially hostile—to the overarching moral aim of impartially maximizing the good of all sentient beings. Singer’s session on effective altruism at Victoria University drew those who were excited by the idea of Impartial Beneficence—but also a group of outraged protestors repelled by Instrumental Harm. To the extent that the positive aim of utilitarianism has greater moral priority, utilitarians would be advised to downplay the negative component of their doctrine and may even find a surprisingly pliant audience in the religious population.

But while such a strategy may be a more effective way of promoting aspects of utilitarian thinking, the apparent psychological disconnect between the ‘positive’ and ‘negative’ dimensions

of utilitarianism—an emotionally-based impartial concern for all, and a ‘hard-hearted’, unemotional attitude to instrumental harm—suggests that the prospect of full-blown, unreserved acceptance of utilitarianism by more than a small minority faces formidable psychological obstacles.

## References

Anscombe, G. E. M. (1981). *Ethics, Religion and Politics: Collected Philosophical Papers Volume III*. Oxford, England: Basil Blackwell.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479. https://doi.org/10.1037/0003-066X.54.7.462

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161.

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer*. Hillsdale, NJ: Erlbaum.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology: external validity in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Bekkers, R., & Wiepking, P. (2011). A Literature Review of Empirical Studies of Philanthropy: Eight Mechanisms That Drive Charitable Giving. *Nonprofit and Voluntary Sector Quarterly*, *40*(5), 924–973. https://doi.org/10.1177/0899764010380927

Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior*, *28*(3), 168–175. https://doi.org/10.1016/j.evolhumbehav.2006.10.003

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, England: Oxford University Press. (Original work published 1789)

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825–829.

Bloom, P. (2017). *Against empathy: The case for rational compassion*. London, England: Random House. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=_eslDAAAQBAJ&oi=fnd&pg=PT2&ots=gYNcLQGJhd&sig=IfoG8Ew500aOWLx9lz1WQgaoyps

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies*, *170*(3), 465–500.

Brañas-Garza, P. (2006). Poverty in dictator games: Awakening solidarity. *Journal of Economic Behavior & Organization*, *60*(3), 306–320. https://doi.org/10.1016/j.jebo.2004.10.005

Brethel-Haurwitz, K. M., Stoycos, S. A., Cardinale, E. M., Huebner, B., & Marsh, A. A. (2016). Is costly punishment altruistic? Exploring rejection of unfair offers in the Ultimatum Game in real-world altruists. *Scientific Reports*, *6*, srep18974. https://doi.org/10.1038/srep18974

Cacioppo, J., Petty, R., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.

Camosy, C. C. (2012). *Peter Singer and Christian ethics: Beyond polarization*. Cambridge University Press.

Cattell, R. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York, N.Y: Plenum.

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives? *Judgment and Decision Making*, *9*(4), 303.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology* (Vol. xiii). New York, NY, US: Guilford Press.

Charities Aid Foundation. (2016). *UK Giving 2015: An overview of charitable giving in the UK during 2015*. Retrieved from https://www.cafonline.org/docs/default-source/personal-giving/caf_ukgiving2015_1891a_web_230516.pdf?sfvrsn=2

Choe, S. Y., & Min, K.-H. (2011). Who makes utilitarian judgments? The influences of emotions on utilitarian judgments. *Judgment and Decision Making*, *6*(7), 580–592.

Comrey, A. L., & Lee, H. B. (1992). *A First Course in Factor Analysis*. Hillsdale, NJ: Erlbaum.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216.

Cullity, G. (2004) *The Moral Demand of Affluence.* Oxford: Oxford University Press.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001

Davies, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*(85).

Dawel, A., O'Kearney, R., McKone, E., & Palermo, R. (2012). Not just fear and sadness: Meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2288–2304. https://doi.org/10.1016/j.neubiorev.2012.08.006

Decety, J., Lewis, K. L. & Cowell, J. M. (2015). Specific electrophysiological components disentangle affective sharing and empathic concern in psychopathy, *Journal of Neurophysiology*, 114 (1) 493-504; DOI:10.1152/jn.00253.2015

de Lazari-Radek, K. de, & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics*, *123*(1), 9–31.

de Waal, F. B. M. (2007). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, *59*(1), 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

Dickens, C. (1853). *Bleak House*. London, England: Bradbury and Evans.

Eckel, C. C., & Grossman, P. J. (1996). Altruism in Anonymous Dictator Games. *Games and Economic Behavior*, *16*(2), 181–191. https://doi.org/10.1006/game.1996.0081

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, *14*(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7

Everett, J. A. C., Haque, O. S., & Rand, D. G. (2016). How good is the Samaritan, and why? An experimental investigation of the extent and nature of religious prosociality using economic games. *Social Psychological and Personality Science*, *7*(3), 248–255. https://doi.org/10.1177/1948550616632577

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology. General*, *145*(6), 772–787. https://doi.org/10.1037/xge0000165

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *The British Journal of Psychiatry*, *126*(3), 237–240.

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

Francis. (2013, July 6). What would Jesus drive? Pope tells priests to buy “humble” cars. *Reuters*. Retrieved from http://www.reuters.com/article/pope-cars-idUSL5N0FC0IR20130706

Fried, C. (1978). *Right and Wrong*. Harvard University Press.

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PloS One*, *8*(4), e60418.

Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. *Judgment and Decision Making*, *5*(7), 497.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Gray, J. (2015, May 21). How & How Not to Be Good. *The New York Review of Books*. Retrieved from http://www.nybooks.com/articles/2015/05/21/how-and-how-not-to-be-good/

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323.

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). Cambridge, MA, US: MIT Press.

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

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Vintage.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science, 316*(5831), 1622–1625. https://doi.org/10.1126/science.1140738

Hare, R. M. (1981). *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press.

Huber, S., & Huber, O. W. (2012). The centrality of religiosity scale (CRS). *Religions*, *3*(3), 710–724. https://doi.org/10.3390/rel3030710

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology*, n/a–n/a. https://doi.org/10.1111/bjso.12199

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hursthouse, R. (1999). *On virtue ethics*. OUP Oxford.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding Libertarian morality: the psychological dispositions of self-identified Libertarians. *PLOS ONE*, *7*(8), e42366. https://doi.org/10.1371/journal.pone.0042366

John Paul II. (1995, March 25). Evangelium Vitae. Papal Encyclical.

Jones, B., & Rachlin, H. (2006). Social Discounting. *Psychological Science*, *17*(4), 283–286. https://doi.org/10.1111/j.1467-9280.2006.01699.x

Jordan, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct? *Emotion*, *16*(8), 1107–1116. https://doi.org/10.1037/emo0000228

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *0*(0), 1–10. https://doi.org/10.1080/17470919.2015.1023400

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

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*, *7*(4), 393–402. https://doi.org/10.1093/scan/nsr005

Kenny, D. (2015, November 24). Measuring Model Fit. Retrieved March 18, 2017, from http://davidakenny.net/cm/fit.htm

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience, 7*(6), 708–714.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*(7138), 908–911. https://doi.org/10.1038/nature05631

Krebs, D. (1982). Altruism: A rational approach. In N. Eisenberg, *The development of prosocial behavior* (pp. 53–76). New York, NY: Academic Press.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, *68*(1), 151–158. https://doi.org/10.1037/0022-3514.68.1.151

MacAskill, W. (2015). *Doing good better: How effective altruism can help you make a difference*. New York, N.Y: Avery.

Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *32*(3), 454–465. https://doi.org/10.1016/j.neubiorev.2007.08.003

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists.

*Proceedings of the National Academy of Sciences*, *111*(42), 15036–15041. https://doi.org/10.1073/pnas.1408440111

McFarland, S., Webb, M., & Brown, D. (2012). All humanity is my ingroup: A measure and studies of identification with all humanity. *Journal of Personality and Social Psychology*, *103*(5), 830.

Mill, J. S. (1863). *Utilitarianism*. London, England: Parker, Son, and Bourne.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgment and Decision Making*, *5*(7), 547.

Oppenheimer, D. M., & Olivola, C. Y. (2011). *The science of giving: Experimental approaches to the study of charity*. New York, NY, US: Psychology Press. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=kCZ6AgAAQBAJ&oi=fnd&pg=PT14&dq=The+science+of+giving:+Experimental+approaches+to+the+study+of+charity&ots=lsQiRMp2Tu&sig=BCKca1gdZMYZ_ToYXasxK-58zTA

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

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are “counter-intuitive” deontological judgments really counter-intuitive? An empirical reply to. *Social Cognitive and Affective Neuroscience*, *9*(9), 1368–1371. https://doi.org/10.1093/scan/nst102

Petrinovich, L., O’Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology*, *64*(3), 467–478. https://doi.org/10.1037/0022-3514.64.3.467

Piazza, J. (2012). “If you love me keep my commandments”: Religiosity increases preference for rule-based moral arguments. *The International Journal for the Psychology of Religion, 22*(4), 285–302. https://doi.org/10.1080/10508619.2011.638598

Piazza, J., & Sousa, P. (2013). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*, 1948550613492826. https://doi.org/10.1177/1948550613492826

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. Penguin UK.

Robinson, J. S. (2012, December 6). *The consequentialist scale: Elucidating the role of deontological and utilitarian beliefs in moral judgments* (Thesis). Retrieved from https://tspace.library.utoronto.ca/handle/1807/33868

Scheffler, S. (1982). *The Rejection of Consequentialism*, Oxford: Clarendon Press. Revised edition 1994.

Sen, A. (1982). Rights and Agency, *Philosophy and Public Affairs*, 11 (1): 3–39.

Schneewind, J. B. (1977). *Sidgwick’s ethics and Victorian moral philosophy*. Clarendon Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Sidgwick, H. (1901). *Methods of ethics*. Kaplan Pub.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, *1*(3), 229–243.

Singer, P. (2011). *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

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

Sinnott-Armstrong, W. (2015), 'Consequentialism', *Stanford Encyclopedia of Philosophy* (online resource).

Smart, J. J. C. (1956). Extreme and Restricted Utilitarianism. *Philosophical Quarterly*, *6*(25), 344–354.

Smart, J. J. C. (1961). *An Outline of a System of Utilitarian Ethics*. [Carlton]Melbourne University Press on Behalf of the University of Adelaide.

Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and Against*. Cambridge, England: Cambridge University Press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245.

Swann, W. B., Gomez, A., Dovidio, J. F., Hart, S., & Jetten, J. (2010). Dying and killing for one's group: Identity fusion moderates responses to intergroup versions of the trolley problem. *Psychological Science*, *21*(8), 1176–1183.

Szekely, R. D., Opre, A., & Miu, A. C. (2015). Religiosity enhances emotion and deontological choice in moral dilemmas. *Personality and Individual Differences*, *79*, 104–109. https://doi.org/10.1016/j.paid.2015.01.036

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395–1415. https://doi.org/10.2307/796133

Van Lange, P. A. M., Bekkers, R., Schuyt, T. N. M., & Vugt, M. V. (2007). From games to giving: social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, *29*(4), 375–384. https://doi.org/10.1080/01973530701665223

Vekaria, K. M., Brethel-Haurwitz, K. M., Cardinale, E. M., Stoycos, S. A., & Marsh, A. A. (2017). Social discounting and distance perceptions in costly altruism. *Nature Human Behaviour*, *1*, 0100.

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition*, *126*(3), 364–372.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly*, *63*(2), 116–132. https://doi.org/10.2307/2695887

## Footnotes

[1] Some non-utilitarian moral views, including Kantianism and some forms of Christian and Buddhist ethics, are also radically impartial (e.g. Kant tells us to give equal respect to all rational beings) but that impartiality is expressed in the goal of maximising the well-being of all.

[2] At the theoretical level, it is possible to understand giving moral weight to different forms of partiality as themselves reflecting a set of deontological rules. But in the psychological context it is more illuminating to treat this dimension separately.

[3] We also asked the experts to indicate "How much do you personally agree with this statement?" (1 = not at all; 5 = very much). This data was not used in the present study.

[4] Interestingly, item 4 on the Impartial Beneficence sub-scale reflects a rejection of a moral distinction between acts and omissions—one of the most counterintuitive aspects of utilitarianism. While this item is directly concerned with impartial concern for the good of all, one of the key applications of this idea is in the context of duties to aid distant others (see e.g. Singer, 1972). Notice also that while this item refers to causing harm, its main point is that allowing harm to occur via omission is *as* morally bad as directly causing harm—by contrast with the items in the Instrumental Harm sub-scale, which involve the claim that certain ways of directly causing harm are *not* morally wrong.

[5] While the OUS scores correlated well with an explicit statement of utilitarianism, it is also worth highlighting at this junction that we would not recommend using such an explicit statement of utilitarianism as an independent measure of utilitarian tendencies. First, the explicit description does not completely distinguish between the positive and negative components of utilitarianism. Second, and relatedly, although OUS scores correlated well with the explicit statement, this was far from a perfect correlation, and OUS scores often had stronger correlations with the other measures than did the explicit statement. For example, as reported below, scores on the OUS-IB had a much stronger correlation with perceptions of wrongness on the Greater Good dilemmas from Kahane et al. (discussed below: $r = .50$, $p < .001$), whereas the correlation with the explicit statement of utilitarianism was substantially weaker ($r = .27$, $p < .001$).

[6] Our 2D model posits that there are two dissociable components to utilitarian decision-making. This is different from Greene's dual process model, which posits two systems that underlie deontology and utilitarianism, but treats utilitarianism as a unitary psychological construct. On Greene's model, the 'dual' refers to automatic vs. controlled processes underlying deontological vs. utilitarian judgments respectively; in our model, the 'two' refers to two different components within utilitarianism.

[7] Fully impartial beneficence regards the well-being of all morally relevant subjects as equally important. Exactly which beings fall within the scope of such concern will depend on further questions about well-being, the relation between well-being and sentience, and which living things possess sentience in the relevant sense (a utilitarian may give equal consideration to the well-being of all sentient beings but hold that e.g. fish are not sentient in the relevant sense).