Impartial Beneficence: The Forgotten Core of Utilitarian Psychology

Jim A.C. Everett
University of Kent

Prior work on moral reasoning has relied on sacrificial moral dilemmas to study utilitarian versus non-utilitarian decision-making. This research has generated important insights into people's attitudes toward instrumental harm—the sacrifice of an individual to save a greater number. But this approach has serious limitations. Most notably, it ignores impartial beneficence—the positive, altruistic core of utilitarianism, characterized by a radically impartial concern for the well-being of others. Here, I describe the two-dimensional model of utilitarianism, showing that instrumental harm and impartial beneficence are both conceptually and psychologically distinct. I review evidence showing that they have different patterns of individual differences, associated underlying processes, and consequences for how moral decision-makers are perceived. Acknowledging the dissociation between instrumental harm and impartial beneficence in the thinking of ordinary people has helped clarify existing debates about the nature of moral psychology, its relation to moral philosophy, and helps generate fruitful avenues for further research.

Everett, J. (2023). Impartial beneficence—the forgotten core of utilitarian psychology. In *The Routledge International Handbook of the Psychology of Morality* (pp. 40-50). Routledge.

Impartial Beneficence: The Forgotten Core of Utilitarian Psychology

Utilitarianism in its purest form is a radically simple moral philosophy: it holds that the whole of morality can be deduced from the single general principle that we should always act in the way that would impartially maximise aggregate well-being. It is at first glance a highly attractive philosophy, intuitively appealing from the start – who *wouldn't* want to maximise the "greatest good for the greatest number"? Utilitarianism has attracted its devotees, people inspired by the theory's deceptively simple basis and its role in progressive calls to do *more* good for *more* people – for example by calling for wealthy Westerners to donate more of their income to those in the developing world (e.g. Singer, 2015) or advocating for animals to be included within our typical moral circle and therefore refrain from eating meat (e.g. Singer, 2015). But utilitarianism has also been resoundingly criticised for the way it can seemingly be used to justify terrible decisions that harm others – torture, murder, even infanticide. Utilitarianism remains deeply controversial: beloved and detested in equal measure.

Utilitarianism is a philosophical theory, primarily discussed by ethicists debating the nature of right and wrong. But it has also been highly influential in moral psychology, with psychologists often describing ordinary people as engaging in *utilitarian reasoning* or making *utilitarian decisions.* When it comes to traditional moral psychological work on utilitarianism, however, one will read little about charity, animal rights, or self-sacrifice. Instead, one will read about runaway trolleys, torturing terrorists, and strange medical experiments. This research has generated important insights into people's attitudes toward instrumental harm—harm to some to help a greater number. However, this approach also has serious limitations. Most notably, it ignores impartial beneficence—the positive, altruistic core of utilitarianism, characterized by the impartial and equal concern for the well-being of others. In this piece I will attempt to plug that gap, highlighting the importance of taking a multidimensional approach to understanding utilitarian psychology. I will show that these two dimensions, of instrumental harm and impartial beneficence, are conceptually and psychologically distinct; that they exhibit different patterns of individual differences; that they seem to rely on different underling psychological processes; and that they have distinct social consequences. I will show that by moving beyond sacrificial dilemmas to this positive, forgotten core of utilitarian psychology, we can shed new light on old questions, and see glimpses of new questions to be asked.

Trolleyology

Sacrificial moral dilemmas are a staple of literature, theatre, films, and – in the last two decades – moral psychology. Whether it is runaway trolleys, burning buildings, or highly dubious medical procedures, philosophy undergraduates have long been forced to grapple with a central question in

ethics: When, if ever, is it acceptable to cause harm to some for the benefit of a greater number? Philosophers have long engaged in vigorous debates how we *should* respond to such questions, contemplating when and why it is acceptable to endorse instrumental harm. But in the last twenty years, psychologists have jumped in too.

To try and understand how, when, and why people engage in (non)utilitarian reasoning[1], moral psychologists have used sacrificial dilemmas like the "trolley problem". Inspired by – and originally directly using - the trolley dilemmas from philosophy (Foot, 1967), in the sacrificial dilemma paradigm participants are typically asked whether it is morally right to sacrifice one person to save the lives of five people. The classic finding is that when the sacrifice is achieved by 'impersonal' means (e.g. switching a runaway trolley to a different track), most people endorse it, though when it is requires 'personal' means (e.g. pushing someone off a footbridge), a large majority rejects it as immoral (e.g. Greene 2001). While classic utilitarianism says we should always sacrifice one to save the greater number, then, troublesome humans do not agree – at least in more direct, personal, physically confronting cases. This, in psychological terms, can be described as the "trolley problem": why it is that we endorse instrumental harm in sacrificial dilemmas in some cases, but not others?

The classic and highly influential answer comes from the dual process model of morality (DPM, e.g. Conway et al. 2018; Greene, 2008, 2014; Greene et al. 2001, 2004; see also Conway, this volume). Dual process models in psychology describe cognition as resulting from the competition between quick, intuitive, and automatic processes, and slow, deliberative, and controlled processes (e.g. Chaiken & Trope, 1999). In the classic trolley problem, Greene et al. (2001) have argued that in impersonal dilemmas the utilitarian-consistent decision to sacrifice is driven by controlled, cognitive processes, while automatic, intuitive processes are activated exclusively in personal dilemmas because of the emotional aversion to harm that such dilemmas involve. When individuals make the pro-sacrificial decision—often called *utilitarian decisions*[1]—, it is thought that they employ deliberative processing to repress their initial intuition and solve the dilemma using utilitarian cost-benefit analysis. Building on this, the DPM suggests non-utilitarian (often referred to as *deontological*) aspects of our moral decision-making are based in intuitive gut-reactions, while utilitarian decisions (e.g. sacrificing one to save a greater number) are uniquely attributable to effortful moral reasoning (see also Conway, this volume, for differences between hard vs. soft dual process models).

As can be seen in Conway's chapter in this same volume, moral psychology has adopted the sacrificial dilemma paradigm with vigor in the 20 years since Greene et al. published their seminal *Science* paper. Research shows, for example, that participants typically take longer to make pro-

[1] For consistency with other chapters in this volume I use the terms moral reasoning and moral decisions or moral decision-making, instead of the more commonly used "moral judgments"—which is used in this handbook to indicate how people evaluate others and others' moral choices. This should not be taken as indicating that when people are responding in sacrificial dilemmas they are (always) engaging in any deliberative reasoning process.

sacrificial decisions – suggesting these decisions are dependent on more controlled, deliberative processes - and pro-sacrificial decisions are associated with stronger activation in brain regions that support controlled, deliberative processes, such as the dlPFC. Most recently (and impressively), recent work by Patil and colleagues (2021) shows across eight studies using a variety of self-report, behavioral performance, and neuroanatomical measures, that individual differences in reasoning ability and cognitive style of thinking are associated with increased pro-sacrificial decisions. Part of the reason that the sacrificial dilemma paradigm and DPM has been so influential is that it aims to shed light not just on how people respond to artificial trolley dilemmas, but how, when, and why people do – or do not – engage in utilitarian reasoning more generally. Sacrificial dilemmas are typically taken as being the central source of conflict of utilitarian and non-utilitarian ethical approaches, with the idea that through studying sacrificial decisions we can understand why people engage in utilitarian reasoning in general (see Kahane & Everett 2022 for extended discussion on the role of the sacrificial dilemma paradigm in psychology). But there is a piece of the puzzle missing, the central, positive core of utilitarianism: impartial beneficence.

A Missing Piece: Impartial Beneficence

From its conception, utilitarianism has been a radically demanding and progressive ethical view. Philosophically, classical utilitarianism is neither solely about sacrificial harm or the rather mundane view that we should think about whether actions have positive consequences for well-being. Utilitarianism makes the much more radical claim that we must impartially maximize the well-being of all persons, rather than the rights of any specific individual, regardless of our personal, emotional, spatial, or temporal distance from the people involved (the *positive dimension,* or what we call *"impartial beneficence"*); and that this aim is not constrained by any other moral rule, including those forbidding us from intentionally harming innocent others (the *negative dimension,* or what we call *"instrumental harm"*). This, in a simplified form, is utilitarianism philosophically. The two dimensions are connected, but dissociable. They are connected because instrumental harm can be seen as consequence of impartial beneficence in utilitarianism: if all that matter is impartially maximising overall, aggregate welfare (impartial beneficence), then sometimes that might mean we have to harm some people in order to bring about a better state of affairs for a greater number of people (instrumental harm). But they are also dissociable: Kantian ethics might, for example, endorse some aspects of impartial beneficence by saying that we must give equal respect to all rational beings (see Mihailov, 2021), but reject instrumental harm by saying it is not acceptable to use one person as a means to an end, as in the classic footbridge dilemma.

Much work on utilitarianism over the last two decades has focused on decisions in the negative dimension of instrumental harm, measured in sacrificial dilemmas. But this is far from the central motivating claim of utilitarianism, nor is it even the only interesting claim. Utilitarianism tells us to

not only maximize our own well-being, or those close to us, but rather to maximize the well-being of all other sentient beings on the planet (Bentham, 1789/1983). It is for this reason that utilitarians have historically been leading figures in efforts against sexism, racism, and 'speciesism;' key advocates of political and sexual liberty; and key actors in attempts to eradicate poverty in developing countries. Take the leading utilitarian philosopher Peter Singer, who argued extensively for the more "positive" aspects of utilitarianism theory through advocating for animals to be included within our moral circle (Singer, 1975) and highlighting the demands of relatively affluent Westerners to do much more good to help those in other countries, even at significant personal cost (Singer, 2015). As I will show in the remainder of this chapter, understanding this positive core dimension of impartial beneficence is central to understanding utilitarian psychology more generally.

The Two-Dimensional Model of Utilitarian Psychology

The *two-dimensional (2D) model of utilitarian psychology* (Everett & Kahane 2020; Kahane & Everett et al. 2018, Kahane & Everett, 2022;) brings together this missing piece in the psychology study of utilitarianism. The model is based on the recognition that there are at least two primary ways in which utilitarianism, as a philosophical theory, departs from our common-sense moral intuitions: first, it permits harming innocent individuals when this maximises aggregate utility (*instrumental harm*); and second, it tells us to treat interests of other individuals as equally morally important, without giving priority to oneself or those to whom one is especially close (*impartial beneficence*).

There is a growing amount of evidence that as well as being conceptually distinct, these two dimensions of utilitarianism are psychologically distinct too. For example, if utilitarianism is a single, unitary psychological construct (which is necessary for us to make conclusions about utilitarianism in general on the basis of questions about sacrificial dilemmas specifically), then we should see similarities in how people respond to different kinds of questions reflecting paradigmatic utilitarian judgments. Unfortunately, the existing evidence suggests that we do not (Kahane & Everett et al. 2015; 2018; see also Conway et al. 2018). For example, people who endorse "utilitarian" sacrifice of one person to save five in trolley-style dilemmas are *not* more likely to also endorse "utilitarian" maximisation of welfare in questions about helping people in far off countries, reducing suffering of animals, or making sacrifices now for future generations. In fact, sometimes people who make the first kind of "utilitarian" judgments are *less* likely to make the second kind of "utilitarian" judgment.

If the psychology of instrumental harm is meaningfully different from the psychology of impartial beneficence, it means that much of our previous work on "utilitarian psychology" has only told half the story at best. By focusing on the sacrificial dilemma paradigm, we have gained important insights into when, why, and how people endorse the instrumental harm of some in pursuit of the greater

good. But we cannot assume that these findings about the psychology of instrumental harm generalize to the psychology of impartial beneficence (the "generalizability question": Everett & Kahane, 2020). In fact, as I will show in the remainder of this piece, there is significant emerging evidence that the findings cannot generalize: that just as someone who endorses instrumental harm is not necessarily more likely to endorse impartial beneficence, the two dimensions are associated with contrasting patterns of individual differences; seemingly dependent on different psychological processes; and result in distinct social perceptions of others.

This insight, that instrumental harm and impartial beneficence are not only conceptually but also psychologically distinct, therefore sets up both a challenge and an opportunity for moral psychology. The disunity of "utilitarian psychology" results in a challenge because it means we need to revisit our conclusions about utilitarian psychology, where findings must be more appropriately reinterpreted as elucidating the psychology of instrumental harm specifically. But it also offers an opportunity: an opportunity to reconsider established findings, shed new light on seemingly settled questions in moral psychology, and an opportunity to come to a more complete understanding of how people come to endorse different aspects of utilitarianism.

Individual Differences

Distinguishing between instrumental harm and impartial beneficence encourages us to reconsider what individual differences have been associated with "utilitarian psychology". Much work, for example, has discussed the association between both clinical and subclinical psychopathy and pro-sacrificial instrumental harm decisions (Bartels & Pizarro, 2011; Kahane et al., 2015; Koenigs et al., 2012; Wiech et al., 2013 – but see Conway et al. 2018 for a contrasting view, as well as Conway, this volume). This has led to a rather unflattering picture in moral psychology of utilitarians as cold, unfeeling, even anti-social. But when we look at the endorsement of utilitarian impartial beneficence, we see a very different pattern. Indeed, participants' scores on the instrumental harm sub-scale of the Oxford Utilitarianism Scale (OUS) are positively associated with subclinical psychopathy and negatively associated with empathic concern (Kahane & Everett et al. 2018). But we find the opposite pattern is for impartial beneficence: people who endorse the utilitarian impartial maximisation of welfare (e.g. "It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal") are actually *less* likely to agree with statements tapping subclinical psychopathy (e.g. "For me, what's right is whatever I can get away with": Levenson et al. 1995) and *more* likely to agree with statements tapping empathic concern (e.g. "I often have tender, concerned feelings for people less fortunate than me.": Davis, 1980). That is, people who feel greater empathy and concern for others can indeed be more likely to endorse utilitarian principles – just in the domain of impartial beneficence, not instrumental harm.

What about socio-ideological attitudes, such as religiosity or political ideology? Religious systems have often focused on the importance of rule-based moral decision making, and utilitarianism has historically conflicted with religious views (Mill, 1863). In line with this, research using sacrificial dilemmas has reported that religiosity is associated with reduced utilitarian endorsement of instrumental harm (e.g. Piazza, 2012; Piazza & Sousa, 2014; Szekely, Opre, & Miu, 2015). When using our OUS measure of impartial beneficence, however, we have shown that religiosity is associated with increased utilitarian endorsement of impartial beneficence (Kahane & Everett et al. 2018). This makes sense when one thinks about the impartial, welfare maximising nature of some religious injunctions, and particularly standard accounts of Christian ethics which generally involve quite radical demands for self-sacrifice and impartiality. Indeed, 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). As we have noted before, Peter Singer has said almost identical things.

When it comes to political ideology, we again see the dissociation between instrumental harm and impartial beneficence. While political liberals are less likely to endorse utilitarian instrumental harm (e.g. thinking it is sometimes morally necessary for innocent people to die as collateral damage if more people are saved overall), they are actually more likely than their conservative counterparts to endorse utilitarian impartial beneficence (e.g. thinking that from a moral perspective, people should care about the well-being of all human beings on the planet equally, Kahane & Everett et al. 2018).

Underlying Processes

Just as treating utilitarianism as a non-unitary psychology construct sheds greater light on how different kinds of utilitarianism-consistent moral decisions are associated with different individual differences, there is also some evidence that the same is true for considering psychological processes underlying these decisions. According to the influential dual process model of utilitarian decision-making, deliberation favors 'utilitarian' reasoning whereas intuition favors 'deontological' decisions (e.g. Greene et al. 2001; Conway et al. 2018; Patil et al. 2021). As with much of the work, however, this has been conducted almost exclusively in the context of attitudes about instrumental harm, measured in sacrificial dilemmas. Would the same pattern emerge for impartial beneficence?

We have tested this in previous work, looking at how using classic manipulations of cognitive process might shift participants' endorsement of both impartial beneficence and instrumental harm (Capraro, Everett, & Earp, 2020). We conducted three studies in which we manipulated participants' cognitive process by priming intuition or deliberation (Levine et al., 2018), telling

participants to answer based "on your first, emotional response and your 'gut-feeling'…just focus on what your intuition tells you" or "on reason, rather than intuition. Focus on thinking and reasoning about the question… think carefully about each question". We then had participants complete both the instrumental harm and impartial beneficence items from the OUS. Across 3 studies and in line with past research, we found that those participants in the intuition-primed conditions endorsed utilitarian statements about instrumental harm significantly less than those who were encouraged to make their decisions through careful deliberation. Importantly, however, this was not the case for impartial beneficence. That is, priming intuition (vs deliberation) reduced utilitarian decisions about instrumental harm, but had no effect on impartial beneficence. Again, studying only sacrificial dilemmas but generalizing to utilitarian psychology at large can give a misleading picture.

Social perceptions

Finally, just as utilitarian reasoning about instrumental harm and impartial beneficence exhibit different psychological profiles with individual differences and potentially rely on different processes, there is also increasing evidence that they have different social consequences for how someone is perceived, and how trustworthy they are seen to be by others.

There is a large body of evidence now showing that people who endorse utilitarian instrumental harm in sacrificial dilemmas (e.g. endorsing sacrificing one person to save five) are seen as less moral and trustworthy, chosen less frequently as social partners, and trusted less in economic exchanges than those who reject it (Bostyn & Roets, 2017; Everett et al., 2016, 2018; Rom et al., 2017; Sacco et al., 2017; Uhlmann et al., 2013: see Crockett et al. 2021 for a review). We have explained this through reference to partner choice models (Everett et al. 2016; 2018), noting that the demands that utilitarianism makes in instrumental harm – the demand to break moral norms and even cause harm when it maximises the greater good – are seemingly incompatible with what we look for in social partners. When looking for friends or thinking about our family, we want to be able to trust that they will support us and do us no harm – even when hurting us would bring about the greater good.

But what about impartial beneficence dilemmas? Should, for example, someone spend their weekend cheering up their lonely mother or instead help re-build houses for families who have lost their homes in a flood? Or should someone give money to help a family member, or donate it to charity to provide life-saving help to many more individuals in a far-off country? Such dilemmas also seem to raise conflicts with what we seek in social partners. The impartial utilitarian standpoint seems to depart from what we want from friends and families because it denies the existence of "special obligations" to those with whom we have a close relationship. Saving my own child over two stranger's children may not maximise the good, but many deontological ethical approaches

suggest that it is morally permissible (even required), because I have *special* obligations to my child, ones that I do not have towards a stranger's children (see Jeske, 2014).

There is evidence that those who help a stranger instead of family members are judged as less morally good and trustworthy than those who did the opposite (McManus et al., 2020), and that this pattern of results is seen even when it is clear that helping strangers would maximise the greater good. For example, Hughes (2017) presented participants with a description of someone facing a dilemma between spending the weekend comforting their lonely mother or instead using the time to help rebuild homes for poor families through Habitat for Humanity. They found that participants who made the impartially beneficent welfare-maximizing decision to volunteer instead of see their mother were seen as having a worse moral character. Similarly, Law et al. (2021) show that socially distant altruists (e.g. endorsing donating money to save the life of a distant stranger in another country) tend to be seen as having a worse moral character than those who are socially close altruists (e.g. endorsing spending their money on a dream vacation for their terminally ill child). Interestingly, however, it appears that this potential negativity towards those who endorse impartial beneficence may be at least somewhat dependent on the type of social role these people have, with people who endorsed impartial beneficence being seen as a worse friend but a *better* political leader (Everett et al. 2018).

We have recently explored the way that endorsing instrumental harm and impartial beneficence might have distinct consequences in trust in political leaders specifically (Everett, Colombatto, et al. 2021). We conducted a Registered Report experiment, recruiting 23,000 participants in 22 countries over six continents. Participants completed both self-reported and behavioral measures of trust in leaders who endorsed utilitarian or non-utilitarian principles about instrumental harm or impartial beneficence in a series of real-world inspired dilemmas concerning the COVID-19 pandemic. For example, one instrumental harm dilemma concerned the permissibility of mandatory privacy-invading tracing devices to reduce the spread of the virus, and one impartial beneficence dilemma concerned whether resources such as medicine should be sent wherever in the world they should do the most good or reserved first for a country's own citizens. Our results showed that across both the self-reported and behavioural measures, endorsement of instrumental harm decreased trust, while endorsement of impartial beneficence increased trust. Just as impartial beneficence and instrumental harm are associated with different individual differences and seem to rely on different psychological processes, so too do they seem to have distinct social consequences for how people are perceived – all of which would be obscured if we treated utilitarianism as a unitary phenomenon, studying only sacrificial dilemmas.

Future Directions

The study of impartial beneficence, this forgotten but central core of utilitarian psychology, is woefully neglected when compared to the astonishing amount of research studying instrumental harm and sacrificial dilemmas. Luckily for the ambitious new researcher, this means that the fruit is ripe for picking. In this final section, I will briefly review just some of many future directions that could be especially exciting.

One particularly important new direction that is already happening is looking at how the endorsement of utilitarian impartial beneficence relates to real-world examples of impartial welfare maximisation. For example, we know that greater concern for animal suffering is positively correlated with impartial beneficence, but not instrumental harm (Caviola et al. 2020), and it will be interested to explore how real-world animal suffering activists think about these utilitarian principles. Similarly, there are strong links between the Effective Altruism movement (MacAskill, 2015) and the utilitarian principle of impartial beneficence, though it remains to be seen how those who fully embrace these principles (e.g. by pledging certain amounts of their income to effective charities in the developing world) come to hold their impartially beneficent views. In this vein, exciting recent work has looked at how real-world extraordinary altruists – those who donated a kidney to a stranger – scored higher on utilitarian impartial beneficence, but not instrumental harm (Amormino et al. 2022). It will be fruitful for future work to consider in more detail the psychology of impartial beneficence in such real-world contexts.

Another open area for future research is to understand the developmental trajectory of the (lack of) endorsement of impartial beneficence, and how early environments might influence adult moral judgments about the impartial maximisation of welfare. While some work has been done looking the development of sacrificial judgments (Caravati et al. 2017; Pellizzoni et al, 2010) or how unpredictable child environments shape adult judgments about instrumental harm (Maranges et al. 2021), we know little about impartial beneficence. How do children start thinking about (im)partiality in moral judgments when this conflicts with motivations to help more people? How and when might children develop ideas about the importance of undergoing small sacrifices for themselves to benefit strangers? It will be interesting for future research to consider such questions.

Finally, it will be interesting to understand "interventions" that promote impartial beneficence and encourage real-world behaviour that impartially maximises overall welfare. By building on our existing knowledge about the psychological barriers to effective altruism (e.g. see Berman, 2018; Caviola, 2021) we can start to consider interventions that can promote the endorsement of impartial beneficence in other behavioural contexts beyond charitable donations.

Conclusion

In this chapter I hope to have convinced you of the conceptual importance and psychological informativeness of treating utilitarianism not as a unitary construct in which we can base conclusions solely on the sacrificial dilemma paradigm. Instead, I hope to have shown the way that treating utilitarianism as a multidimensional construct, consisting of at least two main dimensions of both instrumental harm and impartial beneficence, can shed light on established topics in moral psychology and generate new directions in the field. Compared to instrumental harm, the endorsement of impartial beneficence is correlated with different patterns of individual differences, seems to rely on different underlying processes, and has different consequences for how moral decision-makers are perceived. We have spent two decades studying utilitarianism through focusing on sacrificial dilemmas and this work has generated many important insights. But I believe that this is only part of the story. There is another part of the study which is woefully underwritten – the story of the psychology of impartial beneficence.

References

Amormino, P., Ploe, M., & Marsh, A. (2022). Moral foundations, values, and reasoning in extraordinary altruists. Pre-print. 10.21203/rs.3.rs-1762722/v1

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

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.

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

Berman, J. Z., Barasch, A., Levine, E. E., & Small, D. A. (2018). Impediments to effective altruism: The role of subjective preferences in charitable giving. *Psychological science*, *29*(5), 834-844.

Bostyn, D. H., & Roets, A. (2017). Trust, Trolleys and Social Dilemmas: A Replication Study. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000295

Caviola, L., Schubert, S., & Greene, J. D. (2021). The psychology of (in) effective altruism. *Trends in Cognitive Sciences*, *25*(7), 596-607.

Caravita, S. C., De Silva, L. N., Pagani, V., Colombo, B., & Antonietti, A. (2017). Age-related differences in contribution of rule-based thinking toward moral evaluations. *Frontiers in psychology*, *8*, 597

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

Capraro, V., Everett, J. A. C., & Earp, B. D. (2020). Priming intuition decreases instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology*.

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.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the 4 of philosophers. *Cognition*, *179*, 241–265. https://doi.org/10.1016/j.cognition.2018.04.018

Crockett, M. J., Everett, J. A., Gill, M., & Siegel, J. Z. (2021). The relational logic of moral inference. In *Advances in Experimental Social Psychology* (Vol. 64, pp. 1-64). Academic Press.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy.

Everett, J. A. C., Colombatto, C., Awad, E., Boggio, P., Bos, B., Brady, W. J., Chawla, M., Chituc, V., Chung, D., Drupp, M., Goel, S., Grosskopf, B., Hjorth, F., Ji, A., Lin, Y., Ma, Y., Maréchal, M., Mancinelli, F., Mathys, C., ... Crockett, M. J. (2021). Moral dilemmas and trust in leaders during a global health crisis. *Nature Human Behaviour*.

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence.

Journal of Experimental Social Psychology, *79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Everett, J. A. C., & Kahane, G. (2020). Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.11.012

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

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

Francis, Pope (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

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343.

Greene, J. D. (2008). The secret joke of Kant's soul. In *Moral psychology, Vol 3: The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). MIT Press.

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

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.

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*, *56*(3), 561-577.

Jeske, D. (2014). Special obligations. In E. N. Zalta (Ed.). The Stanford encyclopedia of philosophy (Spring 2014)Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2014/entries/special-obligations/.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

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

Kahane, G & Everett, J.A.C (2022). Trolley Dilemmas: From Moral Philosophy to Cognitive Science and Back Again. Lillehammer, H. (Ed) *The Trolley Problem: Classic Philosophical Arguments Series*. Cambridge, UK: Cambridge University Press.

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

Law, K. F., Campbell, D., & Gaesser, B. (2022). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychology Bulletin*, *48*(3), 426-444.

Levine, E. E., Barasch, A., Rand, D., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General*, *147*(5), 702–719. https://doi.org/10.1037/xge0000399

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.

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, *31*(3), 227-242.

Maranges, H. M., Hasty, C. R., Maner, J. K., & Conway, P. (2021). The behavioral ecology of moral dilemmas: Childhood unpredictability, but not harshness, predicts less deontological and utilitarian responding. *Journal of Personality and Social Psychology*.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*, *120*(2), 443–460. https://doi.org/10.1037/pspp0000281

Piazza, J. (2012). "If you love me keep my commandments": Religiosity increases preference for rule-based moral arguments. *International Journal for the Psychology of Religion*, *22*(4), 285-302.

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*, *5*(3), 334-342.

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental science*, *13*(2), 265-270.

Sacco, D. F., Brown, M., Lustgraaf, C. J. N., & Hugenberg, K. (2017). The Adaptive Utility of Deontology: Deontological Moral Decision-Making Fosters Perceptions of Trust and Likeability. *Evolutionary Psychological Science*, *3*(2), 125–132. https://doi.org/10.1007/s40806-016-0080-6

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.

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.