Moral Psychology and Utilitarianism

Lucius Caviola & Joshua Greene

Guest essays represent only the views of the author(s).

Table of Contents

According to utilitarianism, we should improve the lives of humans and other sentient beings as much as possible. As an abstract ideal, utilitarianism has a natural appeal and may even sound like simple common sense. But utilitarianism has some implications—some merely theoretical, some very practical—that are counter-intuitive. When utilitarianism runs counter to our moral intuitions, is that because of a problem with utilitarianism or with our moral intuitions?

In this article, we discuss the different [1] [2] ways in which human moral psychology and behavior deviate from what utilitarianism prescribes. We focus on psychological deviations from key aspects of utilitarianism: impartiality, maximization, consequentialism, and aggregate-welfarist values. Finally, we

consider the normative implications of psychological science for utilitarianism. We conclude that the science of morality cannot show that utilitarianism is correct but that it can cast doubt on certain intuitive arguments against utilitarianism.

Deviations from Impartiality

Utilitarianism says that we should count everyone's interests equally and that no one's well-being is inherently more important than anyone else's. All individuals should be included in our circle of moral concern and—at least in theory—given equal weight, no matter who they are, where they are, and when they are. However, we know from everyday observation and psychological research that people are not that impartial. And in a world of limited resources, favoring some individuals means disfavoring others.

First, people prioritize ingroup members over outgroup members (**in-group favoritism**). They care more about family, friends, community members, co-religionists, co-nationals, and so on. More generally, people tend to care more about, and feel a stronger sense of responsibility towards, individuals who are close to them on various dimensions of distance. They feel more obligated to rescue a child who is drowning right in front of them than a child who lives in a poor country on the other end of the world. They feel greater empathy for currently alive people than future generations.[3] [4] They attribute higher moral status to humans than to animals, even in cases where animals have equal or higher mental capacities than humans (i.e., speciesism).[5] [6]

The pull of the ingroup is a universal and powerful force.[7] [8] Psychologically, it appears spontaneously when groups are formed arbitrarily based on distinctions such as color preferences.[9] In-group favoritism generally appears early in development,[10] [11] is shared with other primates,[12] [13] and is modulated by the same neuroendocrine pathways across many species.[14] It's even reflected in national laws that explicitly prioritize citizens over non-citizens.[15]

These observations suggest that ingroup favoritism is partly an innate human tendency. However, it is also culturally malleable. Westerners, for example, tend to be more individualistic and feel less bound by in-group relationships than others.[16] Further, the tendency to prioritize humans over animals is weaker in young children than in adults, suggesting that aspects of speciesism are socially acquired[17] and culturally malleable.[18]

Second, people prioritize themselves over most others. Even though humans are, in some ways, an unusually cooperative and altruistic species, we aren't as altruistic as utilitarianism demands. While utilitarianism considers it obligatory to provide substantial personal resources to improve the world and help others regardless of how far away they are, people consider it only supererogatory—nice, but not required. This familiar tendency is at odds with utilitarianism. Few people give everything they don't need to charity. Few people donate their spare kidney to save a stranger's life. And few people become vegan to avoid harming animals. And while an extreme amount of self-sacrifice would be psychologically unsustainable, most of us do far less than we could to promote the well-being of others.

Deviations from Maximization

According to utilitarianism, it's not enough to do *some* amount of good. Instead, we should do the *most* good.[19] For example, if we could save either one or two lives, we are obligated to save two. In other words, effectiveness is a moral imperative.[20] This component of utilitarianism is emphasized by effective altruism.[21]

But people rarely aim to maximize their positive impact on the world. A prime example is charitable giving, where few people support the most effective charities or even attempt to discover which charities are most effective—those that do the most good per dollar. Why is that? There are multiple psychological obstacles to doing the most good.[22][23]

On the one hand, there are epistemic obstacles. For example, most people are unaware that some ways of helping are orders of magnitude more effective than others.[20][21][22][23][24] Similarly, most people are uninformed about the most effective ways of doing good, such as the most effective charities.

But even when people are presented with all the relevant information, they often lack the cognitive tools necessary to discern which options will do the most good. Just as there are cognitive biases that prevent people from maximizing their financial profit,[25] biases can prevent people from maximizing their altruistic impact. Decision-making in a complex world under uncertainty is cognitively challenging. Maximizing one's impact requires employing counter-intuitive concepts from economics and applied mathematics such as expected value, marginal thinking, and opportunity costs. Yet, most people struggle with these concepts and fail to understand their implications for altruistic decisions.[22]

On the other hand, there are motivational obstacles. Even when people understand which options will maximize their altruistic impact, they may not be motivated to choose the option that does the most good. There are several reasons for this.

First, people's emotions don't scale with the numbers,[26] making them scope insensitive: Saving 1,000 lives instead of 10 doesn't feel 100 times better; it feels about the same.[27] Sometimes a single identifiable victim can trigger stronger emotional reactions than 100 "statistical" victims.[28][29] In other words, our emotions are largely "innumerate".

Second, most people want to give to a charity they find emotionally appealing even if they know that other charities could do more good per dollar.[30] Moreover, they view giving based on personal preferences, instead of effectiveness, as morally acceptable or even praiseworthy.[31] This thinking stems from a conception of charitable giving as supererogatory. And if giving

itself is optional, it naturally follows that one is free to give however one pleases, regardless of how (in)effective one's choices may be.[32]

Third, people are often averse to prioritizing some altruistic options over others. However, maximizing altruistic impact requires setting priorities, and prioritizing effectiveness inherently requires deprioritizing less effective options. In a simple, stylized case, this might mean letting two people die in order to save three different people—a classic case of triage. People tend to perceive this as cold or even unfair[33] because they view human life as a sacred value, meaning they are unwilling to make trade-offs involving human lives.[34] [35] Consequently, people prefer solutions that seemingly avoid making such tough trade-offs, such as splitting their donations across multiple causes, even if they are ultimately helping fewer people.[36]

Deviations from Consequentialism

Utilitarianism is a form of consequentialism, and according to consequentialism, the only thing that ultimately matters is producing good consequences. There are various ways in which typical human psychology deviates from this consequentialist standard.[37]

A particularly counter-intuitive aspect of consequentialism is that it allows for (or even requires) actions that seem clearly wrong so long as they do more good overall.[38] [39] [40] [41] At least in theory (but not necessarily in practice; see[42] [43]), utilitarianism rejects **deontological constraints**—rules that forbid certain actions, such as lying or causing direct harm, even when these actions would produce better consequences by preventing even greater harm. In this regard, consequentialism deviates from deontology, an ethical school of thought that conceives of morality not in terms of producing good consequences but in terms of rights and duties[44] (see also 'protected values'[45] and 'taboo trade-off aversion'[46]).

The most widely studied cases in which utilitarianism and deontology diverge are known as sacrificial dilemmas, the most famous of which are "trolley dilemmas".[47][48] In the *switch* case, a runaway trolley is headed toward five people who will be killed if nothing is done. But you can save the five by hitting a switch that will divert the trolley onto a sidetrack, where it will run over a single person. In the contrasting *footbridge* case, you are on a footbridge over the tracks between the oncoming trolley and the five. Next to you on the footbridge is a big man, and the only way to save the five is to push this man off the footbridge and onto the tracks. This will save the five by blocking the trolley, but the man pushed will die.

As philosophers long ago predicted and decades of research have confirmed,[49][50][51] most people approve of killing one to save five in the *switch* case, but not in the *footbridge* case. In the *footbridge* case, killing one to save many seems wrong to many people, even under the assumption that this action will promote the greater good and that there is no better alternative.

What makes people say "yes" to some cases (e.g., the *switch* case) and "no" to others (e.g., the *footbridge* case)? Research indicates that people are sensitive to whether the harm is caused as a *means* or as a *foreseen* side-effect. For example, in the *footbridge* case, the person on the bridge is used as a trolley-stopper, but in the *switch* case, the person is killed as a side-effect of turning the trolley. This distinction between means and side-effect also has a long and distinguished philosophical history as the "doctrine of double effect".[47][52][53]

The means vs. side-effect distinction, however, can't fully explain people's reluctance to push the man off the bridge. For example, in one study, people were far more likely to approve of using the man as a trolley-stopper when this can be done by hitting a switch that opens a trap door (59% approval) rather than by pushing with one's hands (31% approval).[54] Here, the key factor is "personal force," causing harm in a more physically direct way.[50][55][56] Recently, a large cross-cultural study tested the effect of personal force in 45 countries

on all inhabited continents. The results showed that the personal force factor, as well as the means-side effect factor, shapes moral judgments worldwide [57].

What psychological mechanisms explain people's reactions to these dilemmas? According to the **dual-process theory of moral judgment**,[49545859 60] the judgment that it's wrong to push in the *footbridge* case is driven by automatic emotional intuitions, while the judgment that it's acceptable to push in the *footbridge* case (or hit the switch in the *switch* case) are driven by conscious and controlled cost-benefit reasoning. More specifically, deontological responses are principally model-free, meaning that they are based on mental habits, attaching emotional values directly to actions based on their past consequences.[61626364] ("It just feels like the wrong thing to do"). By contrast, utilitarian/consequentialist judgments are considered model-based, generated by a model that represents cause-effect relationships and that generates predictions about consequences in the present context. ("Better to do the thing that will minimize the loss of life").

The original evidence for the dual-process theory came from neuroimaging studies,[4958] and more recent neuroimaging studies have supported and refined the theory.[60656667] Over the last two decades, much of the strongest evidence for the dual-process theory has come from studies of patients with deficits related to one of the two processes posited by the dual-process theory. Patients with damage to the ventromedial prefrontal cortex (VMPFC), like the famous Phineas Gage, have intact general reasoning capacities (often scoring very well on standard IQ tests), but they are unable to generate or integrate intuitive emotional responses into their decisions[68] leading to abnormal social behavior. The dual-process theory predicts that these patients, due to their emotional deficits, will give utilitarian responses to cases like footbridge and not just to cases like switch. This prediction has been confirmed in multiple studies of VMPFC patients,[697071] as well as in similar patients with frontotemporal dementia[72] and traumatic brain injury.[73] The same pattern is

observed in patients with low-anxiety psychopathy[74] and in people with alexithymia,[75][76] a disorder that detaches people from their emotions. A large multi-patient study[77] found that the neural regions and networks identified in the aforementioned fMRI studies using trolley-type dilemmas (specifically for deontological judgment) converge on the brain regions that, when lesioned, most reliably lead to criminal behavior. While patients with emotional deficits are more likely to give utilitarian responses, other patients are less likely to give utilitarian responses. Patients with damage to the hippocampus[78][79] (but see[80]) and the basolateral amygdala[64] both give fewer utilitarian responses. Both sets of patients report that their decision-making is more emotional, and emotion tends to dominate in these patients due to deficits in goal-directed reasoning, which relies on a mental model of cause-effect relationships between actions and outcomes.[61][62]

Behavioral research in healthy people has supported the dual process theory in various ways.[81][82] Research on individual differences using "process dissociation" has been used to assess the relative strength of utilitarian and deontological motivations within people.[63][83][84] Other studies have documented the "foreign language effect",[85][86] whereby people tend to give more utilitarian responses when dilemmas are presented in the respondent's second language —an effect that appears to be due to emotional dampening. Experiments using instructions or concurrent tasks designed to influence how people respond to moral dilemmas have likewise provided evidence for the dual-process theory using a wide range of methods.[87][88]

A further psychological obstacle to consequentialism is omission bias—the tendency to prefer harmful omissions over harmful actions.[89][90] People judge others more harshly for actively causing a harmful outcome than for failing to prevent the same harmful outcome ("doctrine of doing and allowing"[91]). For example, compare a tennis player who recommends poisonous food to his opponent to a tennis player who chooses not to prevent his opponent from

eating the poisonous food.[92] People say that the player who actively recommended the poisonous food to his opponent is morally worse, even though both players knowingly chose the same outcome. As a real-world example, omission bias has been shown to play a role in parents' decisions about vaccinating their children. Some parents are reluctant to vaccinate their children because they anticipate feeling worse if their child is harmed by the vaccine (a result of the parents' action) than if their child is harmed by the disease that the vaccine could have prevented (a result of the parents' omission).[93]

Deviations from Aggregate Welfarism

Utilitarianism favors maximizing well-being in an impartial way. As we've seen, people are uncomfortable with utilitarianism's commitments to full impartiality, prioritizing more effective altruistic strategies, and discounting competing deontological considerations. A further obstacle to embracing utilitarianism is its conception of what counts as good consequences. What consequences matter morally? And how does one assess overall consequences, not just for individuals, but for the world?

A key component of utilitarianism is "welfarism," the idea that outcomes are good or bad to the extent that they promote or undermine well-being—affecting levels of happiness and suffering or the fulfillment of preferences. Crucially, utilitarianism says that effects on well-being are the only consequences that matter *intrinsically*. Other consequences matter *extrinsically* because of their effects on well-being. To what extent do people endorse this view? While research on this question is limited, it appears that most people partially endorse the utilitarian conception of value. Most people think it's good to have more happiness and less suffering in the world.[94] But people also have additional values. People report valuing things such as purity, beauty, complexity, biodiversity, and wisdom, not only for their effects on

well-being but as ends in themselves [8] [95] Thus, people appear to have both welfarist and non-welfarist values.

How do people define well-being? Philosophers disagree about how to define it. So-called "hedonistic" theories see well-being in terms of positive or negative experiences, such as happiness and suffering. Is this all that matters? Robert Nozick's famous Experience Machine thought experiment challenges this assumption, asking whether we would prefer to live life plugged into a virtual reality machine that can stimulate our brains so as to safely and reliably deliver a maximally happy experience.[96] Nozick argues that we have good reasons not to plug in, including the value of having authentic (non-illusory) experiences and behavior. What do people choose? As Nozick predicted, most people say that they would choose real life over the machine.[97] However, when asked to assume that they were already living on an experience machine and told their life outside would be different or worse, most people said they would prefer to remain on the machine.[98] While these findings allow for multiple interpretations, they suggest that people are not just hedonists, valuing other things such as the authenticity of experience.

Consequences can be good or bad for different individuals, but how should we assess the overall value of a set of consequences? A central component of utilitarianism is **aggregationism**, the idea that the overall value is determined by the sum of the well-being of all individuals. To what extent do people hold aggregationist beliefs? If all else is equal, people favor worlds that contain more well-being.[94] But people often prefer to distribute well-being in ways that reduce aggregate well-being but that seem more fair. For example, people may have more egalitarian values, preferring worlds in which everyone has the same level of well-being over worlds where people have higher well-being on average but differ in their well-being levels.[99] [100] [101] [102] One form of egalitarianism is to prioritize the well-being of the least well-off over that of others, even if those others could have larger gains in well-being (prioritarianism).[103] And some people have retributivist intuitions, such that they prefer an immoral

person to suffer even if aggregate well-being is reduced and there are no further societal benefits.[104] Some philosophers think that small differences in well-being distributed across many people cannot outweigh big differences for a few. It's an open question whether this view is widely held.[105]

Comparing outcomes in worlds with a fixed set of individuals is complicated enough. But, in the longer term, actions can affect how many individuals, and which particular individuals, get to exist. The ethics of distributing value across such variable populations is the subject of **population ethics**.[106][107] When directly asked, people find it valuable to add a new happy individual to the world.[94][108] Similarly, people prefer worlds with more happy individuals than fewer happy individuals as long as each individual's happiness level is the same, suggesting that they value improving a population's total level of welfare. However, when individuals' happiness levels differ, people sometimes prefer worlds that have a higher average level of happiness, even if the total level of happiness declines. For example, people may prefer a world with 1,000 maximally happy individuals over a world with 100,000 individuals who are only half as happy.[94] In the suffering domain, people's focus on average welfare can lead them to prefer adding new suffering individuals to an already miserable world so long as the new individuals suffer less than the others, thereby improving the average welfare level.[94] However, people's focus on average levels of well-being instead of the total welfare level is reduced under more careful reflection (i.e., System-2 reasoning). Overall, the existing limited research suggests that people's population ethical intuitions depend on framing and often conflict.[94][108]

Normative Implications

Why should those interested in normative ethics care about the psychology of ethics? Moral philosophers often take their moral intuitions as evidence—as data or even as self-evident proof. But our moral intuitions may not be a perfect moral guide. We know from research on intuitions more generally that

they are systematically fallible.[25] For example, we know that people often make intuitive financial decisions that cost them money, decisions that people wouldn't endorse after more careful reflection. Why should we think that our self-interested decisions are systematically flawed but that our moral intuitions are not? Understanding moral intuitions and moral thinking more generally might help us see flaws in our thinking about utilitarianism and other ethical theories.

Of course, identifying flawed moral judgments is tricky, as there is little agreement on ground truth. But even without agreement on moral truth, psychological research can reveal inconsistencies in our moral thinking. And some ways of resolving those inconsistencies might be more plausible than others.

To take a classic example[109]: A ball and a bat together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost? Most people's immediate and intuitive thought is that the ball costs 10 cents (and that the bat costs $1). But a little reflection reveals that this can't be right (100 – 10 = 90). The correct answer is 5 cents. As it happens, people who do better on counter-intuitive reasoning problems such as this are also more likely to exhibit utilitarian patterns of moral judgment in sacrificial dilemmas.[63][110] It does not follow from this that utilitarianism is a superior moral theory. But it suggests a relationship between utilitarian thinking and the motivation to transcend the limits of intuitive thinking.

In line with this, Joshua Greene has argued that deontological responses in sacrificial dilemmas are grounded in intuitions and that deontological philosophy is best understood as an attempt to organize and justify—or, more provocatively, *rationalize*—those intuitive judgments.[54][111][112][113] By contrast, consequentialism and utilitarianism are, according to Greene, grounded in more reflective moral reasoning combined with a small set of very general evaluative premises (e.g., that happiness is good and that suffering is bad).

As discussed above, the strong emotional aversion to pushing the man off the footbridge is a product of model-free learning and decision-making. ("It just feels like the wrong thing to do"). This aversion is generally good because violence is almost always bad. It's bad for the victim, but it's likely also bad for the perpetrator, who may face retribution and reputational damage.[114] However, in sacrificial dilemmas, the usual relationship between action and consequence has been inverted. The action that is usually a terrible thing to do happens to be the very thing that will minimize harm. Appreciating this requires model-based reasoning. ("Better to do the thing that will minimize the loss of life"). And encouraging people to reason impartially using a version of Rawls' "Veil of Ignorance"[41] makes people more likely to favor the greater good in such dilemmas.[115][116][117] However, many people, including many philosophers, see our model-free emotional response as a justification for rejecting utilitarianism.[40][118][119] Knowing what we now know about how our minds work, does that make sense?

Making judgments based on emotional responses isn't inherently wrong; the question is whether our emotions are sensitive to the right things. Once again, a big reason that people react negatively to saving more lives in the footbridge case is because it involves "personal force," i.e., pushing someone rather than hitting a switch.[50][57] But almost no one believes this should matter morally. If a desperate friend called you from a footbridge looking for advice, you wouldn't ask them whether they would have to push someone with their hands or whether they could use a switch-operated trap door. Our emotions are responding to a morally incidental physical feature of the action. Calling this feature "morally incidental" is, of course, a value judgment and not a claim supported by scientific evidence. But it's a rather commonsensical value judgment and not one that requires a commitment to utilitarianism.

As this example suggests, scientific evidence about moral psychology, when combined with relatively uncontroversial moral assumptions, can point toward more interesting moral conclusions. Empirical research on moral psychology

doesn't *prove* that our anti-utilitarian intuitions lead us astray. But it may make it *harder* to resist that conclusion. Those who have rejected utilitarianism because of its counter-intuitive implications have, at the very least, some reasons to reconsider.

We've focused on trolley dilemmas and the scientific research they have inspired because this is the most developed case study. However, similar normative arguments can be made for other psychological obstacles to accepting (or potentially also rejecting) utilitarianism. Our resistance to utilitarianism's impartiality may be driven not by a good moral reason but by the amoral calculus of evolution, causing us to prioritize ourselves and our close associates over others.[120] This may also explain why we are reluctant to support the most effective charitable causes, which typically benefit physically, temporally, and socially distant individuals. Other psychological tendencies generally regarded as biases outside the moral domain, such as scope insensitivity, make utilitarianism harder to accept. Understanding these moral tendencies as part of a broader pattern of heuristics and biases may make it harder to give them normative weight.[37] Alternative framings of Nozick's Experience Machine argument suggest that his anti-hedonistic-utilitarian conclusion about the nature of well-being has been bolstered by status quo bias.[97] Prioritarian arguments against utilitarianism may be supported by an implicit misunderstanding of utility, confusing it with wealth.[54][121] Punishment as retribution—beyond what is needed to promote the greater good through deterrence, etc.—may be driven by vengeful instincts that can be counterproductive in the modern world.[49][122][123] And so on.

While utilitarianism's foundational principles sound very reasonable, they often seem to go wrong when applied to specific decisions (especially hypothetical ones). This could be because utilitarianism is deeply flawed, as many philosophers have concluded. But it could instead be because our moral thinking is flawed, relying on moral intuitions that are generally adaptive but nevertheless severely limited. A deeper understanding of moral psychology

won't, by itself, prove utilitarianism right or wrong. But it can help us assess utilitarianism in a more informed way.

Practical Implications

In this article, we explored how human moral psychology deviates from the utilitarian theory in the abstract, particularly in the context of hypothetical thought experiments and edge cases. A separate question is how these insights can help us become better utilitarians in the real world, given our psychological and situational constraints. For example, it's not psychologically feasible for most people to be fully impartial, prioritizing strangers over themselves and their loved ones whenever it does more good. A wise and psychologically informed utilitarian understands this and develops heuristics, rules, and virtues that sustainably approximate the utilitarian ideal. Similarly, while utilitarianism rejects deontological constraints in the context of stylized thought experiments free of uncertainty, such constraints provide essential guardrails in the real-world pursuit of better outcomes for humanity. These ideas are explored further in Virtues for Real-World Utilitarians.

About the Authors

Lucius Caviola is a Senior Research Fellow at the Global Priorities Institute, University of Oxford. He specializes in moral psychology and is the co-author of *Effective Altruism and the Human Mind* (Oxford University Press, 2024).

Joshua Greene is Professor of Psychology and a member of the Center for Brain Science faculty at Harvard University. He studies moral decision-making, intergroup conflict, and the nature of thought. He is the author of *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them.*

How to Cite This Page

Caviola, L. & Greene, J. (2024). Moral Psychology and Utilitarianism. In R.Y. Chappell, D. Meissner, and W. MacAskill (eds.), *An Introduction to Utilitarianism*, <https://www.utilitarianism.net/guest-essays/moral-psychology>, accessed 9/18/2024.

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

2. Everett, J. A. C. & Kahane, G. Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends Cogn. Sci.* **24**, 124–134 (2020). ↩

3. Coleman, M. & DeSteno, D. The intertemporal empathy gap: Feeling less distress for future others' suffering. (2024). ↩

4. Syropoulos, S., Law, K. F., Kraft-Todd, G. & Young, L. The Longtermism Beliefs Scale: Measuring Lay Beliefs for Protecting Humanity's Longterm Future. (2023) doi:10.31234/osf.io/e34kv. ↩

5. Caviola, L., Everett, J. A. C. & Faber, N. S. The moral standing of animals: Towards a psychology of speciesism. *J. Pers. Soc. Psychol.* **116**, 1011–1029 (2019). ↩

6. Caviola, L., Schubert, S., Kahane, G. & Faber, N. S. Humans first: Why people value animals less than humans. *Cognition* **225**, 105139 (2022). ↩

7. Hogg, M. A. Social Identity Theory. in *Understanding Peace and Conflict Through Social Identity Theory: Contemporary Global Perspectives* (eds. McKeown, S., Haji, R. & Ferguson, N.) 3–17 (Springer International Publishing, Cham, 2016). ↩

8. Haidt, J. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. (Knopf Doubleday Publishing Group, 2012). ↩ ↩

9. Taijfel, H. Experiments in intergroup discrimination. *Sci. Am.* **223**, 96–102 (1970). ↩

10. Dunham, Y., Baron, A. S. & Carey, S. Consequences of 'minimal' group affiliations in children. *Child Dev.* **82**, 793–811 (2011). ↩

11. Kinzler, K. D., Dupoux, E. & Spelke, E. S. The native language of social cognition. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12577–12580 (2007). ↩

12. de Waal, F. B. M., Leimgruber, K. & Greenberg, A. R. Giving is self-rewarding for monkeys. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13685–13689 (2008). ↩

13. de Waal, F. B. *Chimpanzee Politics: Power and Sex Among Apes*. (JHU Press, 2007). ↩

14. Donaldson, Z. R. & Young, L. J. Oxytocin, vasopressin, and the neurogenetics of sociality. *Science* **322**, 900–904 (2008). ↩

15. Keller, B. Opinion. *The New York Times* (2013). ↩

16. Henrich, J. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. (Farrar, Straus and Giroux, 2020). ↩

17. Wilks, M., Caviola, L., Kahane, G. & Bloom, P. Children Prioritize Humans Over Animals Less Than Adults Do. *Psychol. Sci.* **32**, 27–38 (2021). ↩

18. Ruby, M. B., Heine, S. J., Kamble, S., Cheng, T. K. & Waddar, M. Compassion and contamination. Cultural differences in vegetarianism. *Appetite* **71**, 340–348 (2013). ↩

19. Singer, P. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas About Living Ethically*. (Yale University Press, 2015). ↩

20. Ord, T. The moral imperative towards cost-effectiveness. https://www.academia.edu/download/30735939/moral_imperative.pdf (2012). ↩ ↩

21. MacAskill, W. *Doing Good Better: How Effective Altruism Can Help You Make a Difference*. (Penguin, 2015). ↩

22. Caviola, L., Schubert, S. & Greene, J. D. The Psychology of (In)Effective Altruism. *Trends Cogn. Sci.* **25**, 596–607 (2021). ↩ ↩

23. Schubert, S. & Caviola, L. *Effective Altruism and the Human Mind: The Clash Between Impact and Intuition*. (Oxford University Press, 2024). ↩

24. Caviola, L. *et al.* Donors vastly underestimate differences in charities' effectiveness. *Judgm. Decis. Mak.* **15**, 509–516 (2020). ↩

25. Kahneman, D. *Thinking, Fast and Slow*. (Macmillan, 2011). ↩ ↩

26. Bloom, P. Against empathy: The case for rational compassion. (2017). ↩

27. Desvousges, W. H. *et al. Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy*. (RTI Press, 2010). ↩

28. Small, D. A. & Loewenstein, G. Helping a Victim or Helping the Victim: Altruism and Identifiability. *J. Risk Uncertain.* **26**, 5–16 (2003). ↩

29. Slovic, P. *The Feeling of Risk: New Perspectives on Risk Perception*. (Routledge, 2010). ↩

30. Caviola, L. & Greene, J. D. Boosting effective giving with bundling and donor coordination. (2021) doi:10.31234/osf.io/65fmr. ↩

31. Berman, J. Z., Barasch, A., Levine, E. E. & Small, D. A. Impediments to Effective Altruism: The Role of Subjective Preferences in Charitable Giving. *Psychol. Sci.* **29**, 834–844 (2018). ↩

32. Pummer, T. Whether and where to give. *Philos. Public Aff.* **44**, 77–95 (2016). ↩

33. Caviola, L. *et al.* Utilitarianism for animals, Kantianism for people? Harming animals and humans for the greater good. *J. Exp. Psychol. Gen.* **150**, 1008–1039 (2021). ↩

34. Fiske, A. P. & Tetlock, P. E. Taboo trade-offs: reactions to transactions that transgress the spheres of justice. *Polit. Psychol.* (1997). ↩

35. Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C. & Lerner, J. S. The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *J. Pers. Soc. Psychol.* **78**, 853–870 (2000). ↩

36. Ubel, P. A., DeKay, M. L., Baron, J. & Asch, D. A. Cost-Effectiveness Analysis in a Setting of Budget Constraints — Is It Equitable? *N. Engl. J. Med.* **334**, 1174–1177 (1996). ↩

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

38. Anscombe, G. E. M. Modern Moral Philosophy1. *Philosophy* **33**, 1–19 (1958). ↩

39. Williams, B. *Problems of the Self: Philosophical Papers 1956-1972*. (Cambridge University Press, 1973). ↩

40. Sandel, M. J. Justice: What's the right thing to do. *BUL Rev.* (2011). ↩ ↩

41. Rawls, J. *A Theory of Justice*. (Harvard University Press, 1971). ↩ ↩

42. Mill, J. S. in *Utilitarianism* (Numerous editions, 1861). ↩

43. Hare, R. M. Ethical theory and utilitarianism. *Contemporary British Philosophy* (1982). ↩

44. Kant, I. *Groundwork for the Metaphysics of Morals*. (Oxford University Press, New York, 1785). ↩

45. Baron, J. & Spranca, M. Protected values. *Virology* **70**, 1–16 (1997). ↩

46. Tetlock, P. E. Thinking the unthinkable: sacred values and taboo cognitions. *Trends Cogn. Sci.* **7**, 320–324 (2003). ↩

47. Foot, P. The Problem of Abortion and the Doctrine of the Double Effect. *Oxf. Rev. Reprod. Biol.* **5**, 5–15 (1967). ↩ ↩

48. Thomson, J. J. Killing, letting die, and the trolley problem. *Monist* **59**, 204–217 (1976). ↩

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

50. Greene, J. D. *et al.* Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* **111**, 364–371 (2009). ↩ ↩ ↩

51. Awad, E., Dsouza, S., Shariff, A., Rahwan, I. & Bonnefon, J.-F. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences* **117**, 2332–2337 (2020). ↩

52. Aquinas, S. T. & Catholic Way Publishing. *The Summa Theologica: Complete Edition*. (Catholic Way Publishing, 2014). ↩

53. Quinn, W. S. Actions, intentions, and consequences: the doctrine of double effect. *Philos. Public Aff.* **18**, 334–351 (1989). ↩

54. Greene, J. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them.* (Penguin, 2013). ↩ ↩ ↩ ↩

55. Royzman, E. B. & Baron, J. The preference for indirect harm. *Soc. Justice Res.* **15**, 165–184 (2002). ↩

56. Cushman, F., Gray, K., Gaffey, A. & Mendes, W. B. Simulating murder: the aversion to harmful action. *Emotion* **12**, 2–7 (2012). ↩

57. Bago, B. *et al.* Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nat Hum Behav* **6**, 880–895 (2022). ↩ ↩

58. Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. & Cohen, J. D. The neural bases of cognitive conflict and control in moral judgment. *Neuron* **44**, 389–400 (2004). ↩ ↩

59. Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E. & Cohen, J. D. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* **107**, 1144–1154 (2008). ↩

60. Shenhav, A. & Greene, J. D. Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *J. Neurosci.* **34**, 4741–4749 (2014). ↩ ↩

61. Cushman, F. Action, outcome, and value: a dual-system framework for morality. *Pers. Soc. Psychol. Rev.* **17**, 273–292 (2013). ↩ ↩

62. Crockett, M. J. Models of morality. *Trends Cogn. Sci.* **17**, 363–366 (2013). ↩ ↩

63. Patil, I. *et al.* Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *J. Pers. Soc. Psychol.* **120**, 443–460 (2021). ↩ ↩ ↩

64. van Honk, J. *et al.* Breakdown of utilitarian moral judgement after basolateral amygdala damage. *Proc. Natl. Acad. Sci. U. S. A.* **119**,

e2119072119 (2022). ↩ ↩

65. Glenn, A. L., Raine, A. & Schug, R. A. The neural correlates of moral decision-making in psychopathy. *Mol. Psychiatry* **14**, 5–6 (2009). ↩

66. Cushman, F. & Greene, J. D. Finding faults: how moral dilemmas illuminate cognitive structure. *Soc. Neurosci.* **7**, 269–279 (2012). ↩

67. Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J. & Rangel, A. Emotional and Utilitarian Appraisals of Moral Dilemmas Are Encoded in Separate Areas and Integrated in Ventromedial Prefrontal Cortex. *J. Neurosci.* **35**, 12593–12605 (2015). ↩

68. Bechara, A., Damasio, A. R., Damasio, H. & Anderson, S. W. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**, 7–15 (1994). ↩

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

70. Ciaramelli, E., Muccioli, M., Làdavas, E. & di Pellegrino, G. Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Soc. Cogn. Affect. Neurosci.* **2**, 84–92 (2007). ↩

71. Moretto, G., Làdavas, E., Mattioli, F. & di Pellegrino, G. A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *J. Cogn. Neurosci.* **22**, 1888–1899 (2010). ↩

72. Mendez, M. F., Anderson, E. & Shapira, J. S. An investigation of moral judgement in frontotemporal dementia. *Cogn. Behav. Neurol.* **18**, 193–197 (2005). ↩

73. Martins, A. T., Faísca, L. M., Esteves, F., Muresan, A. & Reis, A. Atypical moral judgment following traumatic brain injury. *Judgm. Decis. Mak.* **7**, 478–487 (2012). ↩

74. Koenigs, M., Kruepke, M., Zeier, J. & Newman, J. P. Utilitarian moral judgment in psychopathy. *Soc. Cogn. Affect. Neurosci.* **7**, 708–714 (2012). ↩

75. Koven, N. S. Specificity of meta-emotion effects on moral decision-making. *Emotion* **11**, 1255–1261 (2011). ↩

76. Patil, I. & Silani, G. Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Front. Psychol.* **5**, 501 (2014). ↩

77. Darby, R. R., Horn, A., Cushman, F. & Fox, M. D. Lesion network localization of criminal behavior. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 601–606 (2018). ↩

78. McCormick, C., Rosenthal, C. R., Miller, T. D. & Maguire, E. A. Hippocampal Damage Increases Deontological Responses during Moral Decision Making. *J. Neurosci.* **36**, 12157–12167 (2016). ↩

79. Verfaellie, M., Hunsberger, R. & Keane, M. M. Episodic processes in moral decisions: Evidence from medial temporal lobe amnesia. *Hippocampus* **31**, 569–579 (2021). ↩

80. Craver, C. F. *et al.* Moral judgment in episodic amnesia. *Hippocampus* **26**, 975–979 (2016). ↩

81. Kahneman, D. Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475 (2003). ↩

82. Stanovich, K. E. & West, R. F. Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* **23**, 645–65; discussion 665–726 (2000). ↩

83. Conway, P. & Gawronski, B. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *J. Pers. Soc. Psychol.* **104**, 216–235 (2013). ↩

84. Conway, P., Goldstein-Greenwood, J., Polacek, D. & Greene, J. D. Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition* **179**, 241–265 (2018). ↩

85. Costa, A., Foucart, A., Arnon, I., Aparici, M. & Apesteguia, J. 'Piensa' twice: on the foreign language effect in decision making. *Cognition* **130**, 236–254 (2014). ↩

86. Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D. & Keysar, B. Thinking More or Feeling Less? Explaining the Foreign-Language Effect on Moral Judgment. *Psychol. Sci.* **28**, 1387–1397 (2017). ↩

87. Capraro, V. The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *J. Pers. Soc. Psychol.* **126**, 719–757 (2024). ↩

88. Klenk, M. The influence of situational factors in sacrificial dilemmas on utilitarian moral judgments: A systematic review and meta-analysis. *Rev. Philos. Psychol.* **13**, 593–625 (2022). ↩

89. Baron, J. *Thinking and Deciding*. (Cambridge University Press, 2007). ↩

90. Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S. & Sinnott-Armstrong, W. Consequences, action, and intention as factors in moral judgments: an FMRI investigation. *J. Cogn. Neurosci.* **18**, 803–817 (2006). ↩

91. Quinn, W. S. Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. *Philos. Rev.* **98**, 287–312 (1989). ↩

92. Spranca, M., Minsk, E. & Baron, J. Omission and commission in judgment and choice. *J. Exp. Soc. Psychol.* **27**, 76–105 (1991). ↩

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

94. Caviola, L., Althaus, D., Mogensen, A. L. & Goodwin, G. P. Population ethical intuitions. *Cognition* **218**, 104941 (2022). ↩ ↩ ↩ ↩ ↩ ↩

95. Greenberg, S. Which intrinsic values set different demographic groups apart? Check out our study results. *Clearer Thinking* https://www.clearerthinking.org/post/2018/08/09/which-intrinsic-values-set-different-demographic-groups-apart-check-out-our-study-results (2018). ↩

96. Nozick, R. *Anarchy, State, and Utopia*. (John Wiley & Sons, 1974). ↩

97. De Brigard, F. If You Like It, Does It Matter If It's Real? *Philos. Psychol.* **23**, 43–57 (2010). ↩ ↩

98. Weijers, D. Nozick's experience machine is dead, long live the experience machine! *Philos. Psychol.* **27**, 513–535 (2014). ↩

99. Fehr, E., Naef, M. & Schmidt, K. M. Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment. *Am. Econ. Rev.* **96**, 1912–1917 (2006). ↩

100. Charness, G. & Rabin, M. Understanding Social Preferences with Simple Tests*. *Q. J. Econ.* **117**, 817–869 (2002). ↩

101. Frohlich, N., Oppenheimer, J. A. & Eavey, C. L. Laboratory Results on Rawls's Distributive Justice. *Br. J. Polit. Sci.* **17**, 1–21 (1987). ↩

102. Frohlich, N., Oppenheimer, J. A. & Eavey, C. L. Choices of Principles of Distributive Justice in Experimental Groups. *Am. J. Pol. Sci.* **31**, 606–636 (1987). ↩

103. Paolacci, G. & Yalcin, G. Fewer but poorer: Benevolent partiality in prosocial preferences. *Judgm. Decis. Mak.* **15**, 173–181 (2020). ↩

104. Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *J. Pers. Soc. Psychol.* **83**, 284–299 (2002). ↩

105. Mogensen, A. L. Against Large Number Scepticism. in *Ethics and Existence: The Legacy of Derek Parfit* (eds. McMahan, J., Campbell, T., Campbell, T., Goodrich, J. & Ramakrishnan, K.) 311–330 O(Oxford University Press, 2022). ↩

106. Parfit, D. *Reasons and Persons*. (OUP Oxford, 1984). ↩

107. Greaves, H. Population axiology. *Philos. Compass* **12**, e12442 (2017). ↩

108. Spears, D. Making people happy or making happy people? Questionnaire-experimental studies of population ethics and policy. *Soc. Choice Welfare* **49**, 145–169 (2017). ↩ ↩

109. Frederick, S. Cognitive Reflection and Decision Making. *J. Econ. Perspect.* **19**, 25–42 (2005). ↩

110. Byrd, N. & Conway, P. Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition* **192**, 103995 (2019). ↩

111. Greene, J. D. The secret joke of Kant's soul. *Moral psychology* **3**, 35–79 (2008). ↩

112. Greene, J. D. Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics* (2014). ↩

113. Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* **167**, 66–77 (2017). ↩

114. Everett, J. A. C., Pizarro, D. A. & Crockett, M. J. Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* **145**, 772–787 (2016). ↩

115. Huang, K., Greene, J. D. & Bazerman, M. Veil-of-ignorance reasoning favors the greater good. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23989–23995 (2019). ↩

116. Huang, K., Bernhard, R. M., Barak-Corren, N., Bazerman, M. H. & Greene, J. D. Veil-of-ignorance reasoning mitigates self-serving bias in resource allocation during the COVID-19 crisis. *Judgm. Decis. Mak.* **16**, 1–19 (2021). ↩

117. Greene, J. D., Huang, K. & Bazerman, M. Redirecting Rawlsian Reasoning Toward the Greater Good. *The Oxford Handbook of Moral Psychology* 246 (2022). ↩

118. Thomson, J. J. The Trolley Problem. *The Yale Law Journal* vol. 94 1395 Preprint at https://doi.org/10.2307/796133 (1985). ↩

119. Kamm, F. M. *Morality, Mortality: Death and Whom to Save from It.* (Oxford University Press, 1993). ↩

120. Singer, P. *The Expanding Circle: Ethics, Evolution, and Moral Progress.* (Princeton University Press, 2011). ↩

121. Greene, J. & Baron, J. Intuitions about declining marginal utility. *J. Behav. Decis. Mak.* **14**, 243–255 (2001). ↩

122. Wiegman, I. The evolution of retribution: Intuitions undermined. *Pac. Philos. Q.* **98**, 193–218 (2017). ↩

123. Sapolsky, R. M. *Determined: Life Without Free Will.* (Random House, 2023). ↩