Are moral judgments rational?

Jonathan Baron*

October 24, 2021

Contents

*Draft chapter for *Cambridge Handbook of Moral Psychology, 2nd edition*, edited by Philip Robbins and Bertram Malle.

1 Introduction

As I write this in the fall of 2021, many hospitals in the U.S. are overwhelmed with COVID-19 patients, most of whom have refused to be vaccinated against the disease. Many of these non-vaccinators appeal to moral principles concerning freedom and rights, which they take to outweigh the consequences of their decision. They claim the right to make decisions about their own bodies, and the right to freedom from government control over personal behavior. Some politicians support these views even to the point of trying to prohibit schools and private businesses from imposing mandates for mask wearing or vaccination. Note that the expected consequences of non-vaccination are bad for everyone. Vaccination reduces the probability of serious illness for the individual, and it reduced the probability of an infected person, even one without symptoms, transmitting the disease to others. If the effect on others is what we consider to be a moral issue, then non-vaccination is both individually and morally irrational, from the perspective of its consequences. Yet the principles at issue are moral principles.

This case is an example of a frequent conflict between moral principles that people advocate and try to follow, on the one hand, and the expected consequences of following those principles, on the other. The moral principles at issue are inconsistent with moral principles based on utilitarianism, which holds that choice options should be evaluated in terms of their expected consequences for all those affected, but this is not all that makes them irrational. The choice of non-vaccination for oneself conflicts with expected-utility theory (discussed below) as applied to individual choices; it is a losing gamble. And opposing vaccinations for others is simply harmful to them, which is inconsistent with any concept of morality.

Apparent examples of this sort of inconsistency in the real world have been extensively documented. In many cases, the analysis of expected consequences is based on economics rather than utilitarian analysis, but the conclusions of economic analysis is generally consistent with those that utilitarian analysis would imply.[1] Apparent inconsistencies have been found in allocation of resources to large humanitarian tragedies (Slovic et al., 2021); in insurance decision by firms and individuals (Johnson, Hershey, Meszaros, & Kunreuther, 1993); in excessive attention to some risks coupled with neglect of others (Breyer, 1993; Kunreuther & Slovic, 1978; Sunstein, 2002; in tax policy (McCaffery, 1997); in economic policies concerning trade, price controls, and wages (Caplan, 2007); and elsewhere.

All these realistic cases (and many more) support the argument that people's moral judgments, when put into practice, lead to consequences that peo-

[1]Traditional economics is concerned with wealth maximization rather than utility maximization. If the winners from some policy change could compensate the losers with enough money so that everyone would rationally agree on the change, the change is recommended, even if the compensation is not paid. Modern "welfare economics" is more consistent with utilitarianism. Both views usually assumes that the utility of money is marginally declining, so that simple redistribution from rich to poor, even at some cost, can be justified.

ple themselves would consider worse on the whole than what might have been achieved. But the real world is complicated. It is possible that the principles can have a utilitarian defense after all. For example, many of these apparently self-defeating policies arise through the functioning of institutions, such as legislatures and courts, that are imperfect yet better than any feasible alternatives, so that any attempt to overturn their results would, in the long run, make matters worse as a result of weakening these institutions. It thus becomes reasonable to ask whether people *really* apply non-utilitarian principles when they make moral judgments. One way to answer this question is to do psychology experiments, and those are the main topic of this chapter. At issue is the question of whether we can demonstrate truly irrational and non-utilitarian reasoning in hypothetical or real judgments under controlled conditions.

1.1 Normative and descriptive "models" in experimental psychology

Since the 19th century, psychologists have studied reasoning in contexts in which right answers are defined by some formal theory such as the logic of syllogisms. A common finding is that reasoning did not conform well to the model, thus, Henle (1962) begins by pointing out that "The question of whether logic is descriptive of the thinking process, or whether its relation to the thinking process is normative only, seems to be easily answered. Our reasoning does not, for example, ordinarily follow the syllogistic form, and we do fall into contradictions." She goes on to muddy the waters. Around the same time (the 1950s and 60s), others were comparing human judgments to other normative models, including probability and statistics (Meehl, 1954; Bruner, Goodnow, & Austin, 1956; Chapman & Chapman, 1969). In retrospect, we can think of such research as comparing "descriptive models" — psychological accounts of what people are doing — to normative models. Perhaps the term "descriptive model" is excessively ambitious, since even now there are few such models that account for more than a few results each, but the term is used.

Kahneman and Tversky (1979; Tversky, 1967; Tversky & Kahneman, 1981) began to apply this approach to decisions as well as judgments (and their 1979 paper proposed a true descriptive model that accounted fairly well for choices among simple gambles). Their normative model was expected-utility theory in the form proposed by Savage (1954), in which both probability and utility were subjective (even if numerical probabilities were included in problem statements). Given this normative model, researchers could not always determine whether a given decision conformed to the model or not. For example, one person might prefer $10 for sure over a gamble with a .6 probability of $25 and a .4 probability of $0. Another person might prefer the gamble. The former person's utility for $25 might be less than twice as high as her utility for $10, and she might think of .6 as "essentially an even chance", so that her subjective probability of winning would be closer to .5. Thus, for her the expected utility (subjective probability times subjective utility) of the gamble would be less than that of $10 for sure.

To overcome this problem and show that choices were inconsistent with the

normative model, Tversky and Kahneman (1981) emphasized the use of framing effects, in which the same choice was offered in different words. If subjects made different choices in the two versions, then they could not be following the normative model. A classic example was the Asian Disease Problem, in which some subjects were told that an Asian Disease was approaching and 600 deaths would be expected if nothing was done. In one version, the subjects chose between "200 saved" and a .33 chance to save 600. In another version the choice was between "400 die" and a .67 probability that 600 would die. Because the subjective utility function is different for gains and losses, most subjects in the gain condition chose "200 saved", and most subjects in the loss condition chose the gamble.

This experiment had two properties that have received little attention in the extensive literature about it. One is that it is essentially a moral problem, not an individual choice like the money gambles used in other studies. It is moral because it is a decision about other people. Research on decisions had slipped from a focus on expected-utility to a focus on utilitarianism. Utilitarianism is the natural extension of expected-utility to decisions for many people.[2] The utilitarian normative model here is to base the decision on the expected number of deaths (usually assuming that the subjective probabilities match the given probabilities).

The second property of the Asian Disease Problem concerns strong preferences for the two options. The expected utilities of the two options are close. Thus, strong preferences for different options violate a feature of utilitarianism (and other moral theories), which is to treat all lives equally. In gains, for example, a strong preference for "save 200" implies that the extra 400, beyond the 200 saved, are given less weight than twice that of the first 200 lives. Slovic (e.g., 2007) has explored this finding of unequal treatment extensively. One way to think of this phenomenon is in terms of the curve relating total disutility to number of deaths. People tend to make decisions as if the slope of this curve decreases: the millionth death matters less than the tenth, or the first.

Here is another example of the move from individual to moral decisions. The pertussis vaccine used to prevent whooping cough in the 1980s would often cause a disease very much like the one it prevented, but at a much lower probability. Despite the clear benefits, many people resisted (and still resist) vaccination (Asch et al., 1994; Sherman et al., 2021). Ritov and Baron (1990) found, in a laboratory study, that many people would not want such a vaccine, because (presumably) they would not want to cause the disease through their action. Ritov and Baron also found that people would also oppose requiring the vaccine as a public health measure. The individual decision was purely a matter of self-interest, but the public-health decision was moral, because it concerned other people.

[2]Utilitarianism requires that we add up expected utilities across people. In situations such as those at issue, where the people are anonymous and drawn randomly from the same population, they can all be treated as if they have the same expectation. Other situations require interpersonal comparisons, trading off the gains for some people against the losses for others, using what we know about the different individuals.

Note that, in this case, the self-interested decision is irrational (from the perspective of expected-utility) because omission of the vaccine increases personal risk. Could we say that the moral omission is also irrational because it means that more people will be sick? Some moral systems have a rule against using people as means to help others, and it could be argued that those who suffer from the side effects will serve as means to prevent disease in a greater number. Yet it seems inconsistent to say that the decision that is rational for each individual is immoral when applied to the population.

In these examples, the general approach of comparing laboratory decisions to normative models can be, and has been, extended from individual decisions to moral decisions, often with the implicit use of utilitarianism as a normative model. Further research, some of which I review here, finds that the departures from utilitarianism are systematic. As noted, some of these departures result from distortions in the way people think about quantities. Many others result from the application of non-utilitarian principles to the problems of interest.

These principles may be absolute or "prima facie", that is, considerations that can be overridden by other considerations (Ross, 1930). Examples are "We have a right to control our bodies", "Do no harm" (meaning do no harm through action, as opposed to omission), "Do not use people as means to achieve better outcomes for others", or "Do not kill innocent people." They are often called "moral intuitions" (Hare, 1981) or "moral heuristics" (Sunstein, 2005). "Heuristics" originally referred to weak methods that might be helpful in solving problems, such as "Do you know a related problem?" (Polya, 1945), but the term was used by Tversky and Kahneman to refer to judgment tasks; an example is judging the probability that someone is a member of a group by the similarity of that person to prototypical members of the group, thus ignoring other relevant attributes such as the size of the group (Tversky & Kahneman, 1974).

When this view is extended to moral judgments, other problems arise. In principle, a heuristic is a "fast and frugal" method that often works but sometimes does not.[3] In morality, though, some of these heuristics seem to become hardened into rules that people knowingly apply, believing that they constitute the best possible moral judgments. Theologians and philosophers defend these rules as normative in this way (e.g., the rule about not using people as means). Such rules are often part of deontology, a class of moral systems based on rules, rights and duties, which go beyond simply bringing about the best expected consequences for all.[4] Thus, much of the research on moral judgment focuses not so much on heuristics but on the contrast between deontological rules and utilitarianism. In this research literature, the terms "deontological" and "utilitarian" are not meant to imply that choices are based on representations of either system, just that they are consistent with what those choices would be.[5]

[3]The term "heuristic" is also used simple algorithms that are more accurate than the more complex algorithms they replace (Gigerenzer et al., 1999).

[4]Deontological systems usually include a role for consequences, but as one criterion among many.

[5]The term "deontology" is variously defined, and many of the intuitions are deontological only in the broadest sense that they refer to properties of an action (or inaction) other than

Henceforth I will use the term "intuition", which I think captures the idea that the relevant moral principles tend to be evoked immediately upon presentation of a moral problem, without any explicit attempt to search memory for relevant considerations.

All normative models are controversial to some degree, including Bayesian probability theory and expected-utility theory (e.g., Ellsberg, 1961), but utilitarianism seems more controversial than most of the others that are studied psychologically, in part because it yields conclusions that seem to conflict with strong moral intuitions held by philosophers and psychology researchers as well as by experimental subjects. Hare (1981) has dealt with this conflict explicitly and in depth. His approach turns out to be surprising relevant to experimental psychology (as I discuss later).

But there are other reasons for looking for biases relative to a utilitarian normative model, even for those who do not accept utilitarianism as truly normative. Specifically, if people consistently violate the utilitarian standard in the same biased way (as in favoring harmful omissions over less harmful acts), we should not be surprised if the real consequences turn out to be worse than if the utilitarian standard were followed. As I suggested, many examples in the real world can be explained in terms of such biases. Thus, the study of violations of utilitarianism can at least help us understand why things in the real world are not as good as they could be. If the violation of utilitarian standards is the result of truly normative moral rules, then we at least learn the potential cost of adhering to those rules.

Of course, much more could be said in defense of utilitarianism (e.g., Hare, 1981, whose other work is nicely summarized by Singer, 2002), but this is not the place for it.

2 Methods and biases

In this section I will discuss several experimental methods and possible biases, organized by method rather than substantive topic, although I comment on the normative approach to some of the topics. All of these methods are potentially capable of showing that judgments or hypothetical decisions are non-utilitarian.

It is worth noting that essentiall all of the non-utilitarian biases I describe here are the result of processes also found in non-moral situations. Cushman and Young (2011) and Greene (2007) have argued explicitly for the parallelism between "cognitive biases" and patterns found in moral judgment.

its consequences. By "consequences", I refer to states of the world to which individuals assign some value. The values must depend on the states alone and not on whatever actions or natural causes brought them about (although actions in themselves can be evaluated as states); values are thus not opinions about what should be done.

2.1 Framing effects

A framing effect, as noted already, is found when two equivalent cases yield different responses. An example from moral psychology is the Schelling effect on judgments about fair tax rates (McCaffery & Baron, 2006a). In the original classroom demonstration by Schelling, he had asked his students to evaluate a tax policy that would allow a larger child deduction to the rich than to the poor. The students objected to this as being unfair. Schelling then pointed out that, if the reference point were a couple with children instead of a couple without children, penalties would be needed for childless households in order to produce the same outcome. The parallel question was then whether the surcharge for poor childless households should be as large as it is for rich ones. Students reversed their preference with the altered presentation. They objected to an equally large surcharge on lower-income households. Note that this method depends only on the potential realization, with reflection, that the two questions are equivalent, but the two answers are not. Nothing here depends on utilitarianism as such.

In another example of a framing effect, Harris and Joyce (1980) told subjects that a group of partners had opened a business (e.g., selling plants at a flea market). The partners took turns operating the business, so different amounts of income were generated while each partner was in control, and different costs were incurred. Subjects favored equal division of profits when they were asked about division of profits, and they favored equal division of expenses when asked about expenses. Because expenses and profits were unequal in different ways, their two judgments conflicted. This result depends on an principle of equality, and the inconsistency does not depend on utilitarianism.

A more complex framing effect concerns the effect of marriage (McCaffery & Baron, 2006a). When asked directly, many subjects favor "marriage neutrality", which means that marriage does not affect the total taxes paid. People also favor progressive taxation, which means that those with higher incomes pay a higher percentage in taxes. Finally, people tend to favor "couples neutrality", which means that couples with the same income pay the same tax, regardless of which earner makes more. Careful reflection (left as an exercise for the reader) implies that these three principles are incompatible. One of them must give.[6] This is, like the Schelling effect, a logical inconsistency, hence a form of framing effect, which involves focusing on the question that is asked, an "isolation effect" (discussed below).

2.2 Contrast of utilitarian and non-utilitarian options

Other methods involve asking subjects to decide between two options, one of which is consistent with utilitarianism and the other of which is not. The non-utilitarian option deviates by exemplifying a particular bias.

[6]Hint: Compare the case where one partner earns $200,000 and the other earns $0 to the case where each earns $100,000. Before marriage, the total tax is higher for the first couple.

2.2.1 Omissions

A great deal of research has concerned action/omission dilemmas such as the vaccine case described above, in which people are more willing to accept the harms caused by omission than the harms caused by action. Although Ritov and Baron (1990) coined the term "omission bias" as a name for this bias, that term was misleading. A simple bias toward omission would be a bias toward the default, whatever it is. Although a default bias does exist, it plays a minor role in the bias at issue (Baron & Ritov, 1994). Another determinant is the amplification effect, in which the consequences of action are simply given more weight than those of omission. If both options involve gains rather than losses, the amplification effect induces a bias toward action, which can be large enough to overcome the default bias.

Recent studies have tended to concern a set of dilemmas originally designed by philosophers as extreme cases on which to test, and try to explain, their moral intuitions (e.g., Foot, 1967/1978). In the simple trolley case, a runaway trolley is headed toward five people and will kill them if nothing is done. You can divert the trolley onto another track where it will kill only one person. Most people think diversion is the best response. In the "footbridge" version, the only way to stop the trolley is to push a large man off of a footbridge, so that he falls on the track and blocks the trolley, being killed in the process. Most people resist this solution, and many experiments have tried to examine and explain this sort of difference.

A potential issue for experiments like these is what question to ask. In many experiments, the research ask, "Is it acceptable to push the man ...?" The problem with this is that "acceptable" applies only to a single option, and utilitarians (and others concerned with decisions) must ask the question "compared to what?" The relevant question for us is which option is better, morally. Deontology often makes distinctions between what is permitted, forbidden, or morally required. Because these categories apply to options, not choices, it is possible for both options in a choice to be acceptable, or both forbidden. Other alternatives that work for everyone are: "Which option should [the agent] choose?" and "Which option should you choose [if you were the agent]?" Some studies have asked replaced "should" with "would". This may be interesting, but some people say, explaining themselves, "I know that I should do it, but I could not bring myself to actually do it" (Baron, 1992).

2.2.2 The nature of "omission bias"

The literature has identified two major determinants of "omission bias": deontological rules, and the use of a limited concept of causality.[7]

[7]Spranca, Minsk & Baron (2001), Experiments 5 and 6, examined other possible determinants and found some evidence for some of them, and other literature has found still others, especially a preference for harm caused by "nature" over harm caused by people (e.g., Kahneman & Ritov, 1994). The possible causes of the basic result are not mutually exclusive and are often confounded.

Rules favoring omission are more common than those favoring action (Baron & Ritov, 2009). Rules that prescribe acts are usually conditional on some role. A physician, once accepting a patient, is morally and legally obliged to try to save the patient's life (unless instructed otherwise) but a rule requiring anyone to try to save every life at risk is impossible to take seriously. Likewise, a rule against performing abortions is easier to follow than a rule requiring prevention of miscarriages in similar situations.

"Utilitarian moral dilemmas" often involve rules of this sort, such as prohibitions against killing, or tampering with human genes that affect future generations. When these rules are understood to be absolute (as discussed in the next section), we would expect that subjects would object to action regardless of how beneficial its consequences are. These results are found (Baron & Ritov, 2009). Thus, one determinant of the usual bias favoring omissions over less harmful acts, is the result of specific rules that are understood to be absolute (or nearly so).

Another determinant concerns causality (Cushman, 2008). We can (loosely) classify judgments of causality into two categories. One category, which includes "but for" causality, may be called "make a difference". You are (perhaps partially) causally responsible for some outcome if something under your control could have made a difference in whether the outcome occurred or not. This view does not distinguish acts and omissions as such. It is often applied to tort law, especially lawsuits against someone who is supposed to take care to avoid harming others. Utilitarianism implies make-a-difference causality, at least when the options are clearly laid out and both possible.[8]

The other category might be called direct causality. By this view, you are causally responsible for some outcome if there is a chain of events between your action and the outcome, with each link in the chain following some known principle of causality, such as the laws of physics (but any science will do, and also beliefs about supernatural causes). By this view, people may sometimes be held morally responsible for outcomes that they could not have avoided. (Spranca et al., 1991, report a few instances of this.). Young children tend to consider outcomes only, thus judging that an act is wrong if it causes harm by accident (Piaget, 1932). The apparent bias toward harmful omissions over less harmful acts seems to be closely related to direct causality.

Supporting a role for perceived direct causality, Baron and Ritov (1994, Experiment 4) compared the original vaccination case (in which vaccination deaths were from side effects) with a "vaccine failure" case, in which the deaths that result if the vaccination is chosen are caused by its failure to prevent the natural disease. The bias against vaccination (action) was much stronger in the original condition than in the vaccine-failure condition.

[8]Utilitarianism is often criticized for implying infinite obligations that cannot possibly be met. One answer (among many) is that utilitarianism need not imply that we are bad people if we fail to maximize utility, as this is inevitable; it is a doctrine of better vs. worse, not best vs. failure. Second, as a part of decision theory, utilitarianism applies to options that are "on the table" in any particular real decision. Some options are not considered, perhaps because of prior commitments to others, or because self-interest is just too strong.

Royzman and Baron (2002) compared cases in which an action caused direct harm with those in which an action caused harm as a side effect (i.e., "caused" only in the make-a-difference sense). For example, in one case, a runaway missile is heading for a large commercial airliner. A military commander can prevent collision with the airliner either by interposing a small plane between the missile and the large plane or by asking the large plane to turn, in which case the missile would hit a small plane now behind the large one. The indirect case (the latter) was preferred. In Study 3, subjects compared indirect action, direct action, and omission (i.e., doing nothing to prevent the missile from striking the airliner). Subjects strongly preferred omission to direct action, but not much to indirect action. Baron and Ritov (2009, Study 3) found similar results, and also that judged causality of the action was the main determinant of bias against action.

Greene et al. (2009) found that direct causality is a matter of degree. The most resistance to action occurred when a physical effect of action (hands-on pushing a man) caused a death, compared to cases in which the causal link between action and outcome involved more steps.

In sum, it seems that the bias against beneficial action is the result of at least two factors other than default bias: the perception of direct causality, as opposed to make-a-difference casuality; and the commitment to particular rules that prohibit certain actions.

All of these studies, it should be noted, are consistent with sometimes extreme individual differences, with some subjects making the utilitarian response almost all the time. These subjects apparently do follow make-a-difference causality. In some experiments, we have found subjects who equate inaction with standing by in the face of evil, as with those German citizens who tolerated Hitler (e.g., Spranca et al., 1991).

Note that some of these studies also ask about "blame" or "responsibility". The latter term is ambiguous between causal, moral, and legal meanings. The former may refer to punishment, which is examined more directly (and less ambiguously) in other experiments (below).

2.2.3 Protected values

Some deontological rules are taken to be absolute (Baron & Ritov, 2009). Tetlock (e.g., 2003), has used the term "sacred values" for essentially the same phenomenon, and Roth (2007) has used the term "repugnant transactions" for moral prohibitions on transactions such as a live donor selling a kidney. These protected values (PVs) are thus "protected" from trade-offs with other considerations. Some PVs are based on religion, but many are held by atheists, such as rules against cloning or genetic engineering of humans. In such cases, people say they should not violate the rule (usually a prohibition) no matter how great the benefits are. However, when asked to try hard to think of cases in which the benefits would be great enough, or when given some possible counterexamples, most people admit that the rules are not in fact absolute, so they seem to be absolute only as a result of insufficient reflection (Baron & Leshner, 2000; Tetlock, Mellers & Scoblic, 2017).

PVs may function as heuristics that serve the purpose of avoiding further thought about whether some trade-off is warranted (Hanselmann & Tanner, 2008). Thus, they appear to be non-utilitarian. However, J. S. Mill (in "On liberty", 1859) argued that we should follow certain moral rules even if it seems clear that the consequences of breaking them in some situation would be better than those of following the rule. Suppression of free speech was an example. The idea here is that our own judgments about expected consequences in such cases are not trustworthy; we are subject to self-serving biases and ordinary error. We do not need to deceive ourselves in order to follow such rules. When asked to join a terrorist cell, a person today might think to himself, "It seems that the cause is just, and that the total harm of the deaths that we would cause would be much smaller than the harm we would prevent by carrying through the plan. But I know that almost all the terrorists throughout history have drawn just this conclusion, and the vast majority of them have been incorrect. Thus, it is probably best if I don't join." Note that everything is conscious here. No self-deception is needed.

This, in experiments on PVs, it is worth giving subjects, as one of the response options, something like the (non-exclusive) options used by Baron and Leshner (2000, Experiment 2). Among the options for possible PVs such as "Cutting all the trees in an old-growth forest" were the following:

1. I cannot imagine any situations in which this is acceptable. (38)

2. I can imagine situations in which the benefits are great enough to justify this, but these situations do not happen in the real world. (7)

3. There are situations in the real world in which the benefits are great enough, but people can not recognize these situations, so it is best never to do this. (9)

4. This is unacceptable as a general rule, but we should make exceptions to it if we are sure enough. (28)

The percents of choices are shown in parentheses, so it seems that apparent PVs are not usually the result of a Mill-type explanation and are truly non-utilitarian principles.

2.2.4 Parochialism and self-interest

From a utilitarian perspective — as well as many other perspectives — a major bias in people's reasoning is parochialism (Baron, 2012a,b; also called "in-group bias"). The term refers to a class of experimental social-dilemma games (Bornstein & Ben-Yossef, 1994). In a social dilemma, each player can help other players in the group at some cost to himself, and the total benefit to the group is greater than the cost. This is called "cooperation". Examples in the real

world are widespread, from doing one's job without shirking, to paying taxes honestly (when it is easy to cheat), to contributing to charities. Parochialism arises when each player's behavior can affect an in-group and an out-group, and some players are willing to help the in-group at some personal cost while hurting the out-group even more (perhaps as a result of ignoring the outgroup).

Consider voting as an example. "Cooperation" means voting for the candidate or proposition that is best for those who are relevant to your vote, which could be you and your family, your compatriots, or everyone in the world. Defection in this example is not voting. Voting has a cost. It is well known, but not well understood, that the probability of being the pivotal (decisive) voter is so low that, even if you gain a large amount of money from your side winning, the expected return of voting is, like that of a lottery ticket, not worth the cost.

However, if you care enough about other people, taking their utilities as part of your own, with some weight for each other person, then voting can be worth the cost (Edlin, Gelman & Kaplan, 2007). Given this mathematical fact, a situation could arise in which it is not worth voting if all you care about is yourself, not quite worth voting if you care about your nation, but well worth voting if you care about humanity. If you are rational, you would then vote for candidates or proposals that are best for humanity. Otherwise, voting is not worth the cost.

"Cosmopolitanism" is sometimes used as a technical term used for the attitude of caring about the world. Although this attitude sounds as idealistic and fanciful as the John Lennon song "Imagine", in fact it is fairly common in the modern world (Buchan et al., 2009, 2011). Arguably, it could arise as a result of reflection (Singer, 1982). What principle can justify caring about some people but ignoring others? Answers could arise, but when we reflect on them (without myside bias) they may seem weak. Surely this sort of reasoning was part of what has led people to oppose slavery and to promote women's political and legal rights. The absence of it allows parochialism to exist.

Other sorts of reasoning lead to parochialism (Baron, 2012a,b). People think they have a duty to support their nation because their nation has given them the vote, or in return for what their nation has done for them. (Of course, most nations do not tell their citizens, even naturalized citizens, that this is expected, and it is well known that some voters, especially in a nation of immigrants, are concerned with particular foreign countries to which they are tied in some way.)

2.3 Attending to irrelevant attributes or ignoring relevant ones

Kahneman and Frederick (2002) proposed that many biases can be explained in terms of "attribute substitution." Two options differ in terms of two or more attributes. Some attributes are normatively relevant and some are not, but the latter are easier to use, and typically correlated (imperfectly) with the relevant ones. So people use the irrelevant ones and sometimes ignore the relevant ones completely.

2.3.1 Allocation

A great deal of research has examined how people think they should allocate benefits and burdens. Allocations can be local, such as the distribution of grades in a class or housework among those living together. But I focus here on policy. These issues include income, wealth, taxes, criminal penalties, tort fines, insurance, and compensation. Much of this research has examined the principles that people use for allocation decisions (e.g., Deutsch, 1975). These include equality (everyone gets the same); contribution (to each according to their contribution, also called "equity"); need (to each according to need); and maximization (e.g., maximization of total wealth — economic efficiency or total utility). But punishment also raises questions about distribution. What principle should determine criminal or tort penalties? Likewise compensation for misfortune, whether at the hands of nature or a harmful act of someone else; compensation is provided by insurance, social insurance, or tort penalties.

Utilitarian theory implies that distributions of goods (e.g., of income or wealth) should be based on maximization of utility, but this principle implies two other criteria: declining marginal utility of most goods; and incentive. A given amount of money means more to the poor than to the rich (i.e., the utility of money is marginally declining; that is, the slope of the curve relating utility to money declines as money increases). Hence, other things being equal, utility would be maximized if we took from the rich and gave to the poor until everyone is equal. However, this would prevent the use of income as an incentive for work (and has "transaction costs" of its own). Hence, maximization requires a compromise between equality and contribution. Such a principle is useless for psychology experiments. Even if it were possible to calculate the optimum trade-off, ordinary people would have no way of knowing the result. However, experiments can show deviations from any such model, even non-utilitarian models that incorporate similar assumptions. Such deviations can be explained in terms of simple heuristics such as equality, or demonstrated by framing effects, such as the Schelling effect, or the Harris and Joyce (1980) study, described above.

People sometimes prefer equality over maximization that involves lives rather than money.[9] Several studies (e.g., Ubel & Loewenstein, 1996) have presented subjects with allocation dilemmas of the following sort: Two groups of 100 people each are waiting for liver transplants. Members of group A have a 70% chance of survival after transplantation, and members of group B have a 30% chance. How should 100 livers — the total supply — be allocated between the two groups. The simple utilitarian answer is "all to group A", but only a minority of the subjects chose this allocation. People want to give some livers to group B, even if less than half. Many want to give half. Many people are are willing to trade lives for fairness to the two named groups. Surely there is some third group that is not in the scheme at all, so inequality is inevitable.

[9] As noted above in connection with the Asian Disease problem, declining marginal utility of lives is difficult to justify by any account.

2.3.2 Compensation and deterrence in tort law and criminal law

Compensation is justified by declining marginal utility. If you have a house fire that requires construction work, or an illness that requires expensive treatment, your utility for money increases. You have an immediate need for more of it. Insurance, including medical insurance and social insurance (such as unemployment compensation) is a scheme for transferring money from those who have a lower utility, those who pay insurance premiums or taxes, to those who have a higher utility. Like progressive taxes, compensation should be limited when its availability can provide incentive for reducing risks. For example, fire insurance could require installation of fire extinguishers. Health insurance may cost more for smokers, but this is justifiable only if this incentive effect actually causes people not to smoke.

Tort penalties and criminal penalties are justified by incentive effects, that is, by the principle of deterrence. If you know that you are likely to be punished or fined for some behavior (including omissions, in some situations), then you are less likely to engage in that behavior. Penalties "send a message" to the person penalized and to others, "Don't do this."[10]

Experiments (e.g., Baron & Ritov, 1993) are hardly needed to demonstrate that these principles are not followed in the real world. Many people advocate health insurance in which people pay premiums according to their individual "risk", even when that risk is caused by factors totally beyond any individual's control, hence not subject to incentive effects (e.g., genetic predispositions to certain diseases). Compensation is often provided to relatives for "wrongful death", even when the death in question reduces their need for money. And tort penalties are often levied even when the incentive effect leads to more harmful behavior (e.g., a lawsuit for side effects of a beneficial vaccine with rare side effects causes the company to withdraw the vaccine entirely; see Baron & Ritov, 1993).

Likewise, criminal punishments are often inconsistent with the principle of deterrence (Carlsmith, Darley & Robinson, 2002). Preferences for penalties are based more on the heinousness of the offense than on factors that should affect deterrence. For example, by utilitarian theory, the severity sof punishment should be higher when the probability of detection is lower; this way, potential offenders are risking a larger loss in the unlikely event that they get caught. But probability of detection plays little role. Littering is lightly penalized.

2.4 Comparison of moral judgments to consequence judgments

In some cases, such as the vaccination case described, the utilitarian answer is fairly clear. When it cannot be specified, a simple alternative approach is to

[10] In most but not all jurisdictions, tort penalties are used to fund compensation for victims. Theoretically, this political convenience is not necessary. Fines for deterrence could be paid into a general fund, and the fund could pay compensation when it is warranted, regardless of whether there is someone to sue.

ask the subject which option, on the whole, has the best overall consequences for everyone affected. When the subject gives one answer to that question and a different answer to the question of what is to be done, then we have pretty good evidence that the subject is giving a non-utilitarian answer, and we can go on to explore the reasons for this discrepancy.

Baron, Ritov and Greene (2013) asked subjects what was best for their nation (or national group, in the case of Arab and Jewish Israelis), what was best on the whole, what was best for the other group (in Israel), and what their moral duty was. Many subjects thought it was their duty to go against their own judgment of what was best on the whole, in the direction of parochialism (in-group bias).

Baron and Jurney (1993), asked subjects if they would vote for various reforms. In one experiment, 39% of the subjects said they would vote for a large tax on gasoline (to reduce global warming). Of those who would vote against the tax, 48% thought that it would do more good than harm on the whole. Of those subjects who would vote against the tax despite judging that it would do more good than harm, 85% cited the unfairness of the tax as a reason for voting against it (for instance, the burden would fall more heavily on people who drive a lot), and 75% cited the fact that the tax would harm some people (e.g., drivers). These subjects are apparently unwilling to harm some people in order to help others, even when they see the benefit as greater than the harm. This effect may be related to "omission bias". Unlike other results summarized here, the principle in question is non-utilitarian but is endorsed by other moral theories. Yet, its application in the real world can make things worse.

2.5 Isolation effects

In "isolation" effects, people attend only (or primarily) to data or issues immediately before them (Camerer, 2000; Kahneman & Lovallo, 1993; Read, Loewenstein & Rabin, 1999). These effects are related to, or identical to, what others have called a *focusing effect* (Idson et al., 2004; Jones et al., 1998; Legrenzi et al., 1993) People know about indirect effects but do not consider them, or do not consider them enough. The idea came from the theory of mental models in reasoning (Legrenzi et al., 1993): people reason from mental models, and when possible they use a single, simple, model that represents just the information they are given. Other factors are ignored or under-used.

McCaffery and Baron (2006) found apparent isolation effects in evaluation of taxes and other policies. For example, people prefer "hidden" taxes, such as a tax on corporations, without thinking about where the money comes from (employees, consumers, stockholders). If people are asked who actually pays, they realize that such taxes are not "free". Caplan (2007) reports similar effects for policies such as rent control, which have an immediate desirable effect on prices but an undesirable secondary effect on the supply of housing. Often people seem to evaluate policies (such as long prison sentences) in terms of their intended effects, even if those are not their main effects.

3 Moral rules and intuitions

Many demonstrations of non-utilitarian biases, or their cognitive-bias cognates, seem to result from intuitive responses rather than any sort of reflection. At issue is whether these biases would be reduced if people engaged in more thinking, or more thinking of a certain sort.

Hare (1981; see https://www.utilitarianism.net/ for additional citations) proposed a related account. In defending utilitarianism, he proposed a two-level theory of moral thinking, with an intuitive and "critical" level. The critical level is utilitarian and is rarely approximated in human thinking, and also rarely needed. To make a decision at this level, a person must sympathetically represent to herself the preferences of all those affected and reach a decision as if the conflicts among their preferences were conflicts among her own preferences. Such decisions can apply to moral principles as well as specific cases; the principles accepted through such reflection are those that would be rationally accepted by anyone, even if that person in real life would lose from application of the principle. Hare argues that the term "moral" implies such universal agreement (an idea he attributes to Kant). The principles are thus universal, but each principle (unlike heuristics or intuitive rules) need not be simple. It may include all morally relevant features of a given case.

The intuitive level consists of simpler rules, which (at best) approximate the conclusions of the critical level for the situations that people usually encounter and sometimes become quite strong (as in the case of PVs). Thus, conflicts between intuition and utilitarian conclusions are most apparent in unusual, hypothetical cases (such as the footbridge version of the trolley problem) that do not arise in most people's lives (but do occasionally arise in fiction, as well as articles in philosophy journals).

3.1 Intuitions and dual-systems

Many approaches to reasoning have relied on the idea of dual systems, intuitive and reflective, with at least the intuitive level being similar to Hare's. That system, by various accounts, is automatic (uncontrollable, but also free of demands on cognitive resources), driven by emotion (or affect), and based on associations rather than rules. The reflective system is on resources, and controllable. Because it is controllable, it may or may not become active after the intuitive system has done, or at least begun, its work. In principle, if the subject knows that reasoning is required, the reflective system could begin right away. Kahneman (2011) argues that a correcitive version of this theory, in which reflection begins after results of intuition are available and can function to correct the intuition, is relevant to a variety of tasks studied in the heuristics-and-biases tradition. The corrective theory has also been proposed as an account of moral judgment by Joshua Greene (e.g., 2007, p. 44, although elsewhere Greene is less specific about the ordering of events in time).

Several lines of evidence seem to support the dual-system theory for moral judgment. First, response times (RTs) for "personal" dilemmas, those that

involve direct killing, such as the footbridge dilemma, are longer, especially when subjects choose the utilitarian option.

A common finding in choice tasks is that RT is slower when the response is rarely made or when the options are similar in attractiveness (hence conflicting). These factors alone can explain RT differences found. Note that the corrective model implies that RT is longer for utilitarian than for deontological responses when their probability is equal (which is also where the two responses are maximally conflicting). Baron & Gürçay (2017), in a meta-analysis of 26 experiments, estimated this RT for each response by assuming that each subject had an "ability" to make utilitarian responses, and each dilemma had a "difficulty" for making that response. (Thus, the footbridge problem is more "difficult" than the simple trolley case.) The two choices would be equally likely when ability was equal to difficulty, according to our measures. A plot of RT for each choice as a function of ability minus difficulty indeed showed the slowest RT when this difference was zero, but, at this point, the utilitarian responses took no longer than the deontological responses. Rosas, Bermúdez, and Aguilar-Pardo (2019) also found that RTs were determined mainly by conflict. These results are inconsistent with any form of the corrective model.

Baron & Gürçay (2017) also noted that subjects who made more utilitarian responses had longer RTs on everything, a result consistent with the claim that reflection-impulsivity is correlated with utilitarian responding. Why this happens may depend on developmental processes that have already occurred before the experiment. For example, people who are generally reflective may come to favor utilitarian solutions over time.[11]

This kind of account in terms of individual differences in reflection is not far from Greene's two-system account, but it does not assume any sequential effects involving suppressing an early response by a late one, so it is thus consistent with the known results, and with versions of two-systems theory that assume that the systems work in parallel rather than sequentially (e.g., Sloman, 1996). It is clear by any account that people differ in some sort of reflectiveness, and these differences are related to differences in at least some moral dilemmas (Patil et al., 2021).

Other results concern the effects of time pressure or cognitive load, which, in some studies, seem to affect utilitarian responses but not deontological responses. A general problem with these studies is that the effects vary for different dilemmas, not only in magnitude but also in direction (as also found by Gürçay & Baron, 2016, despite finding no overall effect of time pressure vs. instructions to reflect). For example, to deal with load or time pressure, subjects may skip or skim the less salient parts of the printed description, and those may vary with how the dilemma is described. Researchers should at least test effects in ways that take into account the variance across dilemmas as well as across subjects, and most researchers have not done this (for an exception, see Patil

[11] A personal example: As a child I was puzzled by the proverb "Two wrongs don't make a right," which suggests that punishment is wrong. Ultimately I figured out that punishment can be justified if it prevents future wrongs. Later I learned that this was the utilitarian solution to the puzzle.

et al., 2021), as well as trying different ways of ordering the information in the dilemma..

These sorts of results concerning time pressure and cognitive load have been difficult to replicate (e.g., Bago & De Neys, 2019). Rosas and Aguilar-Pardo (2020) found that utilitarian responses can occur under extreme time pressure. Moreover, studies that track the position of the pointer (mouse) during experiments with moral dilemmas do not show any tendency to switch from utilitarian to deontological responses during the time (usually 10–20 sec) that the subject deliberates (Gürçay & Baron, 2016; Koop, 2013).

4 Future directions

A lot remains to be known about moral reasoning. The reader who has gotten this far will not be surprised that I think this topic should be part of cognitive psychology, which has been studying reasoning more generally for well over a century. Many of the methods of psychology remain to be fully applied to moral reasoning. But moral reasoning is important for practical purposes too. It is tied up with politics and public policy.

Political judgments of citizens are often moral judgments. These merit special attention because the actions or omissions of citizens affect other citizens and non-citizens at home and abroad. Many of the world's problems, within and among nations, can be traced to policies approved by citizens. The utilitarian argument I made earlier applies here. If citizens collectively follow non-utilitarian moral intuitions, then we should not be surprised if the final results they influence are deficient, for all those affected, everywhere.

Differences in thinking about politics arise in individual development (as studied by Kohlberg, 1963, Adelson, 1971, and many others) and in cultural evolution. Hallpike's (2004) analysis, which is analogous to that of Kohlberg, suggests that something like developmental stages occurred over the course of cultural evolution, with the earlier stages still present. Early people, and those who still live as they did, and young chidren, do not distinguish morality, laws and social conventions, and etiquette. They are just "the way we do things". With the growth of cities and writing, codified, impartial, laws came to exist and were soon "written in stone" or in parts of what is now the Old Testament. Similar developments may occur in early adolescence (depending on culture, see Haidt et al., 1993). The development of a concept of morality, independent and outside of laws and conventions, came relatively recently, roughly the last 3,000 years, and is still not fully understood by many people. For most who make this distinction now, it comes in adolescence. The existence of a concept of morality raises the possibility of rational thought about what it should be.

It is apparent that culture has a large effect not only on moral beliefs but also on how (or whether) people reason about them, or about anything. A question of interest is how cultural traditions persist over generations and over historical time (even within generations), and how they change. Attitudes toward homosexuality, for example, have changed enormously in the last 50 years,

in some countries. And it is clear that there are cultural influences on beliefs about what good thinking is. One way to study cultural change over time is to examine written documents, both for their content and for the type of reasoning they exhibit. Some of this sort of work has been done (Suedfeld, 1985; Suedfeld, Guttieri & Tetlock, 2003), but it has been confined to specific groups that are not representative of any larger cultural tradition.

It is clear that education can be designed to encourage rational thinking (e.g., Baron, 1993). Liberal education at the university level is often explicit in its attempts to encourage questioning, consideration of diverse views, and understanding of the nature of expert knowledge. Many secondary schools do this too (e.g., Metz et al., 2020).

Several efforts have been made to teach moral thinking in a way that views it as a type of thinking rather than a set of rules. Kohlberg, in particular, encouraged widespread experimentation with moral discussion in high schools around the world (Snarey & Samuelson, 2008). Much of this work disappeared with Kohlberg's death, and with claims that his ideas were biased against women (claims that were consistently shown to be unfounded, as Snarey and Samuelson point out).

Education is one important domain where people's thinking can be influenced. Others, probably to a lesser extent, are journalism and politics itself. Ultimately, individuals and cultures change from a variety of influences, and we cannot expect applied research on one domain or another to provide the key. Change is slow, but the world would benefit if people's moral thinking became more rational.

References

Adelson, J. (1971). The political imagination of the young adolescent. *Daedalus, 100*, 1013–1050.

Asch, D., Baron, J., Hershey, J. C., Kunreuther, H. C., Meszaros, J., Ritov, I., & Spranca, M. (1994). Determinants of resistance to pertussis vaccination. *Medical Decision Making, 14*, 118–123.

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.

Baron, J. (1992). The effect of normative beliefs on anticipated emotions. *Journal of Personality and Social Psychology, 63*, 320–330.

Baron, J. (1993). Why teach thinking? — An essay. *Applied Psychology: An International Review, 42*, 191–237.

Baron, J. (2012a). Parochialism as a result of cognitive biases. In R. Goodman, D. Jinks, & A. K. Woods (Eds.), *Understanding social action, promoting human rights*, pp. 203–243. Oxford: Oxford University Press.

Baron, J. (2012b). The "culture of honor" in citizens' concepts of their duty as voters. *Rationality and Society, 24*, 37–72.

Baron, J. & Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory and Cognition, 45*(4), 566–575.

Baron, J., Gürçay, B., & Luce, M. F. (2018). Correlations of trait and state emotions with utilitarian moral judgments *Cognition and Emotion, 32*(1), 116–129.

Baron, J., & Leshner, S. (2000). How serious are expressions of protected values. *Journal of Experimental Psychology: Applied, 6*, 183–194.

Baron, J. & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty, 7*, 17–33.

Baron, J. & Ritov, I. (1994). Reference points and omission bias. *Organizational Behavior and Human Decision Processes, 59*, 475–498.

Baron, J., & Ritov, I. (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral Judgment and decision making*, Vol. 50 in B. H. Ross (series editor), *The Psychology of Learning and Motivation*, pp. 133–167. San Diego, CA: Academic Press.

Baron, J., Ritov, I., & Greene, J. D. (2013). The duty to support nationalistic policies. *Journal of Behavioral Decision Making, 26*, 128–138.

Bhatia, S., Walasek, L., Slovic, P. & Kunreuther, H. (2021), The more who die, the less we care: Evidence from natural language analysis of online news articles and social media posts. *Risk Analysis, 41*, 179–203.

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review, 7*, 129–145.

Bornstein, G. & Ben-Yossef, M. (1994). Cooperation in intergroup and single-group social dilemmas. *Journal of Experimental Social PsYchologsy, 30*, 52–57.

Breyer, S. (1993). *Breaking the vicious circle: Toward effective risk regulation.* Cambridge, MA: Harvard University Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Buchan, N. R., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., & Foddy, M. (2009). Globalization and human cooperation. *Proceedings of the National Academy of Sciences, 106*, 4138–4142.

Buchan, N. R., Brewer, M., Grimalda, G., Wilson, R., Fatas, E., & Foddy, M. (2011). Global social identity and global cooperation. *Psychological Science, 22*, 821–828.

Caplan, B. (2007). *The myth of the rational voter: Why democracies choose bad policies.* Princeton, NJ: Princeton University Press.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284–299.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271–280.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380.

Cushman, F., & Young, L. (2011), Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*, 1052–1075.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues, 31*, 137–149.

Edlin, A., Gelman, A., & Kaplan, N. (2007). Voting as a rational choice: Why and how people vote to improve the well-being of others. *Rationality and Society, 19*, 293–314.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics, 75*, 643–699.

Foot, P. (1978). The problem of abortion and the doctrine of the double effect. In P. Foot, *Virtues and vices and other essays in moral philosophy*, pp. 19–32. Berkeley: University of California Press. (Originally published 1967 in *Oxford Review*, no. 5, 1967.)

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Greene, J. D. (2007). The secret joke of Kant's soul, in W. Sinnott-Armstrong, Ed., *Moral psychology, Vol. 3: The neuroscience of morality: Emotion, disease, and development*, pp. 36–79. MIT Press, Cambridge, MA.

Greene, J. D., Cushman, F. A., Stewart. L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009) Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*, 364–371.

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*, 2105–2108.

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-systems model of moral judgment. *Thinking and Reasoning, 23*, 49–80.

Haidt, J. (2001). The emotional dog and its rational tale: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, Fall 2004*, 55–66.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or, Is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.

Hallpike, C. R. (2004). *The evolution of moral understanding.* London: Prometheus Research Group.

Hanselmann, M., & Tanner, C. (2008). Taboos and conflicts in decision making: Sacred values, decision difficulty, and emotions. *Judgment and Decision Making, 3*, 51–63.

Hare, R. M. (1952). *The language of morals.* Oxford: Oxford University Press (Clarendon Press).

Hare, R. M. (1981). *Moral thinking: Its levels, method and point.* Oxford: Oxford University Press (Clarendon Press).

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69*, 366–378.

Harris, R. J., & Joyce, M. A. (1980). What's fair? It depends on how you phrase the question. *Journal of Personality and Social Psychology, 38*, 165–179.

Johnson, E. J., Hershey, J. C., Meszaros, J., & Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty, 7*, 35–51.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Strauss and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*, pp. 49–81. New York. Cambridge University Press.

Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods: A study of the headline method. *Journal of Risk and Uncertainty, 9*, 5–38.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Kohlberg, L. (1963). The development of children's orientations toward a moral order. I. Sequence in the development of human thought. *Vita Humana, 6*, 11–33.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making, 8*, 527–539.

Kunreuther, H., & Slovic, P. (1978). Economics, psychology, and protective behavior. *American Economic Review, 68*, 64–69.

McCaffery, E. J. (1997. *Taxing women.* Chicago: University of Chicago Press.

McCaffery, E. J., & Baron J. (2006). Thinking about tax. *Psychology, Public Policy, and Law, 12*, 106–135.

McCaffery, E. J., & Baron J. (2006). Isolation effects and the neglect of indirect effects of fiscal policies. *Journal of Behavioral Decision Making, 19*, 1–14.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a look at the evidence.* Minneapolis: University of Minnesota Press.

Metz, S. E., Baelen, R. N., & Yu, A. (2020) Actively open-minded thinking in American adolescents. *Review of Education, 8*, 768–799.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment.* New York: Cambridge University Press.

Mill, J. S. (1859). *On liberty.* London: J. W. Parker & Son.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology, 120*(2), 443–460.

Piaget, J. (1932). *The moral judgment of the child.* Glencoe, IL: The Free Press.

Polya, G. (1945). *How to solve it: A new aspect of mathematical method.* Princeton: Princeton University Press.

Rosas, A., & Aguilar-Pardo, D. (2020): Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking and Reasoning, 26*, 534–551.

Rosas, A., Bermúdez, J. P., & Aguilar-Pardo, D. (2019). Decision conflict drives reaction times and utilitarian responses in sacrificial dilemmas. *Judgment and Decision Making, 14*, 555–564.

Ross, W. D. (1930). *The right and the good.* (Reprinted 2002 by Oxford University Press, Oxford.)

Roth, A. E. (2007). Repugnance as a constraint on markets. *Journal of Economic Perspectives, 21*, 37–58.

Royzman, E. B. & Baron, J. (2002). The preference for indirect harm. *Social Justice Research, 15*, 165–184.

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Shou, Y., & Song, F. (2017). Decisions in moral dilemmas: The influence of subjective beliefs in outcome probabilities. *Judgment and Decision Making, 12*, 481–490.

Sherman, G. D., Vallen, B., Finkelstein, S. R., Connell, P. M., Boland, W. A., & Feemster, K. (2021). When taking action means accepting responsibility: Omission bias predicts parents' reluctance to vaccinate due to greater anticipated culpability for negative side effects. *Journal of Consumer Affairs*, 1–22.

Singer, P. (1982). *The expanding circle: Ethics and sociobiology.* New York: Farrar, Strauss & Giroux.

Singer, P. (2002). R. M. Hare's achievements in moral philosophy. *Utilitas, 14* (3), 309–317.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making, 2*, 79–95.

Sunstein, C. R. (2002). *Risk and reason: Safety, law, and the environment.* New York: Cambridge University Press.

Snarey, J., & Samuelson, P. (2008). Moral education in the cognitive developmental tradition: Lawrence Kohlberg's revolutionary ideas. In L. P. Nucci & D. Narvaez (eds.), *Handbook of moral and character eductation*, pp. 53-79. Routledge.

Suedfeld, P. APA presidential addresses: The relation of integrative complexity to historical, professional, and personal factors. *Journal of Personality & Social Psychology, 49*, 1643–1651.

Suedfeld, P., Guttieri, K., & Tetlock, P. E. (2003). Assessing integrative complexity at a distance: Archival analyses of thinking and decision making. In J. M. Post (Ed.), *The psychological assessment of political leaders: With profiles of Saddam Hussein and Bill Clinton*, pp. 246–270. Ann Arbor, University of Michigan Press.

Sunstein, C. R. (2005). Moral heuristics (with commentary). *Behavioral and Brain Sciences, 28*, 531–573.

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Sacred versus pseudo-sacred values: How people cope with taboo trade-offs. *American Economic*

Review, 107(5), 96–99.
Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions *Trends in Cognitive Sciences, 7*(7), 320–324.
Thaler, R. H. (2015). *Misbehaving: The making of behavioral economics.* New York: W. W. Norton & Co.
Tversky, A. (1967). Additivity, utility, and subjective probability. *Journal of Mathematical Psychology, 4*, 175–202.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.
Ubel, P. A., & Loewenstein, G. (1996). Distributing scarce livers: The moral reasoning of the general public. *Social Science and Medicine, 42*, 1049–1055.