Moral judgment: Acts, omissions, and rules

Jonathan Baron*

September 20, 2013

Contents

*Draft chapter for *Blackwell Handbook of Judgment and Decision Making*, edited by G. Keren and G. Wu. Rejected by the editors.

Moral (mis)judgment is at the heart of many problems in the world. People who commit horrendous acts, and those who tolerate them, are often behaving out of pre-existing moral conviction. Such phenomena are, as they should be, studied from several perspectives.

Here, I review only research in the tradition of the field of judgment and decision making (JDM). Thus, I largely omit the extensive related work in developmental, social, and clinical psychology. Moreover, I concentrate on a couple of representative issues, specifically the role of acts, omissions, and moral rules in accounting for judgments.

The JDM approach is based on normative, descriptive and prescriptive models. Application of this approach to moral judgment is more difficult than to other types of judgment, because the normative models themselves are particularly controversial. Here I assume utilitarianism as a normative model, and I discuss research that at least implicitly accepts the value of using utilitarianism as a point of comparison. Utilitarianism says that the best option is the one that does the most total good. If judgments lead to choices that do less good than is possible with other options, then it seems worthwhile to know whether the cause of such deficiencies lies in the moral principles that people follow. In particular, *deontological* principles concern rules that make reference to features of action other than consequences. Such principles might in some sense be morally better than utilitarianism, but, by comparing them to the rule of choosing the option with the best expected outcome, at least we learn the cost of such moral superiority.

Utilitarianism needs to be defended as a normative model, just as other normative models require defense (Baron, 1994). I shall not do that here. For the defenses that are perhaps most relevant to JDM, see Baron (1993a, 1996, 2006, 2008) and Hare (1981).

1 Moral heuristics and biases

Some research in moral judgment has followed the JDM tradition of looking for biases that could be attributed to the use of heuristics. Most authors are quite cautious about the term "bias", because they do not want to commit themselves to a normative theory, but much of the research on "moral heuristics" makes the most sense if we think that utilitarianism — the theory that the best option is the one that does the most overall good (or, equivalently, least harm, since there is no zero point for utility) — is at least tentatively adopted as a point of comparison (e.g., Sunstein, 2005).

This approach was (to my knowledge) first applied explicitly by Ritov and Baron (1990) and Spranca, Minsk, and Baron (1991), and defended more generally by Baron (1994). The first two studies were concerned with the distinction between acts and omissions. For example, Spranca et al. asked subjects to compare two endings to a story, in which Ellen had decided to lie to the police to protect her friend. In one ending, Ellen lied. In the other, she was about to lie when she realized that the police had decided incorrectly that her friend was

innocent, and she failed to correct their false belief. Subjects were asked to compare Ellen's morality in the two endings. Although many subjects thought she was equally wrong in both cases — given that her intent was the same and the outcome, which she could easily have prevented, was the same — others thought that she was more wrong when she acted.

Ritov and Baron (1990) asked about giving a vaccine that would prevent a disease but would cause side effect that were equivalent to the disease in some smaller number of people. Failing to give the vaccine was a harmful omission, which was more harmful than the act of vaccinating, yet many subjects favored not vaccinating.

Another pair of scenarios (based on examples used by Foot, 1978, and other philosophers) used extensively in research concerns a runaway trolley that will kill five people if nothing is done. In one version, you can switch the trolley to a different track, where it will kill one person instead of five. In the "fat man" version, you can stop the trolley only by pushing a large man off of a bridge on to the trolley track, thus killing the man but saving the five others. Most (but not all) subjects think it is morally right to switch the trolley in the first case (the utilitarian answer, because more lives are saved) but not in the second.

2 Protected (sacred) values (PVs)

Another example of a possible non-utilitarian bias in mofal judgment is the existence of moral rules that are taken as absolute, not to be violated no matter what the cost. For example, people may think that it is always wrong to kill innocent people, no matter what benefits result from doing so (including, perhaps, the benefit of saving others from death). When people explicitly affirm such rules, even when they admit that the benefits exceed the costs of violations, then this affirmation violates utilitarianism. Absolute values as such do not necessarily violate it: people might say that it is impossible for the benefits of breaking the rule to ever exceed the costs. While such judgments do not violate utilitarianism, they often seem odd, and they sometimes conflict with other judgments, e.g., when two absolute rules conflict with each other.

Baron and Spranca (1997) called such rules "protected values" (PVs) because they were protected from trade-offs with other values. Many of these values involve morally questioned behaviors, such as cloning people for reproduction or destroying the natural environment. Other researchers have used the term "sacred values" for what seems to be the same phenomenon, although they measure it differently (Fiske & Tetlock, 1997; Tetlock, Lerner, & Peterson, 1996). Roth (2007) has also used the term "repugnant transactions" for moral prohibitions on transactions such as a live donor selling a kidney.

Rules of this sort, if they were taken seriously, would cause great difficulty for evaluation of public policies through measurement of utility, because they would amount to infinite utilities and would thus allow one person to determine a decision for everyone — unless someone else had a conflicting rule, in which case the choice could not be determined. People who endorse more than one

such rule — and many people endorse several — could find themselves in a similar dilemma.

It appears that absolute rules are often over-generalizations. When people are asked to try to think of counterexamples, cases in which the benefits would be great enough to justify breaking the rule, they can usually do so, and they change their mind about whether the rule is absolute (Baron & Leshner, 2000). Thus, PVs could be explained psychologically as the result of failure to think critically about rules. A rule against killing sounds good, until one tries to think of counter-examples. Of course, over time, people may become committed to some rules, so that they resist counter-arguments. Thus, rules may be maintained by "myside bias" (Baron, 2008).

PVs are closely related to, and correlate with, other types of values or goals, which express themselves in other ways. In particular, they are relates to *moralistic goals*. These are goals or values are those that people want others to follow, regardless of whether the others endorse the same goals and regardless of whether the consequences are, on the whole, worse as a result. Baron (2003) found that people endorsed moralistic goals for banning actions like the following:

- testing a fetus for IQ genes and aborting it if its expected IQ is below average
- cloning someone with desired traits so that these may be passed on, such as an athletic champion or brilliant scientist
- modifying the genes of an embryo so that, when it is born, it will have a higher IQ
- giving a drug (with no side effects) to enhance school performance of normal children

In many cases (22% of examples like these), subjects (recruited on the World Wide Web) would ban these actions even if the consequences of allowing the actions were better on the whole than the consequences of banning them, if the subjects could imagine that the consequences might be better, and if "almost everyone in a nation thought that the behavior should be allowed." In sum, they were willing to impose their moral(istic) principles on others, whatever the consequences, and whatever the others wanted. Moralistic values, taken seriously, are protected from trade-offs with values involving consequences, so they are likely to be thought of as PVs.

3 Omission, action, and related issues

Omission bias, another potential bias, is the tendency to judge acts that are harmful (relative to the alternative option) as worse than omissions that are equally harmful (relative to the alternative) or even more harmful (as in the

vaccination case) (Baron & Ritov, 1994). In any given case, some people display this bias and others do not.

Omission bias is related to controversial issues, such as whether active euthanasia should be allowed. Passive euthanasia, the withholding of standard medical treatment for those who are judged to be no worse off dead than alive is widely permitted, but active euthanasia and assisted suicide are banned in most places even for those who wish to die. Opponents of active euthanasia can, of course, find other arguments against it than the fact that it is "active". But it is possible that these arguments would seem less compelling if the act/omission distinction were not involved.

Omission bias could also justify a lack of concern with the problems of others (Singer, 1993). For example, many people — even people who take an interest in social issues — often think that they are not "responsible" for this poverty and thus need do nothing about it. Singer (1993) argues, however, that with a little effort we can think of many helpful things we can do at very low cost to ourselves, such as supporting beneficial policies. Failure to do these things can be seen as a harm, but many people do not see it that way.

3.1 Determinants of omission bias

Omission bias has been studied extensively, and we can draw some conclusions about why it happens. In particular, research has identified several moderators of the effects, i.e., manipulations that affect its frequency, and correlates of its presence.

3.1.1 Perceived causality

An early and much replicated result is that omission bias is correlated with act/omission differences in perceived causality. When people say that harms of action are worse than equivalent harms of omission, they also tend to think that the causal relation between the person and the harm is greater in the action (Spranca et al., 1991; Baron & Ritov, 2009).

Two concepts of causality compete here and elsewhere: "but for" (*sine qua non*) causality; and physical causality. But-for causality is relevant in tort law and sometimes elsewhere in the law. We say that person P causes outcome O, in this sense, if O was affected by P's choice.[1] But-for causality does not distinguish acts and omissions. Thus, people can be held liable for omissions. But-for causality is also relevant in utilitarian morality, which is about choices and consequences and which ignores other properties of options aside from consequences.[2]

In physical causality, P's behavior is linked to O through a series of physical events, each of which presumably follows some physical principle. When people

[1] If P's choice affected O probabilistically, we say that it was a partial cause or a contributing cause.

[2] Of course, different ways of bringing something about may have different side effects, and these consequences must also be considered.

distinguish acts and omissions morally, they seem to be basing moral judgments on this kind of causality.

But-for causality may be harder for young children to learn. The appreciation of it requires evaluation of counterfactuals. Although children as young as 5 can distinguish cases of causality by omission from non-causality (Schleifer, Shultz, & Lefebvre-Pinard, 1983), the distinction appears to be weak. The cases were very simple. For example, a store owner did not put salt on the ice in front of his store. In one condition, a customer slipped on the ice and was injured. In the control condition, the customer fell before reaching the ice, and was injured.

3.1.2 Physical proximity, contact, and personal force

People consider harm to be worse when it involves physical contact with the victim, such as pushing a man off a bridge to stop a trolley, as opposed to switching the trolley to a different track. Greene et al. (2009) showed that, in one situation, this sort of case depends on "personal force" more than on contact itself, and proximity alone had no effect at all. Personal force means that the force that directly affects the victim "is generated by the agent's muscles, as when one pushes another with one's hands or with a rigid object."

3.1.3 Protected values

PVs are also related to omission bias. Absolute rules would, to put it mildly, be difficult to follow if they were neutral between acts and omissions. If you think that is abortion is absolutely wrong no matter what the benefit, it is easy for you to take no action that causes an abortion, but it is extremely difficult to avoid omissions that lead to abortion. If you tried to do this, you would be morally obligated spend your time doing little else. If you had two such PVs against omission, then you would be in real trouble. People seem to recognize this logical asymmetry, and, as a result, they endorse PVs against action more than PVs against omission (Baron & Ritov, 2009).

When people have PVs against some action, then that action usually shows strong omission bias. If asked, for example, whether it is right to kill one person in order to save 5, 50, or 500 others, the numbers don't matter. A person with a true PV against active killing will always say no, and many subjects do exactly this (Ritov & Baron, 1999; Baron & Ritov, 2009).

3.2 Related biases

Several biases in moral judgment are related to omission bias. In many cases, the biases are confounded, so that it is impossible to say which one is present. I know of no attempt to distinguish individual differences in these. It might turn out that some common factor accounts for many of them, such as perceived causality. (Baron & Ritov, 2009, provide a more detailed review.)

3.2.1 Indirectness and the double effect

The indirectness bias is illustrated in the doctrine of the double effect. For example, when a mother's life is threatened by a pregnancy, some Catholic hospitals will permit a hysterectomy to save the mother, but they will not permit an abortion. The fetus dies in either case, but, in the case of the hysterectomy (which of course leaves the mother infertile), the killing is seen as an indirect by-product (Bennett, 1966; Kuhse, 1987). In the abortion the death of the fetus is the means to save the mother, so the fetus is being harmed directly. This indirectness bias has been studied by Royzman & Baron (2002).

3.2.2 Agent relativity

If you have an obligation to help your cousin, do I also have an obligation to advise you to help your cousin? If helping cousins is agent-relative, then the answer is no. Utilitarians argue against such agent-relative obligations. If it is a good thing for people to help their respective cousins, then it is good to promote it, and it is agent-general.

Omission bias is agent relative when the harm from omission is the result of someone else's action. In the classic case of shooting one prisoner to save ten from being shot by a horrible dictator, the choice of not shooting is implicitly agent relative, because shooting will happen anyway.[3] Baron and Miller (2000) found no evidence for agent-relativity in a cross-cultural study done in the U.S. and India.

3.2.3 Naturalism

Naturalism is the bias toward nature. It is also related to omission bias, because "nature" often defines the default situation, the result of inaction, as in the case of the vaccination, where the disease can be assumed to be natural. Of course, the result of omission is not always natural, as in the case of the dictator just described. (Rudski, Osei, Jacobson and Lynch, 2011, provide a recent review.)

4 The psychological basis of moral biases

The results for omission bias are perhaps the clearest illustration of the idea of moral heuristics and biases, so at this point I will step back and discuss some general issues in the psychology of these biases. Then I will review additional results.

[3]This is not a pure test of agent relativity, though, because the options also differ in doing something versus doing nothing. The agent should be required to shoot at a target to indicate to the dictator that he will not shoot the prisoner.

4.1 Two-systems theory

Many approaches to moral judgment have relied on various sorts of two-systems, or two-levels theory, a lower one and a higher one. The lower system is, by various accounts, automatic, unreflective, driven by emotion (or affect), based on associations rather than rules, and undemanding on limited cognitive resources. The higher system is the opposite, and is sometimes said to kick in only after the lower system has produced a tentative judgment.

4.1.1 Hare's two-level theory

It is interesting to view these theories in the light of the approach of Richard Hare (1981), a philosopher. In defending utilitarianism, he described essentially 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 himself the preferences of all those affected (weighing people equally) and reach a decision as if the conflicts among the preferences were conflicts among her own preferences. This method follows, Hare says, from analysis of what we mean by moral judgments, but I shall not explain that here other than to say that moral principles are by definition universal (among their other features). That is, they may describe situations in detail but cannot contain people's identities, as these are morally irrelevant; a moral principle must apply to anyone in exactly the same situation.

Intuitive principles can be justified at the critical level. They are the principles we follow most of the time, and those we should follow if they are well justified. Such justification may imply that we should always follow certain principles, even when we think it would be better for everyone if we did not (as I explain later). Most of the time, intuitive principles are *prima facie* and can be overridden by critical-level thinking or by other intuitive principles. Intuitive principles, like ecologically useful heuristics, are designed for real, representative cases.

At the intuitive level, principles can conflict, leading to feelings that all options are wrong. Conflict of this sort cannot occur at the critical level. Intuitive principles are like heuristics in that they are rules that should, when they are well chosen, help us make the judgments that we would make with critical thinking, with far less effort.

The critical level can justify following some intuitive principles as absolute rules. When we face unusual situations in real life, and we think we know the utilitarian right answer, we should, if we are good utilitarians, also consider the probability that our perception is incorrect, which can sometimes be quite high. Most adulterers and terrorists think that their behavior serves the greater good (that their spouses will not find out and thus will not suffer harm, or that a massacre of non-combatants will bring on Utopia), yet they might temper their judgments if they considered that practically every adulterer and terrorist before them thought the same thing, and in hindsight most of them were wrong.

People faced with unusual *hypothetical* situations might carry over such caution without fully knowing where their caution comes from. Or they might know very well.

In this way, absolute rules such as PVs might have a utilitarian justification. In experiments on PVs, it is worth giving subjects, as one of the response options, something like, "Although I can imagine cases in which this ought to be done, I do not think that anyone is capable of recognizing such cases accurately, so they should always follow the rule against doing it."

Intuitive principles are drummed into us before we could even understand what critical thinking is, either by our parents or perhaps even by evolutionary selection. Because of this, we feel compunction (or guilt) when we violate them, and we anticipate such feelings when we think about violating them. Such anticipated feelings may affect the choices we make in experiments. Moreover, such strong intuitions may be part of a good utilitarian character. It may be very difficult to induce people to feel strongly about morality yet at the same time be willing to make exceptions for such unusual special cases. And the apparent utilitarian answer may not even be the best utilitarian answer, which may take into account the effect of choices on others.

Hare's theory has methodological implications for philosophical inquiry. Many philosophers argue that moral intuitions provide useful data for developing a normative theory of morality. Hare argues that these intuitions can be understood as products of the intuitive level. Some of them are highly dependent on culture (e.g., exposure to Christianity, which has influence social norms even in non-Christians). Other intuitions work well (from a critical-level view) in realistic cases but not in the fantastical cases of the sort used in experiments and in philosophical argument. Thus, the philosophical approach of criticizing utilitarianism by finding conflicts with intuition in such unusual situations is invalid from the outset. Greene (2007) makes a similar argument.

4.1.2 Greene's theory and some problems with it

In the same article (following earlier articles), Greene proposed a two-system theory that is related to Hare's. System-1 is, fast, automatic and effortless. System-2 requires effortful thinking. Green also proposes that system-1 is influenced by emotional responses, more than system-2. Thus, in a dilemma such as the fat-man version of the trolley problem, people have an immediate, automatic emotional response to the idea of pushing a man to his death, and this leads them to want to say that it would be wrong to do so. Then, some people will reflect before they make this response, using system-2, and decide that they would not want to let five others die through their inaction. The reflection that occurs in Greene's theory is not necessarily the same as what occurs at Hare's critical level. More likely, it is the intrusion of what Hare would call a competing intuition, that it is better to save more lives. Such a principle might well be adopted as overriding at the critical level, but we have almost no reason to think that anything approaching Hare's idea of critical thinking is occurring in an experiment where people spend less than a minute thinking on what is, for

a utilitarian, a simple case.

Several lines of evidence support this theory, yet questions can be raised about each one, particularly about whether the two systems operate sequentially.

First, response times (RTs) for "personal" dilemmas like the fat-man are longer, especially when subjects endorse the utilitarian option. Baron, Gürçay, Moore & Starcke (2012) argue that this result can be explained in terms of conflict. When choices are difficult, so that the subject is as likely to respond yes as no, RT is long. (Most subjects respond "no" to the fat-man have fast RTs.) By the two-system theory, at this point, it should still be longer for yes than no responses, because yes responses require an extra step. This result is not found.

Second, cognitive interference slows down RT to utilitarian responses but not deontological (non-utilitarian, rule-based) responses (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). The interference, however, involved arithmetic, and it was necessary to process the numbers to give the utilitarian response. Also, the deontological response could be based on a less thorough reading of the dilemma, considering only the type of action to be taken and not the numbers, which form part of the background that makes a normally immoral action morally reasonable.

Third, Suter and Hertwig (2011) claimed that instructing people to go fast or slow affected their responses. Yet they found this only for a few selected dilemmas, and a reanalysis of their data shows no overall effect. Moreover, Gürçay and Baron (in preparation) have failed to find such an effect in two studies.

Fourth, utilitarian responding correlates with cognitive reflection as a trait (e.g., Paxton, Ungar & Greene, 2011). Cognitive reflection is measured by test consisting of three short arithmetic problems with "obvious" answers that turn out to be incorrect, thus apparently requiring correction of a system-1 response by system-2 (Frederick, 2005). This is a more interesting result, but a correlation like this does not imply that correction is involved in the moral judgment task, or even in the arithmetic test itself. A person who is not distracted by the trick answers in the arithmetic test might just adopt an attitude of using system-2 from the outset, analyzing each problem without even being tempted to take a guess at the answer. Similarly, in moral judgment, people may set out to look at all the information, including side effects of doing nothing, before even making a tentative judgment. More generally, the correlation could result from individual differences in reflection-impulsivity (Baron, Badgio & Gaskins, 1986), a measure of cognitive style concerned with the relative preference for accuracy (reflection) versus speed (impulsivity).[4]

This kind of account 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 results discussed so far in this

[4]It is also possible that other factors could affect the observed correlations, such as different kinds of education or upbringing.

section, 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.

4.2 The role of emotion

The evidence on the role of emotion is also strong. Much of it concerns individual differences, some of these the result of brain damage. Damage to regions of the brain that involve emotional responses is correlated with more utilitarian judgments (e.g., Moretto, Làdavas, Mattioli & di Pellegrino, 2010, who find evidence that the effects are correlated with physiological manifestations of emotion). Psychopathy, a trait associated with immoral behavior, is also associated with blunted emotional reactions, and is also correlated with utilitarian judgments in the usual dilemmas, as are other similar traits (e.g., Bartels & Pizarro, 2011). Interestingly, the tendency to feel anger, unlike other emotions, is positively correlated with utilitarian responses (Choe & Min, 2011).

Several attempts have been made to manipulate emotion and show an effect on moral judgment. To my knowledge, none of these attempts has convincingly showed that such a direct effect of emotion on moral judgment can occur. I discuss some of the methodological problems later.

Notice that the difficulty of finding a causal effect of emotion on judgment causes problems for a currently popular theory of moral judgment (Haidt, 2001), which holds that moral judgments are mainly the result of system-1 reasoning evoked by an emotional response and that system-2 reasoning is almost always rationalization rather than reasoning that could cause a change of mind. Nor can this view be saved by removing the emotion, so long as it assumes that the system-1 response is almost always deontological. Such a proposal is inconsistent with the observation of large individual differences (e.g., Baron, Gürçay, Moore & Starcke, 2012). The evidence indicates that reasoning (as subjects perceive it) is involve in producing answers to moral dilemmas (Bucciarelli, Khemlani & Johnson-Laird, 2008).

The association of emotion with deontological responding is also suspect. This seems to be the result of the attention given to a few cases in the psychology literature, cases such as the fat man and other similar dilemmas in which the utilitarian response is joined with something emotionally disturbing. It is not at all clear that real world dilemmas are predominantly of this type, as opposed to the opposite type in which the utilitarian response is the one supported by emotion, particularly emotions such as empathy, while the deontological response is the result of rigid application of a rule, such as Kant's famous example in which he argues that it would be wrong to lie in order to save someone's life (Baron, 2011a). Very few people today would agree with Kant on this point, but people often apply rules of their religion or culture even when they must force themselves, against obvious human passions, to do so.[5] One example might be

[5]See Kahane, Wiech, Shackel, Farias, Savulescu, & Tracey, 2012, for a similar argument,

following a rule against abortion, even when it saves the mother's life and when the fetus would die anyway.

5 Methodological issues

Many of the controversies in the literature result from the use of different methods.

5.1 Judgment vs. behavior, should do vs. would do

Social psychologists and behavioral economists are interested in what people do. Many JDM researchers, by contrast, see the study of moral judgment as part of cognitive psychology, so we are interested in how people *think* about morality. This is also the interest of developmental psychologists in the tradition of Jean Piaget, Lawrence Kohlberg, Deanna Kuhn, and others.

Of course both approaches are worthwhile, but for somewhat different purposes. The study of behavior is most relevant for situations that involve conflicts of interest, especially conflicts between self and others or between self-interest and (legal or moral) duty. These include both good and bad behavior: volunteering to do good works, as well as breaking the rules or taking more than one's share.

5.2 What question to ask

When self-interest is not involved — as is approximately true in judgments about public policy, and many other judgments — it seems reasonable to assume that people behave consistently with their moral judgments. No "temptation" of the simple sort is involved. Kohlberg (e.g., Kohlberg, Levine & Hewer, 1983) even argued that, as Socrates said, he who knows the right does the right. Although that is surely an overstatement, it may close enough to the truth so that the study of moral judgment, divorced from behavior, is valuable. This is especially likely for judgments about public policy. Asking people to think about such judgments is not all that different from asking them what they think in an opinion poll, a referendum, or a general election.

Within studies of judgment, we still have choices, and it matters (Kahane & Shackel, 2010). Here are several ways to ask questions about a hypothetical scenario, where X is an actor in the scenario and A and B are options:

1. Should X do A? (yes or no, or some graded scale) (or, Should X do A or B?)

2. If you were X, should you do A? (or the scenario could be written in the second person)

3. Would X do A?

with some results.

4. Would you do A?
5. Is A permissible?
6. Is A forbidden?
7. Which is morally better, A or B?
8. Rate the (im)moralty of X.
9. Is A wrong?
10. Is A appropriate?

Question 3 is rarely used. It is asking for a prediction of other people's behavior. This is of interest when we want to study social norms (Bicchieri, 2006) or conventions.

Question 4 asks for a prediction of the subject's behavior. Usually people will predict that they will do what they think they should do, but discrepancies are found even when self-interest is not obviously involved. For example, Baron (1992) asked subjects about a dilemma in which an evil dictator asked them to shoot one person in order to prevent 5 others from being shot by the dictator. Subjects were more likely to say that they should shoot than that they would shoot, and the difference was correlated with an answer to the question about which option would lead to greater guilt feelings. Thus, avoidance of guilt feelings could act as a form of self-interest, creating a conflict of interest.

In general, I think that questions 1, 2, 7, and 8 are defensible in most studies. In theory, 1 and 2 should never differ, if moral judgments are universalizable in Hare's sense; this is what it means to universalize moral judgments. However, it is conceivable (but never tested to my knowledge) that question 2 would trigger potential guilt feelings and would lead subjects to look for reasons not to do what they would acknowledge that X should do in question 1. Thus, question 1 may be better for most purposes.

Question 8 is useful for experiments in which options being compared are separated, or in which we want to allow subjects to say that both options are (im)moral but one might be better (worse) than another. The advantage of questions 7 and 8 might be that they call attention to morality, as distinct from law or convention, when this is a potential problem. By contrast, questions about what is permissible, forbidden, or wrong (5, 6, 9, and 10) invite the subject to think about law, external rules, or conventions, as distinct from moral judgments.

Moreover, questions 5 and 6 assume that the subject has clear concepts of moral permissibility and moral prohibition (as distinct from analogous legal concepts), as well as related concepts of duty, obligation, and supererogation (going beyond the call of duty). Arguably, some moral theories do not naturally make such distinctions. For example, utilitarianism as advocated by Hare (1981) does not make them at the critical level, where one option is always morally best and this is overriding. Hare does explain, however, how the critical-intuitive

distinction can lead to such concepts in practice, by taking into account such factors as the normal capacity for self-sacrifice in the formulation of intuitive rules. For example, we might say that people have a duty to pay taxes (so it is morally forbidden not to pay taxes) but not to give so much to the poor as to become almost as poor themselves (even though this is the best option at the critical level), because it is a real constraint on behavior that people have a limited capacity to sacrifice their self-interest. Such convoluted justifications of the concept of duty are surely not natural to people who think like utilitarians. Such people seem to exist. When confronted with questions in an experiment, however, they may not be willing to think so deeply, and they may simply pull a legal concept, not a moral one, out of a hat in order to answer the question.

Likewise, questions 7–10 have an additional disadvantage. They ask for judgments about a person's behavior. Again, if utilitarianism is the implicit normative model, then the question should be clearly meaningful in utilitarian terms, and questions about judgment are somewhat unclear. They should be understood as referring to the goodness/badness of the behavior, but they could also refer to judgments about blame or punishment, which are separate decisions made by others. In utilitarian theory, blame and punishment should be determined not primarily by the badness of the behavior that is blamed or punished but rather by the future effects of the blame and punishment on other behavior.

5.3 Showing that judgments are not normative

I have emphasized the idea of comparing moral judgment to normative models, just as is done elsewhere in JDM with probability, decision making, and quantitative judgment. And I have assumed utilitarianism as an appropriate normative model for this purpose. As elsewhere in JDM, if we are going to say that some judgment is inconsistent with a normative model, we need to set up the experiment to make sure that this is true.

A special problem is that subjects often try to weasel out of moral dilemmas by changing the description of the situation. Perhaps the features they add are to be expected (Keys & Schwartz, 2007), as in the case of assuming guilt feelings after causing harm, despite preventing greater harm. In other cases, subject admit that they did not follow instructions. The instructions typically say to assume that the situations presented are real as stated, but subjects do not do this. For example, when presented with the dilemma about whether to shoot one person to save 5 from being shot by an evil dictator, people say that the dictator would shoot the five anyway, even though the scenario says that he will keep his promise not to shoot. Such responses often arise when subjects try to give a utilitarian rationalization of a deontological rule that they want to follow (e.g., don't shoot people), or they might arise just to avoid the problem of both options appearing to be wrong (shooting, vs. letting 5 people die).

Experimentally, it may be necessary to ask subjects explicitly whether they answered as if everything in the scenario were true, without adding anything. I have found that they willingly answer such questions, sometimes with apparent

relief at being able to explain why their answer might otherwise seem irrational.

Another, similar issue, is that people often perceive differences between cases or options in normatively relevant attributes that are typically correlated with the manipulated attribute, despite the experimenter's effort to remove the correlation. The clearest example is the role of intention in omission bias. Omissions are typically less intentional than acts. And intention is relevant to utilitarian moral judgment of the choices of others, at least insofar as these judgments are relevant to our choices about praise, blame, punishment, and reward (Spranca et al, 1991).

For example, the Ellen scenario used by Spranca et al. (1991, described at the beginning of this chapter), attempted to hold constant Ellen's intention to deceive the police, and its strength, in both the omission case and the (adjacent) action case in which she goes through with the lie. Here, Spranca et al. asked whether intention was the same in both cases, and we found that most subjects did think it was, so we could interpret any moral judgment they made about action being worse as non-normative, at least insofar as it could be explained by differences in intention. Although a few of our subjects did see a difference in intention here, other studies find more subjects showing such differences (e.g., Royzman & Baron, 2002). These cases must be eliminated in order to demonstrate omission bias. Other researchers do not eliminate them, and their results thus fail to show anything conclusively non-normative. It is of course possible that asking directly about intention is not the right question, as people do not interpret it in the same way that philosophers do. For example, it seems that judgments of intention are greater for immoral acts, other things being constant (Knobe, 2003).

This problem is avoided if the subject is not asked to make judgments about behavior but is, rather, asked about which of two options is morally better or should be chosen, as discussed in the last section.

5.4 Within vs. between subjects

The frequently-asked question of whether to compare conditions within or between subjects has particular significance in the study of moral judgment. The question concerns transparency. If we want to make it clear to subjects what conditions we are comparing, then the best way is to present them adjacently. We typically do this when we ask about choices. We often present all the options. (Sometimes researchers ask for responses to each option, without showing the others.) If we do not want to make the comparison transparent, then presenting conditions to different subjects is one way to do it. Another way is to use many different examples of each condition, and separate the ones we compare, in hopes that subjects will forget their response to their first response when they get to the second. (We can test whether this worked.) Still another way is to use a combined design, in which we use several cases, with two versions of each. Each subject gets only one version of each case but gets an equal number

of each condition in the study as a whole.[6]

When we ask for comparisons transparently, we are asking for the subject's moral principles. The contrast calls attention to some particular issue, such as the distinction between acts and omissions. When we ask opaquely we are combining this issue of the subject's principles (which will still affect the response) with the question of what features of a case are salient. For example, Ritov and Baron (2011) found differences in the determinants of responses as a function of whether cases were presented jointly or separately (as have many others, in studies not involving moral judgment). In particular, the probability of apprehension affected judgments of appropriate punishment (as utilitarian and economic theory imply that it should) only when the manipulation of probability was transparent. Some people thus saw the relevance of probability, but mostly when it was brought to their attention.

For the purpose of comparing moral judgments to normative theory, transparent judgments would seem to be the obvious methodological choice, as they provide the clearest measure of the principles that people endorse, but opaque designs might be useful for studying other issues.

5.5 Generality and individual differences

Studies of moral judgment typically present several scenarios (cases) to each of several subjects, ideally in an order that is randomized for each subject (so that order effects and scenario effects may be tested separately). This leads to two related questions. One is whether effects are general across subjects and cases, and the other is whether there are systematic and interesting differences.

Sometimes we are not interested in the differences, and we expect the effect of interest (e.g., omission bias) to be generally true for both subjects and cases. In these cases, we need to think first about whether our hypothesis is one-tailed or not. If it is, then, arguably, we need a fairly large number of cases but not so many subjects, and we need only a test across cases (Baron, 1975). If the effect exists in a few subjects, then it exists. We are not concerned about some effect in the opposite direction.[7]

In cases where our hypothesis predicts generality across subjects and items, we can take both into account by treating them as crossed random effects (Bates, 2005; Bates et al., 2011).

In other cases, we are interested in individual differences among subjects. We may even think that subjects show effects in different directions. For example, we may think that some people show omission bias and other show the opposite,

[6]The lmer() function in the lme4 package of R (Bates, 2005; Bates, Maechler, & Bolker, 2011), and other similar programs, have made it easy to analyze such data by fitting models in which subjects and cases are treated as crossed random effects, as discussed in the next section.

[7]In many cases, we may have reason to think that some effect will go in the opposite direction, because we have stacked the deck on that side. For example, in many studies of omission bias, we make the numbers favor action, but people still favor omission. But this does not imply that we should do a two-tailed test. If the numbers win, then the experiment failed to show what it was designed to show.

action bias. In such cases, it is difficult to generalize to a population of people unless we sample "the population" randomly. Most basic research is about human beings. If we are optimistic, most of the population is yet unborn.

We can, however, ask whether people show effects in opposite directions. With a sufficient number of cases given to each subject, we can test significance within each subject (Baron, 2010). We can ask whether the number of results that are significant at $p < .05$ are greater than the chance expectation of 5% (one tailed) in each direction. Baron (2010) discusses other approaches to finding subjects whose results go in each direction, and presents examples of such results, including results showing that action bias, the opposite of omission bias, is found in some subjects in some types of studies.

Another approach is to study individual differences among cases and subjects simultaneously. For example in omission-bias problems, such as variants of the trolley problem, we can suppose that cases differ in the extent to which they evoke utilitarian responses (action). We can think of this as the "difficulty" of the problems, in the sense that more difficult problems are less likely to elicit the correct (utilitarian) answer. For example, the decision to someone off a bridge to stop a trolley is more difficult than switching the trolley to a different track. We can also think of subjects as differing in "ability" to produce utilitarian responses.

Baron, Gürçay, Moore and Starcke (2012) suggested that such a situation is analogous to ability testing. A simple model of data from ability testing (Rasch, 1961) assigns to each subject an ability measure and to each test item a difficulty measure and then assumes that the probability of a correct response to the item is a logistic function of ability minus difficulty. When ability and difficulty are equal, the probability is .5. This simple model is only approximately true, but often close enough to be useful.[8] When this model was applied to several data sets, Baron et al. found that both ability and difficulty were smoothly distributed (i.e., unimodal).

5.6 Response times (RTs)

We also found that mean response times (RTs) for omission-bias-type scenarios were slowest when ability and difficulty were about equal. Thus, the slowest responses thus came at a different level of difficulty for different subjects, and slow responses could be interpreted as a sign that both options seemed equally (un)attractive.

In use of RT data, two issues must be addressed (among others). One is when to start the clock. Of interest is the time spent deliberating, more than the time spent reading, although the latter might be of interest too. In order to estimate the former, it may be necessary to present all background information, allow the subject to read it and then click for the rest of the information, which should be presented as briefly as possible, with the clock starting at this point. The rest of the information might, for example, include the number of people

[8] More flexible Rasch-type models are available.

harmed if an action is chosen and the means by which the action occurs.

The other issue is the result of the skewness of RT distributions, which implies that mean RTs will be excessively influenced by the longest times, thus increasing variability and reducing power for detecting effects. A standard way to deal with this problem is to take the logarithm of each time before doing any analysis. This transformation makes sense if we think that variation in RTs as the result of random perturbations, each of which increases or decreases the time by multiplication, that is, by changing the RT by a percentage of what it would otherwise be. The central limit theorem of statistics assumes that perturbations are additive, not multiplicative. Taking the logarithm makes the theorem fit better, under the stated assumption.

5.7 Emotion and affect

As noted, many theories of moral judgment assume a role for emotion or affect. These terms have different meanings in psychology. Affect is usually understood as referring to a good-bad dimension of evaluation. Thus, an affective response to a stimulus (such as a response option) might be a simple association between the stimulus and such concepts as "good" or "bad". Writers who postulate a role for affect in judgment seem to have this concept in mind (e.g., Slovic, Finucane, Peters & MacGregor, 2002).

Emotion, by contrast, is a mental state with a hedonic value, a typical eliciting stimulus, and a typical effect on behavior. Classic emotions are fear, anger, disgust, sadness, regret, elation, and mirth. Fear, for example, has a negative hedonic state, is typically elicited by some sort of threat or extreme surprise, and causes high arousal, freezing, and constriction of peripheral blood vessels. Emotions typically can be distinguished by their physiological manifestations, especially when facial musculature can be assessed. And they typically have an arousal component, which can be measured with skin resistance, heart rate, blood volume, etc. Unfortunately, even physiological measures of arousal (e.g., Moretto, et al., 2010) could be sensitive to emotions aroused by non-moral aspects of the problem, such as frustration at having to make a difficult decision, or to deployment of mental effort.

It is clear that the scenarios used in moral judgment research evoke affective responses, but little research has examined whether they evoke emotional responses. The fact that they are hypothetical need not in itself be a problem — people laugh and cry in response to fiction — but it isn't clear that they are sufficiently involving as they are usually presented. (It would be interesting to use real cases as stimuli, cases previously unknown to the subjects.)

The idea that emotion has a causal effect on moral judgment (e.g., Haidt, 2001) is difficult to test. Several lines of evidence are relevant, some of these reviewed earlier. Non-utilitarian responses to personal omission-bias scenarios are associated with psychopathy, which is itself associated with reduced responsiveness of some emotions, with other such measures of individual differences, and with brain damage that impairs emotional responsiveness. But these studies are all correlational. Whatever causes the reduced emotional responsiveness

may also affect the moral responses.

Experiments that manipulate emotion might be more convincing, but it is difficult to manipulate just emotion and nothing else that might affect the judgments. For example, most manipulations of disgust involve putting the subject in an unpleasant situation such as a smelly or messy room. This may evoke disgust, but it may also evoke moral condemnation of the experimenter. A possible solution, not tried to my knowledge, is to manipulate the intensity of emotion with epinephrine, as famously done by Schachter and Singer (1962).

Other studies use manipulations that supposedly reduce emotion by asking people to be dispassionate, but the wording of such instructions also encourages subjects to question their initial judgment cognitively. And even the studies that show results with these problems unsolved turn out to be difficult to replicate.

6 Conclusion: Why it matters

The study of moral judgment is important for many of the same reasons why we study judgments of any sort. A major reason is to improve the human condition. If it turns out that judgments are biased in some way relative to normative models, and if we can develop prescriptive approaches to reduce the biases or work around them, then we can improve matters. Although some people think that the discovery of biases is reason for pessimism, I think it is the opposite. If we cannot find ways to improve human judgments and decisions, then one path to improvement of the human condition is closed to us.

Moral judgments have special significance because they have externalities. Biases in moral judgments affect people other than the judges themselves. (Other judgments have this property, of course, such as expert judgments.) This happens in situations that involve conflicts of interest, particularly when people find ways to rationalize pursuit of their own interests at the expense of others. But it also happens in politics. Political opinions, and the sometimes violent actions that express them, are rich with moral judgments, even when no self-interest is involved. Successful efforts to improve the quality of moral judgments in politics may do a great deal of good.

References

Baron, J. (1975). Stimuli and subjects in one-tailed tests. *Bulletin of the Psychonomic Society, 6*, 608–610. http://www.sas.upenn.edu/~baron/onetail/.

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

Baron, J. (1993a). *Morality and rational choice*. Dordrecht: Kluwer (now Springer).

Baron, J. (1994). Nonconsequentialist decisions (with commentary and reply). *Behavioral and Brain Sciences, 17*, 1–42.

Baron, J. (1996). Norm-endorsement utilitarianism and the nature of utility. *Economics and Philosophy, 12*, 165–182.

Baron, J. (2003). Value analysis of political behavior — self-interested : moralistic :: altruistic : moral. *University of Pennsylvania Law Review, 151*, 1135–1167.

Baron, J. (2006). *Against bioethics.* Cambridge, MA: MIT Press.

Baron, J. (2008). *Thinking and deciding*, 4th edition. New York: Cambridge University Press

Baron, J. (2010). Looking at individual subjects in research on judgment and decision making (or anything). (Special issue on "Methodological concerns of the experimental behavioral researcher: Questions and answers," edited by S. Li). *Acta Psychologica Sinica, 42*, 1–11.

Baron, J. (2011a). Utilitarian emotions: Suggestions from introspection. *Emotion Review* (special issue on "Morality and emotion" edited by Joshua Greene) *3*, 286–287.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 3.* Hillsdale, NJ: Erlbaum.

Baron, J., Gürçay, B., Moore, A. B., & Starcke, K. (2012). Use of a Rasch model to predict response times to utilitarian moral dilemmas. *Synthese* (special issue on Psychological Models of (Ir)rationality and Decision Making, edited by C. Witteman & W. van der Hoek), *189*, Supplement 1, 107–117.

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

Baron, J. & Miller, J. G. (2000). Limiting the scope of moral obligation to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology, 31*, 705–727.

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., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*, 1–16.

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

Bates, D. (2005). Fitting linear mixed models in R. *R news, 5*, 27–30.

Bates, D., Maechler, M. & Bolker, B. (2011). lme4: Linear mixed-effects modegls using S4 classes. R package version 0.999375-42. http://CRAN.R-project.org/package=lme4

Bennett, J. (1981). Morality and consequences. In S. M. McMurrin (Ed.), *The Tanner Lectures on human values* (Vol. 2, pp. 45–116). Salt Lake City: University of Utah Press.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms.* New York: Cambridge University Press.

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision Making, 3*, 121-139.

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

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology, 18*, 255–297.

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.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 24–42.

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

Haidt, J. (2001). The emotional dog and its rational tale. *Psychological Review, 108*, 814–834.

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

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgment. *Mind and Language, 25*, 561–582.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cogntive and Affective Neuroscience, 7*, 392–402.

Keys, D. J. & Schwartz, B. (2007). "Leaky" rationality: How research on behavioral decision making challenges normative standards of rationality. *Perspectives on Psychological Science, 2*, 162–180.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*, 309–324.

Kohlberg, L., Levine, C. & Hewer, A. (1983). Moral stages: A current formulation and a response to critics. In J. A. Meacham(Ed.), *Contributions to Human Development* (Vol. 10). Basel: S. Karger.

Kuhse, H. (1987). *The sanctity of life doctrine in medicine: A critique.* Oxford: Oxford University Press.

Moretto, G., Làdavas, E., Mattioli, F., & di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience, 22*, 1888–1899.

Paxton, J. M., Ungar, L., & Greene, J. D. (2011). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*, pp. 321–334. Berkeley, California: University of California Press.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*, 263–277.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes, 79*, 79–94.

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.

Rudski, J. M., Osei, W., Jacobson, A. R., & Lynch, C. R. (2011). Would you rather be injured by lightning or a downed power line? Preference for natural hazards. *Judgment and Decision Making, 6*, 314–322.

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379–399.

Schleifer, M., Shultz, T. R., & Lefebvre-Pinard, M. (1983). Children's judgements of causality, responsibility and punishment in cases of harm due to omission. *British Journal of Developmental Psychology, 1*, 87–97.

Singer, P. (1993). *Practical ethics* (2nd ed.). Cambridge: Cambridge University Press.

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

Slovic, P., Finucane, M., Peters, E. & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman, (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 397–420. New York: Cambridge University Press.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105.

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

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

Tetlock, P. E., Lerner, J. & Peterson, R. (1996). Revising the value pluralism model: Incorporating social content and context postulates. In C. Seligman, J. Olson, & M. Zanna (Eds.), *The psychology of values: The Ontario symposium, Volume 8*. Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.