Decision theory for non-consequentialists

Wolfgang Schwarz
Draft, 2 August 2014

1 Decision problems

Consider a trolley problem. A runaway trolley is heading towards five people tied to the tracks. You can flip a switch that would redirect the trolley onto another track where it would run over a construction worker repairing the lights in a tunnel. What should you do?

Different moral theories give different answers. A consequentialist theory might say that you should flip the switch because five deaths are worse than one. A deontological theory might instead say that you shouldn't flip the switch because killing one person is worse than letting five die.

Now consider the following variation of the problem. A runaway trolley is heading towards five people tied to the tracks. You can flip a switch that would redirect the trolley to another track, where it *might* run over a construction worker scheduled to repair the lights in a tunnel.

This time, a crucial piece of information is missing. We don't know whether the construction worker is actually there and whether she would be harmed by redirecting the trolley. The missing information is especially important from the deontological perspective according to which you shouldn't flip the switch in the original problem. In the new problem, we don't know whether flipping the switch would amount to a killing or not. So even if we assume an uncompromising ban on killing, we can't say what you should do.

Is this a problem? One might say that moral theories don't need to give verdicts for arbitrary underspecified cases. Consider the following scenario. There are two buttons; which one should you press? Plausibly, that question can't be answered. We need more information about the buttons. Similarly, a deontologist might say that in the variant trolley problem, flipping the switch is obligatory *if the construction worker is not in the tunnel* and forbidden *if the construction worker is in the tunnel*, and that's all there is to be said. On this view, a moral theory only needs to provide an evaluation of fully specified decision situations, where all relevant facts are settled.

But this is an unsatisfactory position. In real life, we almost never have perfect information about all relevant facts. We don't know whether there are people on the tracks. We don't know how the economy will react to a stimulus plan. We don't know whether civilians would be harmed in a military operation. How should our actions be

guided by our moral theory, if our moral theory refuses to give a verdict until all factual uncertainty is resolved? But shouldn't endorsing a moral theory have some effect on one's choices? Otherwise what's the point of morality?

One might even argue that the verdicts familiar moral theories give for trolley problems and the like often presuppose that the agent *knows* that the relevant facts obtain. Consider another variant. This time, the second track is empty, and you have a further option: you can stop the trolley by throwing a large rock in its way. Doing so would cause significant damage to the trolley. A moral theory might thus reasonably say that it would be better to flip the switch and let the trolley come to a natural stop on the empty track. But suppose you don't know that the track is empty. Suppose all you know is that *one of the two tracks* is empty and the other has five people tied to it. Surely then you should throw the rock.

There might be an important sense of 'ought' or 'right' in which the right choice (what you ought to do) is in fact to flip the switch. On this conception, agents who always make the right choice (and who don't also happen to be omniscient) are agents who constantly take incredible moral risks and always come out lucky. Even if this is correct on some usage of 'right', it is clear that a conscientious moral agent should not aim at this peculiar ideal. A conscientious moral agent would throw the rock in our trolley problem. She would do so independently of whether the five people are in fact tied to the one track or the other. But this means that we can evaluate the options in a decision problem without fixing all relevant facts. We know what a morally conscientious agent would do, given only the information available to her. Our moral theory does not fall silent.

2 Moral theories

I should probably say a little more on what I mean by a "moral theory". A moral theory is a system of substantive moral principles, rules, guidelines etc. that determine an assignment of a moral status to possible choices in decision situations. The moral statue might be 'right' or 'wrong', 'obligatory', 'permissible', or 'forbidden', or it might be more fine-grained, involving degrees of rightness and wrongness.

The totality of our moral convictions is a moral theory. So is the totality of objective moral truths, if there is such a thing. Genuine moral disagreement is disagreement over moral theories.

I take for granted that there is more than one logically consistent moral theory. Even if it is objectively true that you should flip the switch in a our first trolley problem, this is not a truth of logic. A theory that says you shouldn't flip the switch may be false, but it is not incoherent.[1]

1 One view this rules out is that 'right', for example, simply *means* 'maximizes overall happiness', in the

A moral theory can be more or less comprehensive, depending on the range of (types of) decision situations it covers. A theory that deals exclusively with trolley problems would not be very comprehensive. A theory that only deals with situations in which the outcomes of all options are fully specified would also not be very comprehensive, although in a different way. The observations in the previous section suggest that our moral commitments are more comprehensive than that. We can evaluate the options in decision situations without knowing exactly what would happen if a given option were chosen. And that's not a coincidence: if we couldn't, our moral theory would be useless for guiding our actions.

I want to make no assumptions about the internal structure of moral theories. In principle, a moral theory could be an endless, gerrymandered list of verdicts about specific decision situations, without any unifying principles. It is clear, however, that our own moral commitments do not take this form. At the other extreme, a moral theory might base all its verdicts on a unifying principle – Kant's categorical imperative, say, or the principle that pleasure is good and pain bad. A moral theory might also address decision situations indirectly, *via* character traits that a morally ideal agent should possess. That alone would not yet constitute a moral theory in my sense, since it would give no verdicts about acts. It turns into a moral theory once we add a chapter on how the relevant character traits manifest themselves in an agent's choices.

The question I want to address is internal to moral theories. I want to know what credible moral theories should say about decisions with imperfect information. This is relevant, for example, when we consider uncertainty about morality. Suppose you are undecided between a consequentialist theory on which you should flip the switch in our original trolley case and a deontological theory on which you shouldn't. What should you do? Well, *according to the consequentialist theory* you should flip the switch, and *according to the deontological theory* you shouldn't. From the internal perspective of moral theories, moral uncertainty generally does not pose any interesting problem.[2]

3 Moral decision theory

I will proceed on the assumption that credible moral theories should provide an evaluation of the options in a significant range of decision situations with imperfect information. The

way in which 'vixen' means 'female fox'. Another view it excludes is that normative and descriptive facts are so thoroughly entangled that one can't type decision problems without already settling their moral evaluation. (I do not assume that the typing can be done in ordinary language, although I see no good reason that it can't.)

2 Admittedly, a moral theory might have special rules for cases of moral uncertainty. But note that (1) this would not help agents who are uncertain whether the relevant moral theory is true, and (2) we certainly don't want to say that whenever an agent is *certain* of some moral theory M, then she (morally) ought to act in accordance with M.

cases I will focus on are situations in which we are at least given a probability measure over the different possibilities.

For the sake of concreteness, I will assume that this measure stands for the agent's personal degrees of belief in the relevant situation. The relevant decision problems thus take the following form: what should an agent do if she has such-and-such options and such-and-such degrees of belief over relevant ways the world could be?

Other interpretations of the probability measure lead to other questions – to other typings of decision situations –, many of which are equally interesting and important. For example, we could let the probability measure stand for the degrees of belief that would be rational in light of the agent's evidence. Or we could let it stand for degrees of belief that would be rational in light of the evidence the agent ought to have. Most of what I will say easily carries over to these interpretations.[3]

One might want even more. One might want a comprehensive moral theory to say what an agent should do even if no probabilistic information is given. Personally, I don't think that this is possible. Nor is it motivated by the assumption that moral theories should be action-guiding. Wherever there is evidence, and wherever there are degrees of belief, there is probabilistic information. A more important class of situations are ones in which the relevant probabilities are not *sharp*, or ones where they fail to satisfy the axioms of probability. I set these cases aside because the issues I want to focus on already arise for the easier – albeit somewhat idealized – cases in which a sharp, coherent probability measure is available.

So let's return to the kind of question on which I want to focus. We want our moral theories to evaluate the options in decision situations that are given by a range of available options and a subjective probability measure over relevant ways the world might be. How could that be done? There is always the possibility of an endless, gerrymandered list. But suppose we want something more systematic.

A tempting idea is to adapt the formal framework of decision theory. In rough outline, the relevant part of decision theory goes as follows. We assume that the probability measure P is defined over a set of propositions, called *states (of the world)*, each of which determines an *outcome* for every available option. We also assume that there is a utility measure U, representing the agent's preferences, that assigns to each outcome a number. Let $O[A, S]$ denote the outcome brought about by option A in state S. If we hold fixed the option A, the utility function effectively maps each state S to the utility

3 We don't have to make a choice. It would be a mistake, I think, to debate whether "subjective moral oughts" are relative to the agent's degrees of belief or to the degrees of belief she ought to have. A more plausible story is that the moral *ought* is a "circumstantial modality" roughly in the sense of Kratzer [1991]. On this account, *ought*s are always relative to contingent circumstances: *given that you promised to call* (and other relevant facts about the world), you ought to call; *given that your friend was robbed* (etc.), you ought to help her; *given that your degrees of belief are such-and-such* (etc.), you ought to do Φ; *given that your evidence is such-and-such* (etc.), you ought not to do Φ.

$U(O[A, S])$ of choosing A in S. In the jargon of probability theory, this mapping is a "random variable". Its *expectation*, a.k.a the *expected utility of A*, is the average of the values, weighted by their probabilities. That is, if $S_1, S_2, S_3, \ldots$ are the relevant states, then the expected utility of an option A is given by

$$(*) \qquad EU(A) = \sum_i P(S_i)U(O[A, S_i]),$$

or a corresponding integral if the states are continuous. Decision theory then says that the agent should choose an option with maximal expected utility.

The obvious method to adapt this to the present tasks is to replace the agent's subjective utility function by a *moral utility function* that ranks outcomes not by the agent's personal preferences but by their degree of moral goodness. The rest can stay the same. Given a range of options and a suitable probability measure, (*) defines the *expected moral utility* for each option. And so we can let our moral theory say that the agent ought to choose an option with greatest expected moral utility.

This treatment of decision situations with imperfect information is widely endorsed by consequentialists. At least on the surface, it certainly seems to presuppose a broadly consequentilist account of morality. For one thing, it assumes that the moral evaluation of acts is derivative of the goodness attached to outcomes the act might bring about. Moreover, it presupposes that moral value is a "cardinal" quantity that can be meaningfully compared, aggregated and averaged across possible circumstances. Well-known formulations of decision theory further suggest that this quantity is agent-insensitive and time-insensitive in a way that doesn't allow us to capture the fact that agents stand in different moral relations to themselves, their family, and their contemporaries than to other people at other times. The focus on expected utility also assumes a questionable symmetry between good outcomes and bad outcomes. Finally, the decision-theoretic framework seems to generate a lot more obligations than many non-consequentialist would like: in most decision situations, the available options will have different expected moral utility; if the agent ought to choose an act with maximal expected utility, our lives will therefore tightly constrained by moral obligations. (There will be no room for superogatory acts.)

My aim in this paper is to defend the use of decision theory in non-consequentialist moral theories. I will argue that all the worries just raised can be answered. In fact, I will argue that moral decision theory, when spelled out carefully, is *better* suited for non-consequentialist theories than for strictly consequentialist theories.

As a first step, I need to explain what sort of decision theory I think we should use.

4 Decision theories

Decision theory is not a settled discipline. There are many different formulations, and even more interpretations of these formulations.

One point of divergence concerns the interpretation of the probability measure. In the tradition of von Neumann and Morgenstern [**?**], the probability P in decision problems is often understood as some kind of objective probability. The question then arises what an agent should do if the objective probabilities aren't known. By contrast, "Bayesian" decision theorists in the tradition of Ramsey [1931] and Savage [1954] take the probabilities to measure the agent's subjective degrees of belief. As I mentioned above, different interpretations are also possible for the application to moral theories, although I will generally follow the Bayesian interpretation of probabilities as subjective degrees of belief.[4]

Even more contested, and more important for the present topic, is the interpretation of the utility measure. A least three approaches should be distinguished.

The first, popular among psychologists and economists, assumes that the utility associated with an outcome is a function of the agent's wealth or well-being. This leads to the idea that decision theory – "rational choice theory", as it is then often called – is a theory of narrow-minded, self-interested homines economici.

On the second approach, which is probably dominant among working decision theorists, an agent's utility measure is derived from her preference order over a large space of possible acts. It can be shown that if this order satisfies certain qualitative conditions ("axioms"), then it can be represented by a utility measure over the outcomes in such a way the agent prefers an act X to an act Y just in case the expected utility of X is greater than that of Y. Various sets of axioms have been proposed. A characteristic example, from [Savage 1954], looks as follows.

(S) Suppose two acts A and A' lead to the very same outcomes unless some condition C obtains. Let B and B' be acts that coincide (in terms of their outcomes) with A and A' respectively whenever C obtains, and that coincide with one another whenever C does not obtain. Then the agent prefers A to A' iff she prefers B to B'.

Note that this axiom, like all the others, is purely structural. On the present approach, decision theory therefore no longer settles which features of outcomes should matter to an agent.

A third approach, popular mainly in theoretical philosophy, simply interprets the utility function as capturing the agent's degrees of desire, without assuming that they can be

4 It is important to keep this stipulation in mind, as it does not fit the use of 'probability' in ordinary English. On the Bayesian interpretation, the natural-sounding claim that an agent doesn't know that the probability of Heads is 1/2 has the curious meaning that the agent doesn't know that her own degree of belief in Heads is 1/2.

operationalized away in terms of preferences or choice behaviour. Decision theory is here understood as a theory of purely *instrumental* rationality, saying how rational agents should go about pursuing their goals or desires, without making any prescriptions or assumptions about the content of these goals or desires. In particular, it is not assumed that the agent only cares about local, ahistorical features of outcomes.

Consider the following well-known scenario, going back to [**?**]. A mother has a treat which she can either give to her son or to her daughter. Instead of directly handing it to either of them, she decides to toss a coin. Evidently the mother does not just care about herself. Nor does she only care about who eventually gets the treat, for then she wouldn't prefer tossing the coin over giving it outright to one of her children. The reason why she decides to toss the coin is plausibly that she cares about whether the treat is allocated by a fair procedure. That is a perfectly coherent desire.

To accommodate desires like these, the decision-theoretic "outcomes" have to be individuated very inclusively. We must distinguish between the outcome that the daughter gets the treat directly and the outcome that she gets the treat as the result of a coin toss. Following Jeffrey [1965], a common move is to identify the outcome $O[A, S]$ of an option A under state S with the totality of everything that would be the case if the agent chooses A in S. Among other things, the outcome therefore contains the act that brought it about, as well as the choice situation in which the act was chosen: one thing that would certainly be the case if you choose A in a given situation is that you will have chosen A in that situation.

This has far-reaching consequences for the technical formulation of decision theory. For one thing, on the present view the very same outcome can never be brought about by different options. Traditional decision-theoretic axioms such as (S) are therefore no longer applicable: one can never compare an agent's preferences between different acts that may lead to the very same outcomes. It might still be possible to derive interesting qualitative constraints on preference orders from the norm to maximize expected utility, but the constraints look very different from the traditional axioms. (See [Joyce 1999] for the details.)

It is easy to see why this last approach – which I will call *minimal decision theory* – has not been popular among economists and psychologists: it makes no interesting predictions at all about what rational agents will do. It does not say that if you are prepared to choose an apple over a banana, and a banana over a carrot, then you would choose the apple over the carrot. You might well have a desire for apple-if-the-alternative-is-banana, but no desire for apple-if-the-alternative-is-carrot. Similarly, minimal decision theory does not say that if you weren't willing to pay some amount for a concert ticket, then you should be willing to sell your ticket for that amount. In terms of your desires, situations in which you give something away need not be treated as equivalent to situations in which you don't get it in the first place. Nor does minimal decision theory make interesting

predictions about an agent's attitudes towards risk. Like the mother with the treat, you may sometimes prefer a gamble over sure outcomes, because you value something about the gamble. In other cases, you might think gambling would be frivolous and for that reason prefer a sure outcome.

All this is just what we want from a decision theory in our moral theories. We don't want the formal framework of decision theory to settle substantive moral questions. We want to allow for fairness as a moral value, and we want to allow for the possibility that *giving away X* has a different moral status than *not getting X in the first place*.

Minimal decision theory still comes in many flavours. (Some classical treatments are [Jeffrey 1965], [Gibbard and Harper 1978], [Lewis 1981], [Skyrms 1984], [Sobel 1986], and [Joyce 1999].) A lot of these details will not be important for us, but it will still be useful to get a broad sense of the issues, and to have a concrete proposal at hand.

5 A minimal decision theory

Here is a rough, and somewhat idiosyncratic, overview of Lewis's decision theory, as presented in [Lewis 1981] (also drawing on [Lewis 1979]).

According to Lewis, subjective utilities are defined over properties the agent might have at the time of decision – on a very inclusive understanding of "properties" on which just about any predicate corresponds to a property. Thus *having children* and *eating a muffin* are properties, but so are highly extrinsic and gerrymandered things like *either shovelling snow or eating a muffin 2000 years after Julius Caesar crossed the Rubicon*. Call the conjunction of all properties instantiated by an agent at a time the agent's *total profile* at that time. Only few of these properties will typically matter to the agent. If two profiles differ only with respect to the arrangement of matter in distant galaxies, or with respect to what Julius Caesar had for breakfast before he crossed the Rubicon, you probably won't care which of them is realized. Let's say that a conjunction of properties is a *complete profile* with respect to a utility function if all total profiles that entail the conjunction have equal utility.

So assume we have a utility function over the space of all properties. Given a range of options, the states in a well-defined decision problem should now specify, for each option, all utility-relevant properties the agent would come to instantiate if she were to choose that option. In other words, a state should specify for each option a complete profile that would be realized if the option were chosen.

Here we face one of the problems to which there is no agreed upon solution. Suppose one of your options is to toss a fair coin. You'll get a reward iff the coin lands heads. Among the things you care about is whether you will get the reward. Since you know that the coin is fair, you know that *if you were to toss the coin, it might land tails*. According to Lewis, this is incompatible with the assumption that *if you were to toss the*

coin, it would land heads and thus also with *if you were to toss the coin, you would get the reward*. However, we still want to take into account the possibility that you'll get the reward. To do so, Lewis suggests that in cases like this, the states may only specify the objective chances that would pertain to complete profiles. Unfortunately, this doesn't cover cases in which a choice would not even bring about a particular assignment of chance. It also gives intuitively wrong verdicts for cases in which the agent's subjective probabilities come apart from the chances. So we'll have to bracket these cases if we want to use Lewis's theory.

Another famous point of divergence concerns the computation of expected utilities. Lewis endorses the classical formula (*). Given his definition of states, this makes his decision theory a so-called *causal decision theory*, in contrast to Jeffrey's [1965] *evidential decision theory*. Jeffrey replaces the unconditional probability $P(S_i)$ of states with the conditional probability $P(S_i/A)$ of states given the relevant option:

$$(**) \qquad EU(A) = \sum_i P(S_i/A)U(O[A, S_i]).$$

Under plausible assumptions about the utility function, (**) turns out to be invariant under different partitionings of the states. Thus Jeffrey does not face any difficulties about defining an adequate partition of states. On the flip side, Jeffrey's theory is widely (though by no means universally) thought to give wrong recommendations in cases like Newcomb's problem where an option is evidence for something good without being causally relevant to bringing about the good (see e.g. [Joyce 1999: sec. 5.1]).

There are further choice points. For example, Lewis's theory seems to give implausible verdicts in cases where a given decision would provide the agent with information relevant to the expected utilities (see e.g. [Egan 2007]). In response, one might supplement the basic theory by a "ratifiability" constraint according to which an option should only be chosen if it maximizes expected utility conditional on it being chosen (see e.g. [Harper 1984], but see also the apparent counterexample in [Skyrms 1984: 84–86]). Alternatively, one might hope that a dynamical model of deliberation could take care of the problems (see e.g. [Skyrms 1990], [Arntzenius 2008]). Further questions arise, for example, in cases where expected utilities go undefined (see e.g. [Easwaran 2008]). Again, there are many possibilities, and none of them is universally accepted.

In practice, almost all formulations of minimal decision theory agree about almost all decision problems we encounter in everyday life. So we don't have to wait until all controversies in decision theory have been settled before applying decision theory to ethics. Of course, we will eventually run into the same problems about what to say, for example, in moral Newcomb problems, or when moral expected utilities go undefined. For reasons that will become clear, I think the right answers will coincide with the right answers in the theory of instrumental rationality, so at least we won't have to start from scratch.

Using something like Lewis's minimal decision theory immediately resolves a lot of the worries often raised about the use of decision theory in non-consequentialist moral theories. In Lewis's framework, the moral status of acts is not derived from the goodness of outcomes. (We don't have to "consequentialize" our moral theory before applying decision theory.) Moral utilities can pertain directly and primarily to "properties" such as *eating a muffin*, *intentionally killing an innocent*, or *resisting a strong desire to gossip*. Minimal decision theory also allows us to make special provisions for risk, or to give special weight to worst outcomes. Even more obviously, it automatically gives a special place to the agent and the time of her decision. We can easily say that it is worse for an agent to *put a burden on others* than to *put a burden on herself*. After all, these are very different properties.

There is still work to be done. To apply decision theory, we need a numerical measure of moral utility. Where does that come from? Secondly, we still need to defend the idea that what matters in cases of uncertainty is the expected moral utility of the available options. Why not follow some other decision rule? Thirdly, we still need to make room for the intuition that not any difference in expected moral utility gives rise to moral obligations.

6 Moral utility

Standard decision theory requires a "cardinal" measure of utilities in the sense that one can meaningfully speak about how much better or worse (right or wrong) one act is compared to another. More precisely, it should be possible to say of four acts A, B, C, D that A is as much better than B as C is better than D.

To illustrate why this is needed, suppose we have only three kinds of moral status: obligatory, permissible, forbidden. Our moral utility function U has to convert them into numbers, so that they can be multiplied and added, as required by (*). Moreover, we should convert them into numbers in such a way that higher numbers represent better moral status, since we want to recommend choices with greater expected utility. One way to do this would be to encode 'obligatory' by the number 1, 'permissible' by 0, and 'forbidden' by -1. But we could also use the numbers $\langle 3, 2, 1\rangle$, or $\langle 1000, 10, 0\rangle$, or all kinds of others. The problem is that our choice of numerical representation will now make a difference to the recommendations of decision theory. So if we want to give decision-theoretic recommendations, we must narrow down the legitimate numerical representations. In standard formulations of decision theory, the remaining representations must all be positive linear transformations of one another.

In my view, this is itself a good argument why moral theories should make fine-grained, "cardinal" evaluations of profiles. The argument is that otherwise it becomes hard to give systematic verdicts about cases with imperfect information.

Take an example. You are on a boat and notice what seems to be a white shark approaching a group of surfers. Fortunately, you have a harpoon. Let's say you are 80 percent certain that the fish is a dangerous white shark, and give 20 percent credence to the possibility that it is a harmless, but equally endangered species of shark. What should you do? Here are the most relevant things that might happen, depending on your choice:

(1) You kill a white shark that would otherwise have attacked the group of surfers.
(2) You kill a harmless and endangered fish.
(3) You let a white shark attack the group of surfers.
(4) You let a harmless fish approach the group of surfers.

To simplify, let's say that shooting would come with an 80 percent probability of (1) and a 20 percent probability of (2). Not shooting would come with an 80 percent probability of (3) and a 20 percent probability of (4). Surely all this is relevant to the evaluation of your options. In particular, what makes your decision problem non-trivial is that (1) and (4) are much better than (2) and (3). It also matters that (3) is worse than (2), although both are deplorable. Moreover, it matters that (3) is *a lot* worse than (2), whereas (1) is not a lot worse than (4), by comparison with (3) and (2). If our moral theory doesn't allow us to make these comparative verdicts, it is hard to see how it could give sensible advice about what you should do. But if have verdicts like these, we can meaningfully apply decision theory.[5]

There are also independent reasons to think that our moral theories should provide fine-grained, "cardinal" evaluations of profiles. I mention three.

First, the features of acts that matter to morality are almost never clear-cut, and often come in degrees. Imagine a sorites sequence leading from a clear case of murder to a clear case of innocent behaviour. It is hard to believe that at some point, a minute difference between two cases makes an enormous moral difference – that one act in the sequence is forbidden and the next one permissible. Or consider all the morally relevant acts that come in degrees. If helping people is right and harming people wrong, then surely helping more is better than helping a little and harming more is worse than harming a little.

Second, moral theories should somehow be connected to our practice of blame and praise, punishments and rewards; and these come in "cardinal" units. We tend to criticise people less for cheating on their taxes than we criticise people for sexual assault. It is implausible that this does not track any moral difference between their acts.

5 The moral status of ordinary properties is often sensitive to probabilities. For example, one might say that letting the shark attack the surfers is even worse if one of the surfers is about to discover a cure for cancer. A *pure* moral theory that stays neutral on contingent factual hypotheses (as it must if it is to give recommendations for arbitrary decision problems) should therefore focus on complete property profiles. This blocks the usual decision-theoretic route of getting cardinal utilities from ordinal utilities.

Third, we often need to balance moral considerations against other moral and non-moral considerations. Most of us think that extra-ordinary circumstances can make it permissible to lie or even kill an innocent if that is the only way to prevent a horrendous tragedy. Moreover, we think that people's obligations are sensitive to the amount of sacrifice fulfilling the obligations would require. Some obligations justify more sacrifice than others. Either way of balancing would be impossible if we didn't have approximately "cardinal" moral values.

You may agree with all that and still feel uneasy about putting numbers on the possible profiles in a decision problem. I feel the same. The use of numerical utility measures is mathematically convenient, but it obscures the fact that what is measured has a quite different structure than the numbers. A clear indication of this is that standard utilities are unique only up to positive linear transformation. Without an arbitrary choice of zero and unit, it is strictly meaningless to say that the utility of an act A is 12, or that A has twice as much utility as B. Moreover, numerical measures seem to imply an implausible requirement of complete commensurability of all profiles. If all profiles are assigned a numerical utility, then it looks like any profile A must be either better than another profile B, or worse than B, or equally good. But we may well want to allow for incommensurable profiles. Formally, this is easy to accommodate in the kind of decision theory I have outlined (see e.g. [Joyce 1999]), but it makes precise numerical representations of utility even more arbitrary and artificial. One might also raise objections about the Archimedian structure of real numbers, and the fact that real numbers are never "infinitely" greater or less than one another, which makes it hard to capture, say, lexical orderings. These are all fair concerns. They reveal a true weakness in popular formulations of decision theory. But they apply not only to decision theory in ethics but also to decision theory as a theory of practical or instrumental rationality.

7 Expectations

Moral decision theory says that an agent should maximize expected moral utility. Why should we accept this?

Here is one argument. Assume some form of minimal decision theory is in fact the correct theory of instrumental rationality – that is, the correct theory of how to rationally pursue some goals in the face of limited information. If we can think of our basic moral theory as putting forward certain goals – not to kill innocents, etc. – then the corresponding moral decision theory says how one should rationally pursue these moral goals in the face of limited information. Arguably, this is exactly what we are looking for when we want to extend our moral theory to cases of imperfect information.

Here is another, related argument. Morality should not require practical irrationality. Imagine an agent who cares about nothing but doing what's morally right. It would be

odd to think that such an agent would have to violate the constraints of minimal decision theory.

The most important test for the expectation rule is whether it yields plausible verdicts in concrete applications. For non-consequentialists, it is here crucial that we use minimal decision theory, which avoids the agent-neutrality, risk-neutrality and history-neutrality (among other things) of classical decision theories. Recall that minimal decision theory by itself makes essentially no predictions about how agents will behave. Conversely, this means that just about any pattern of behaviour you can think of is compatible with minimal decision theory – including any pattern of behaviour you might deem morally ideal.

One might worry that if minimal decision theory makes no predictions about choice behaviour, it will not be useful in a moral theory that is supposed to put constraints on what people should do. However, the substantive constraints are not meant to come from the decision rule. They come from the underlying moral utility function. The power of the expectation rule is that it systematically turns a utility measure over complete profiles into recommendations for almost any decision problem an agent could face.

This systematicity is itself a great advantage. The space of possible decision problems is enormous. There are infinitely many ways in which you could be uncertain about the number of people on the tracks in a trolley problem; there are many more possibilities if we also take into account who these people might be; you may be further uncertain whether you face a trolley problem in the first place; you may be uncertain whether the switch is broken, whether the information you were given is correct, and so on. None of these cases is absurdly far-fetched or unrealistic. If we want our moral theory to give a verdict for a substantive range of decision problems people actually encounter, and if we don't want it to do that by giving an infinite list of *ad hoc* decision rules for individual problems, we need a unified set of principles.

To be sure, there are systematic rivals to the expectation rule. One is the *maximin* rule, which says that one should choose an option with the best worst-case profile. Another is the *shortcut* rule, which says that one should choose an option for which it is likely (say, probability $> 1/2$) that it yields the greatest utility. But these rules give horrendous recommendations for a large range of problems. Suppose you have a vaccine against a very crippling but non-fatal disease that is rampant in a certain community. The vaccine is very safe: in millions of applications it has never had any undesirable side-effects. Your credence that someone might die as a consequence of a vaccination is therefore very low, although (reasonably) not quite zero. Maximin then suggests that you should not vaccinate the members of the community. For the shortcut rule, vary the case so that (1) the "disease" only produces a minor temporary inconvenience in affected patients, and (2) the vaccine is extremely dangerous, leading to torturous deaths in 30 percent of all cases. For each member of the community, the shortcut rule then recommends the

vaccination.

What these rules get wrong is that they don't adequately balance probabilities and moral utilities. It is wrong to look only at the worst possible case, irrespective of its probability. It is equally wrong to look only at the most likely cases and ignore all risks. We should consider all things that might happen, and we should take them into account in proportion to their probability. This is just what the expectation rule says.

Non-consequentialists sometimes offer rules for how to deal with specific cases of imperfect information or risk. Jackson and Smith [**?**] discuss the rule for magistrates to convict people only if it is more than 95 percent certain ("beyond reasonable doubt") that they are guilty. One could think of similar rules for combatants in a war: to only shoot at someone if it is more than 95 percent certain that she is an enemy combatant. As Jackson and Smith show, such rules tend to get all sorts of edge cases wrong, and don't generalize well to cases where your options include convicting or killing several people at once. Such rules are in principle *compatible* with moral decision theory. They can be validated by the general trick of assigning low utility to profiles that involve violations of the rule. But the problems highlighted by Jackson and Smith show that this is probably not what we should do. The better choice is to assign suitably low moral utility to profiles that involve accidental killings of civilians (or unintended convictions of innocents) and let the general principles of decision theory decide how to handle individual cases with imperfect information.

There are also long-run arguments in favour of the expectation rule. Imagine a certain decision problem is repeated indefinitely. If each time you maximize expected moral utility, then by the strong law of large numbers you can be rationally certain that you maximize the overall *actual* moral utility. Similarly, if everyone followed the advice to maximize expected moral utility, then the laws of large numbers imply that people would on average choose morally better acts than if they followed any other rule.

Finally, we could try to adapt the classical decision-theoretic argument for the expectation rule, which tries to show that the rule follows from independently plausible conditions on an agent's preference order. In our context, the task is simplified by the fact that we can take the probability measure and the utility ranking over complete profiles as given. Nevertheless, we must be careful not to impose substantive conditions on preference orders. The Bolker/Jeffrey axioms are often regarded as acceptable even for minimal decision theories. The "preference relation" $>$ constrained by the axioms would be our relation of moral betterness between property profiles, but extended to arbitrary unspecific properties. I will not go through all the axioms; a characteristic example is the "averaging" axiom

(B) If two properties X and Y are incompatible and $X > Y$, then $X > (X \vee Y) > Y$.

The idea is that an unspecific property $X \vee Y$ cannot be better or worse than both of

the more specific properties X and Y. For example, *killing an innocent* cannot be worse than *killing an innocent man* and also worse than *killing an innocent woman*.[6] Assuming that an option is itself a property – perhaps a fairly gerrymandered property such as *either killing a white shark who would otherwise have attacked the surfers or killing a harmless end endangered type of shark* – it is then natural to suggest that an agent should choose the *best* option, by the extended betterness ranking. It turns out that this is equivalent to saying that the agent should choose the option with maximal expected utility as computed by Jeffrey's evidential decision theory. [Joyce 1999] discusses how the story needs to be adjusted to yield a causal decision theory.

8 Decision theory for consequentialists

I have given a number of arguments in defence of the expectation rule. Personally, I think they are good arguments, and so I think the expectation rule should play a role in credible moral theories. But what becomes of these arguments if we assume a consequentialist perspective?

Consequentialism paints an attractively simple picture of morality. On this picture, morality is concerned with a certain objective good – happiness, preference satisfaction, absence of suffering, or a combination of things like these. The basic moral imperative is to promote this good. All other moral principles should be derived from this. And here lies the problem: None of the arguments in the previous section showed that the expectation rule follows from the principle to promote the good. A strict consequentialist should therefore reject all of my arguments as irrelevant. But then what is *her* argument for the expectation rule?[7]

Of course, she might cheat. One way to cheat would be to replace pleasure etc. as the objective good by the sole objective good of maximizing expected good*, where good* is pleasure etc. That would clearly go against the spirit of consequentialism. A more subtle way to cheat would be to *define* 'promoting the good' as 'maximizing expected good'. But why should this be the right definition? In other words, isn't there still a question why we should "promote" the good rather than "promote*" the good, where 'promote*' is defined, say, by the shortcut rule? Consider an agent who always chooses on the basis of what is likely to maximize good. Or consider an agent who always chooses the option that actually maximizes good (by luck). Surely these agents also "promote the good", in some reasonable sense. What is it that makes their choices objectionable? The question does not seem empty. Yet the answers I gave in the previous section all

6 Unspecificity must not be conflated with uncertainty. It is compatible with (B) that *killing an innocent whose gender is unknown* is worse than *killing an innocent known to be male* and also worse than *killing an innocent known to be female*.

7 This problem was brought to my attention by Alexander Pruss.

seem quite irrelevant from a consequentialist perspective. Granted, the shortcut rule sometimes recommends actions that are intuitively abhorrent. But how does that show that it isn't the right rule for a given occasion? Maximizing expected utility performs well in the long run. – Again, how is this relevant for an *act* consequentialist? Morality should not require agents to be practically irrational. – How does that follow from the principle that one should promote the good?

The problem gets even worse if we remember that speaking of "the expectation rule" hides the fact that the precise formulation of minimal decision theory is a subtle and difficult matter, involving a lot of choice points. For example, I believe that moral decision theory should be causal rather than evidential, because evidential theories would give intuitively wrong recommendations in Newcomb-type problems. Others would disagree and argue that, on the contrary, evidential decision theory is preferable in moral theories (see e.g. [**?**]). As a non-consequentialist, my firm judgement about Newcomb cases is enough to prefer moral theories that embed the principles of causal decision theory. On what basis could a consequentialist say that 'promote' has to be read causally rather than evidentially? Similarly, I think causal decision theory needs some form of ratifiability constraint to avoid the implausible consequences discussed e.g. in [Egan 2007]. Again, how could these constraints be motivated directly from the goal that one should promote the good?

In my view, the best move for consequentialists is to relax their position and admit that the decision-theoretic part of their moral theory is a non-trivial addition, independently judged on the grounds of simplicity and systematicity, intuitive credibility, long-run performance, etc. The consequentialist moral theory then has two parts: an assignment of moral utility to full profiles based on the amount of good realized in these profiles, and a particular decision-theoretic rule to guide actions. There is nothing wrong with such a theory, except that it delivers verdicts that we non-consequentialists deem clearly wrong. But one should not pretend that the whole theory is somehow derived from the simple idea that pleasure (or welfare or whatever) is good and pain (or whatever) is bad.

9 Expectations and obligations

One problem is left from those I mentioned in section 3: the problem of too many obligations. If we are always morally obligated to choose an option with greatest expected (moral) utility, it seems that morality will constrain our lives much more than we'd like to think. There are different ways to respond to this – apart from accepting the consequence. One way is to attach moral utility to pursuing one's own goals and desires. Another, not necessarily exclusive, way is to refine the connection between expected utility and obligation.

Expected utilities provide a moral ranking of an agent's options. This ranking allows

us to say that option A would be better, from a moral perspective, than option B, that both of them would be much better than C, etc. I don't find it objectionable that our moral theory should provide such rankings. But it is a further question whether one is always morally obligated to choose the (morally) best option. One could instead say, for example, that one is obligated to choose an option only if it is *a lot* better than the alternatives (more generally, that one ought to choose one of several options if all of them are a lot better than all the others). As long as the ultimate verdicts are determined by the expected (moral) utilities of the options, any such account still vindicates the use of decision theory.

An interesting possibility is to not only compare the expected utility of an option relative to other options, but relative to some fixed points in the utility scale. For example, if we say that all options with an expected utility below a certain point are wrong, there can be tragic scenarios in which all options are wrong.

In fact, it is not essential to my proposal that the taxonomy of right and wrong, obligatory and forbidden, even applies to the options in a decision problem. One might have a conception of morality in which all true *ought*s are "objective oughts". There is still a question of what a morally conscientious agent should do if she lacks information to determine what she objectively ought to do. Decision theory can answer this question.

10 Surprising expectations

In his "Rule utilitarianism, rights, obligations and the theory of rational behavior" [1980], Harsanyi argues that rule utilitarianism leads to better outcomes than act utilitarianism, because rule utilitarians would contribute to the common good even if it their individual contribution is unlikely to make a difference. Here is one of his examples.

> 1000 voters have to decide the fate of a socially very desirable policy measure M. All of them favor the measure. Yet it will pass only if all 1000 voters actually come to the polls and vote for it. But voting entails some minor costs in terms of time and inconvenience. The voters cannot communicate and cannot find out how many other voters actually voted or will vote.
>
> Under these assumptions, if the voters are act utilitarians then each voter will vote only if he is reasonably sure that all other 999 voters will vote. Therefore, if even one voter doubts that all other voters will vote then he will stay home and the measure will fail. Thus, defeat of the measure will be a fairly likely outcome.

The reasoning looks plausible, but let's work through the case. Let's say that each member of the community would lose 1 util by voting. If everyone votes, this means

that the group has lost 1000 utils in total. To get an interesting social dilemma (or an argument for rule utilitarianism), we want the state in which everyone votes to be better than the state in which everyone stays at home. So the net utility of the "very desirable measure" M must exceed 1000 utils. Let's say it is 2000 utils. The decision matrix for an individual act-utilitarian voter then looks as follows.

	...	**998 others vote**	**999 others vote**
vote	...	-999	1000
don't vote	...	-998	-999

If all the others vote, the utility of voting exceeds the utility of not voting by 1999. If not all others vote, the utility of not voting exceeds that of voting by 1. As a result, voting comes out as having greatest expected utility even if the probability of all others voting is as low as 0.0005. A voter does not have to be "reasonably sure", as Harsanyi claims, that all the others will vote. Only if she is extremely confident that some of the others will stay at home does act-utilitarianism advise her to abstain.

Admittedly, in real life it may be reasonable to assign a probability of less than 0.0005 to the assumption that everyone else in a large group will do their share. But in real life we also rarely need absolutely everyone to do their share in order to reach a desirable outcome. Moreover, if we may be confident that at least one person will not show up, this is typically because it is reasonable to expect that at least one person isn't motivated or has forgotten or is unable to come. But if there is substantial chance of such disturbances, then groups of (act-utilitarian) non-voters actually perform better, in the long run, than groups of (rule-utiliarian) voters who often waste almost everyone's efforts.

So Harsanyi's example doesn't work. Nor do his other examples. I mention this not to pick on Harsanyi or to defend act utilitarianism, but to highlight two important facts. First, estimating expectations is hard – even for Nobel price winning decision theorists. Second, in many cases where an act is almost certainly pointless, the expected difference it makes is large enough to justify its choice.

This is one of the areas where decision theory can be of real service to our moral thinking. Without the formal framework of decision theory, it is tempting to fall back on stupid heuristics, such as only looking at the most probable states. No doubt the (by far) most probable state is that your vote will make no difference. Likewise, reducing your carbon-offprint won't affect climate change, donating to cancer research won't affect whether new cures will be found, going vegetarian will not prevent the rise of antibiotic resistant bacteria. And so on. But there is no credible moral theory on which this implies that you shouldn't go ahead and do your part anyway. Even if our moral theory cares about nothing but difference-making (as it shouldn't), what matters is not what difference your act would *most likely* make. That is the indefensible "shortcut" rule. What matters is the *expectation* of the difference. If the stakes are high, even states with negligible probability can drastically alter the expectation.

11 Conclusion

Moral theorists have often neglected decision problems with imperfect information. There is a lot of work on trolley problems, but very little work on trolley problems with imperfect information. In a way, the proposal I have made vindicates this negligence. Once we have specified the moral status of complete acts, the framework of decision theory will automatically cover all cases with imperfect information. Substantive moral theorizing can restrict itself to judgements about how various properties contribute to an act's "moral utility".

These properties can include aspects of the agent's epistemic state. It is wrong to throw a bottle out of the window if one doesn't know whether it might hit somebody, even if in fact no-one is harmed. The specification of moral utilities is "objective" in the sense that it considers all features of the relevant acts, not only features of which the agent is aware. But it is not "objective" in the (dubious) sense that it ignores the agent's epistemic state, or in the sense that it imagines the agent to be omniscient.

References

Frank Arntzenius [2008]: "No Regrets, or: Edith Piaf Revamps Decision Theory". *Erkenntnis*, 68: 277–297

Kenny Easwaran [2008]: "Strong and weak expectations". *Mind*, 117: 633–641

Andy Egan [2007]: "Some Counterexamples to Causal Decision Theory". *Philosophical Review*, 116: 93–114

Allan Gibbard and William Harper [1978]: "Counterfactuals and Two Kinds of Expected Utility". In C.A. Hooker, J.J. Leach and E.F. McClennen (Eds.) *Foundations and Applications of Decision Theory,* Dordrecht: D. Reidel, 125–162

William Harper [1984]: "Ratifiability and Causal Decision Theory: Comments on Eells and Seidenfeld". *PSA*, 2: 213–228

John C Harsanyi [1980]: "Rule utilitarianism, rights, obligations and the theory of rational behavior". *Theory and Decision*, 12(2): 115–133

Richard Jeffrey [1965]: *The Logic of Decision.* New York: McGraw-Hill

James Joyce [1999]: *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press

Angelika Kratzer [1991]: "Modality". In A. von Stechow and D. Wunderlich (Eds.) *Semantik Handbuch,* Berlin: de Gruyter, 639–650

David Lewis [1979]: "Attitudes *De Dicto* and *De Se*". *The Philosophical Review*, 88: 513–543. Reprinted in Lewis's *Philosophical Papers*, Vol. 1, 1983.

— [1981]: "Causal Decision Theory". *Australasian Journal of Philosophy*, 59: 5–30

Frank Ramsey [1931]: "Truth and Probability". In R.B. Braithwaite (Ed.) *Foundations of Mathematics and other Essays,* London: Routledge & P. Kegan

Leonard Savage [1954]: *The Foundations of Statistics*. New York. Wiley

Brian Skyrms [1984]: *Pragmatics and Empiricism*. Yale: Yale University Press

— [1990]: *The Dynamics of Rational Deliberation*. Cambridge (Mass.): Harvard University Press

Jordan Howard Sobel [1986]: "Notes on decision theory: Old wine in new bottles". *Australasian Journal of Philosophy*, 64: 407–437. Reprinted with revisions in [Sobel 1994: 141–173]

— [1994]: *Taking Chances*. Cambridge: Cambridge University Press