Decision theory for non-consequentialists*

Wolfgang Schwarz
Draft, 03 December 2014

1 Introduction

Moral philosophers have said a lot about what acts would be right or wrong in decision problems where all relevant facts are specified. Any such account of "objective" moral status should arguably be accompanied by an account of "subjective" moral status: about what would be right or wrong given the limited information available to the agent. Here consequentialists have often appealed to the framework of decision theory. If the objective moral status of an act is measured by the net amount of good the act brings about, then it is tempting to think that an act's subjective moral status is measured by the *expectation* of the good it might bring about – that is, by the weighted average of the amount of good the act might bring about, weighted by the respective probabilities.

Natural though it may appear, the decision-theoretic approach is not without problems. How can we quantify "amounts of good"? Does subjective moral status really go by the expectation of good – why not factor in, for example, the risks incurred by a choice? What if an agent doesn't know the probabilities, or lacks the cognitive capacity to compute the expectations? From a non-consequentialist perspective, the approach may seem even less appealing. How can we use decision theory if the rightness of acts is not determined by the goodness of outcomes? How could we accommodate agent-relative values, or supererogatory options?

I will argue that all these problems can be answered. Properly construed, the decision-theoretic approach is compatible with a very wide range of moral views. In fact, in some respects the approach is easier to justify in a non-consequentialist setting.

2 Moral decision problems

Consider a trolley problem. A runaway trolley is heading towards three people tied to the tracks. You can flip a switch that would redirect the trolley onto another track where

* Ancestors of this paper were presented at the ANU, the University of Leeds and the University of Bielefeld. I'm indebted to the audiences on these occasions, as well as Christian Barry, Holly Lawford-Smith, and especially Seth Lazar for very helpful comments.

it would run over a construction worker repairing the lights in a tunnel. What should you do? Different moral theories give different answers. Some say you should flip the switch because three deaths are worse than one. Others say you shouldn't flip the switch, perhaps because killing one person is worse than letting three die.

Now consider the following variation. A runaway trolley is heading towards three 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. What should you do? 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 if you redirected the trolley. The missing information is especially important from the perspective of theories 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? Moral theories don't need to give verdicts for arbitrary underspecified cases. "There are two buttons; which one should you press?" – That question can't be answered: we need more information about the buttons. Similarly, one might claim that in the second 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 that view, a moral theory only needs to provide an evaluation of fully specified decision problems, 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? And shouldn't endorsing a moral theory have some effect on one's choices – otherwise what's the point of morality? Parallel problems arise for our role not as agents but as advisers or judges. Suppose you ask me what you should do, perhaps because you know that I have more (but still incomplete) information. Shouldn't the moral theory I endorse manifest itself in my advice?

Discussions on this issue sometimes turn into debates about semantics. There is a sense of 'right' and 'ought' – often called *objective* – on which the right choice (what you ought to do) is the choice that is right in the light of all the facts, known and unknown. But there also seems to be a *subjective* sense on which an act is right if it is right relative to some contextually relevant information. For my topic, it isn't really important whether there is such a reading for deontic modals in English, or whether one of the readings is somehow prior to the other. What's important is that moral theories should not content

themselves with verdicts about objective moral status.[1]

I mentioned one motivation for caring about subjective moral status: that endorsing a moral theory should guide our actions and reactions. Another motivation (emphasised e.g. in [Jackson 1991]) is that there is something troubling about agents who always make the objectively right choice without also happening to be omniscient. Such agents would constantly take incredible moral risks and always come out lucky. To illustrate, consider another variation of the trolley scenario, where the tunnel track is empty and you have a further option to stop the trolley by throwing a large rock in its way. Doing so would cause significant damage to the trolley. Objectively, it would then be best 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 three people tied to it. In that case, surely the only defensible choice is to throw the rock. A morally conscientious agent would not do the objectively right thing and flip the switch. Note also that we can say what she would do without knowing to which track the three people are tied. We can evaluate the options in her decision problem without fixing all morally relevant facts. Our moral theory is not silent.

I should perhaps 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 determines a ranking of possible choices in decision problems. The ranking need not be complete, and it may be very coarse-grained, classifying some of the available options as right and others as wrong. 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 make no general assumptions about the internal structure of moral theories. In principle, a moral theory could be an endless, gerrymandered list of verdicts about specific decision problems, 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 single unifying principle – the categorical imperative, or the principle that pleasure is good and pain bad. A moral theory might also address decision problems indirectly, perhaps via character traits that ideal agents should possess. That alone would not yet constitute a moral theory in my sense since it would give no verdicts about decision problems. It turns into a moral theory once we specify how the relevant character traits manifest themselves in an agent's choices.

Let's say that the *domain* of a moral theory is the class of decision problems for which it provides a ranking of available acts; the domain together with the associated rankings

1 I will continue to use the entrenched labels "subjective" and "objective", but I should point out that my usage does not conform to every usage in the literature; in particular, my account of subjective moral status is not meant to satisfy the adequacy conditions Holly Smith puts forward in [Smith 2010], for reasons I will explain in section 8.

is the theory's *extension*. A theory that deals exclusively with trolley problems would have a very limited domain. A theory that only deals with decision problems in which the outcomes of all options are fully specified also has a limited domain, although in a different respect. The above observations suggest that our moral commitments are more comprehensive than that. We can evaluate the options in decision problems without knowing exactly what would happen if a given option were chosen.

I will proceed on the assumption that credible moral theories should provide an evaluation of the options in a significant range of decision problems with imperfect information. I will also assume that the relevant kind of imperfect information can always be modelled by a probability measure. The most salient interpretation of this measure, if we're concerned with action guidance, takes it to encode the agent's subjective degrees of belief. For example, if the agent is 80 percent confident that the construction worker is in the tunnel, the relevant decision problem will include a probability measure assigning probability 0.8 to that state of the world.

Here I use the term 'probability' in the mathematical sense in which any function that satisfies certain formal conditions is a probability measure. I do not assume that agents "know the probabilities" in the ordinary sense where that would require knowledge of relevant statistics or natural laws. One can be 80 percent confident that a construction worker is in the tunnel without having any such information about objective probability.[2]

It won't hurt if throughout this paper you interpret all probabilities as the agent's degrees of belief. My own view is that moral theories should evaluate decision problems without fixing the interpretation of the probabilities. We can ask what an agent should do in the light of such-and-such information without assuming that the agent, or anyone else, is aware of the information. At the limit, where we take into account all the facts, we thereby get the "objective oughts" on which moral philosophers have traditionally focused. If instead we evaluate the options relative to the information captured by the agent's degrees of belief, we get a kind of "subjective ought". If we evaluate them relative to the degrees of belief that would be appropriate given the agent's evidence, or relative to an adviser's or onlooker's degrees of belief, we get different kinds of "subjective ought".

That evidence or degrees of belief can be represented as a unique, coherent probability measure is of course an idealisation. More realistically, all we may have are unsharp qualitative constraints on probabilities: that A is much more probable than B, that C is approximately independent of D, etc. For the sake of simplicity, I will nevertheless focus on the idealized case in which the relevant probabilities are sharp. Cases of unsharp probabilities might then be treated by the method of supervaluation, but I will not

2 Following [von Neumann and Morgenstern 1944], some authors take the probabilities in decision theory to be objective; the question then arises how one should act if these probabilities are unknown. The answer is given by the decision theoretic tradition of [Ramsey 1931], [Savage 1954], and [Jeffrey 1965], where the probabilities are interpreted as degrees of belief.

discuss whether that treatment is fully adequate, and I will completely ignore cases of incoherent probabilities.

So the decision problems I want to consider are given by a range of options and a probability measure over relevant possibilities. Moral theories should determine a ranking of these options. How should they achieve that? There is always the possibility of an endless, gerrymandered list. But suppose we want something more systematic. Can we use the framework of decision theory?

3 The decision-theoretic approach

Here, in rough outline, is how standard decision theory evaluates an agent's options. Suppose the agent faces a choice between some options $A_1, A_2, \ldots$. Suppose also that her degrees of belief are represented by a probability measure P over possible *states of the world* $S_1, S_2, \ldots$. Each of these, together with any option, determines an *outcome*. Let $O[A_i, S_j]$ denote the outcome brought about by option A_i in state S_j. Suppose we have a utility function U (on which more soon) that assigns to every outcome a number. If we hold fixed a given option A_i, the utility function effectively maps each state S_j to a number $U(O[A_i, S_j])$: to the utility of choosing A_i in S_j. In the confusing jargon of probability theory, this mapping is called a 'random variable'. Its (equally ill-named) *expectation*, a.k.a. the *expected utility of* A_i, is the average of the values, weighted by their probabilities:

$$EU(A_i) = \sum_j P(S_j)U(O[A_i, S_j]).$$

Decision theory now ranks the available options by their expected utility. The best options are those with greatest expected utility.

Decision theorists disagree on the interpretation of the utility function U. There are two main approaches. The first, popular in economics and psychology, identifies the utility function with a measure of the agent's personal wealth or welfare or well-being. Decision theory thereby turns into a theory of narrow-minded, self-interested *homines economici*. Unsurprisingly, the theory does not accurately predict real people's choices, and it is doubtful that it provides a plausible standard of rationality. From a Humean perspective, there is nothing irrational about agents who, say, care more about the well-being or ill-being of others than about themselves.

Whatever its relevance to a theory of rationality, the present approach lends itself naturally to consequentialist moral theories. We simply have to replace the personal utility function U by a *moral value function* V. Just as a rational *homo economicus* would promote her personal utility U, a moral agent should promote the moral value V. One might even hold that the moral value V is simply some kind of aggregate of the personal utilities U, perhaps by the reasoning of [Harsanyi 1978]. In any case, the advice

to a morally conscientious agent would be to maximize expected moral value. For the limit case where all information is given, this means to choose whatever option actually maximizes V. If only probabilistic information about the possible consequences of each option is available, the goodness of the possible outcomes (as captured by V) has to be weighted by the corresponding probabilities.

One problem with this approach concerns the value function V. Standard decision theory assumes that utility is a *cardinal* measure, meaning that it not only represents whether one outcome is better than another, but also by *how much* the first is better or worse than the second. More precisely, it should be meaningful to say that an outcome O_1 is as much better than an alternative O_2 as a third outcome O_3 is better than O_4. It is not obvious that every consequentialist moral theory can provide such a cardinal measure of goodness. To give an objective ranking of an agent's options for decision problems where all relevant information is given, a merely ordinal measure is enough.

Another problem is the justification of the expectation rule. Why should we interpret "promoting" moral value as maximizing its expectation? This is a substantive moral assumption. To be sure, there are long-run arguments in support of the expectation rule: if everyone follows the policy of maximizing expected moral value, then by the laws of large numbers it is almost certain that *actual* moral value will be maximized in the long run, compared to any other policy. But reflections on what would happen if everyone followed a given policy are not the kind of reason that should impress an *act* consequentialist. On the other hand, rule consequentialists will not agree that the right act is the one that maximizes expected moral value in the first place.[3]

One might try to justify the expectation rule by the fact that it gives intuitively plausible verdicts in actual cases. But even that looks doubtful. For example, the expectation rule implies that agents should be risk-neutral about moral value: if an act A_1 would lead to 100 units of good, while A_2 would lead to either 0 or 200, each with probability 1/2, then the two options have equal expected moral value. But in real cases, they often don't seem equally good.

These problems only get worse if we consider non-consequentialist theories, which don't see the goal of morality in the promotion of some good attached to possible outcomes. To be sure, the discussion over "consequentializing" has shown that many non-consequentialist theories concerning objective moral status are extensionally equivalent to theories of a consequentialist structure. The trick is to individuate outcomes very inclusively so that the "outcome" of an act includes not only what it causally brings about, but also the act itself as well as the conditions under which the act was chosen. Any right- or wrong-making feature of acts can then be located in the corresponding outcomes, so that one can define a notion of goodness that is maximized if and only

3 The present point is raised by Alexander Pruss at http://alexanderpruss.blogspot.de/2014/01/consequentialism-and-doing-what-is-very.html.

if an act is obligatory according to the original, non-consequentialist theory. But this does not vindicate the use of decision theory to tackle decision problems with imperfect information. For one thing, the cooked-up goodness measure tends to be merely ordinal, not cardinal.[4] Moreover, risk-neutrality and other such consequences of decision theory look even more dubious from a non-consequentialist perspective.

In response to these worries, one might try to adapt the second approach to subjective utility, which is often regarded as the orthodox approach among decision theorists. Here an agent's utility function is not identified with any fixed measure of wealth or welfare. Rather, it is derived from the agent's preferences over possible acts, partially revealed by her choice dispositions. So-called *representation theorems* show that if the preference order satisfies certain qualitative conditions (known as decision-theoretic *axioms*), then one can construct a utility function U such that the agent prefers an act A_1 to an act A_2 just in case the expected utility of A_1 is greater than that of A_2.[5]

For the ethical case, the idea would be that the moral value function V is derived from a more basic relation of moral betterness between possible acts. If we could show that the betterness relation satisfies the decision-theoretic axioms, the existence of a suitable moral value measure and the expectation rule would then fall out from the representation theorems.

However, on closer inspection the preference-based (or "betterness-based") approach also has a number of problems. First, it makes the application of decision theory in ethics utterly opaque. Consider a trolley problem with imperfect information. It would be nice if we could use the tools of decision theory to get clear about what the agent should do. That is, we would like to derive the subjective moral ranking of the agent's options on the basis of the probabilities and the moral value pertaining to the possible outcomes. The present approach suggests that the derivation should instead go in the opposite direction: we first have to know the answer to all decision problems with imperfect information in order to derive the moral value function.

Relatedly, the present approach threatens to trivialise the idea that subjective moral status is determined in a decision-theoretic manner. The idea now reduces to the hypothesis that the moral ranking of acts in decision problems conforms to the decision-

4 By this I mean that any monotonic (order-preserving) transformation of the goodness measure would do just as well. Cardinal information is irrelevant if we only want to reproduce a theory's objective oughts.

5 Classics in this tradition are [Ramsey 1931], [von Neumann and Morgenstern 1944] and [Savage 1954]. The preference-based approach has its roots in the positivist doctrine that scientifically legitimate quantities must be defined in terms of directly observable phenomena. As such, it is widely ridiculed in contemporary philosophy (e.g. [Meacham and Weisberg 2011]). However, it has a more respectable successor in the functionalist position, sometimes called "interpretivism", defended e.g. in [Lewis 1974], [Stalnaker 1984], and [Blackburn 1998] (and indeed [Ramsey 1931]). That position takes an intermediary route between assuming subjective utilities as basic and deriving them from choice dispositions. I will advocate a similarly intermediary route for the case of moral value.

theoretic axioms. But arguably the axioms are too weak to characterize genuinely decision-theoretic moral theories.

In other respects, the axioms may be too strong – especially if we want to accommodate non-consequentialist theories. Recall that to make decision theory applicable to such theories, "outcomes" must be individuated very finely, ideally so that they include the acts that brought them about. It turns out that this move is incompatible with standard axiomatizations of decision theory in the tradition of von Neumann and Morgenstern [1944] or Savage [1954], for which it is essential that the very same outcome can be brought about by different acts.[6]

Finally, even if we grant that outcomes don't include the acts that brought them about, most of the "possible acts" over which preferences are assumed to be defined in classical axiomatizations do not correspond to options any agent could possibly face – let alone in a realistic, single decision situation. But all we can take for granted is that moral theories rank the options within a significant number of decision problems that agents may realistically face.[7]

In sum, neither of the traditional approaches to subjective utility looks promising if we want to apply decision theory to moral theories, including non-consequentialist theories.

There are further worries with the decision-theoretic approach, no matter how the moral values are defined. First, non-consequentialist theories often postulate agent-relative norms. Suppose some moral theory – call it M – says that in a certain situation Alice should dance with Bob, although Bob should not dance with Alice. How is this possible if the rightness of acts is derived from the impersonal goodness of outcomes? Either the

6 The assumption that outcomes are logically independent of acts is usually not presented as an axiom, but hidden in the definition of the act space. To see the importance of the assumption, consider the central *independence* axiom (also known as the *separability* axiom or the *sure-thing principle*) of classical decision theory:

> If two acts A and A' lead to the very same outcomes unless some condition C obtains, and if two further acts B and B' coincide in terms of outcomes with A and A' respectively whenever C obtains, and coincide with one another whenever C does not obtain, then the agent prefers A to A' iff she prefers B to B'.

If different acts can never lead to the same outcome, there are no acts at all that satisfy the stated conditions on A, A', B, and B': the axiom becomes empty.

One might try to accommodate non-consequentialist values (or, for the theory of rationality, non-consequentialist desires) without going all the way to include acts in outcomes, but it is an open question to what extent this is compatible with the spirit and/or the letter of classical axiomatic decision theories. There is a large literature on this topic; see e.g. [Tversky 1975], [Broome 1991: chs.5–6], [Pettit 1991], [Dreier 1996b], and [Joyce 1999].

7 A promising response to this problem (advocated e.g. in [Joyce 1999] for rational choice and in [Broome 1991] for consequentialist moral theories) is to interpret the preference relation not directly in terms of choice behaviour, but as comparative desirability, which is only partly and indirectly related to choice. However, it is not obvious that every credible moral theory must provide a sufficiently rich relation of comparative moral desirability.

outcome in which Alice dances with Bob is better than the one in which she doesn't dance with Bob, or it is not better. How could it be better relative to Alice and worse relative to Bob? One might say that different moral theories – different standards of goodness, different value functions – pertain to Alice and to Bob, but M is supposed to be a single moral theory, simultaneously giving verdicts for both Alice and Bob.

Second, the expectation rule seems to over-generate moral obligations. Suppose you think about going for a bike ride in the afternoon. If you go, there is a greater risk that you might get in an accident and injure (say) an innocent child than if you stay at home. Assuming that injuring the child has significant moral disvalue, staying at home plausibly comes out as having greater expected moral value. But are you really obligated not to go on the bike ride, merely due to the small risk that you might injure a child? Relatedly, if the subjective moral imperative is to maximize expected moral value, there seems to be no room for supererogatory options: options that are morally better than the alternatives and yet are not morally required. Similarly, there seems to be no room for moral dilemmas in which all options would be wrong. Non-consequentialists often want to allow for supererogation and moral dilemmas.

Finally, it has often been pointed out that as a practical guide, the principle to maximize expected moral value is pretty useless, as it seems to presuppose not only that the decision maker is conscious of the relevant probabilities and moral values, but also that she is capable of performing the required calculations, which quickly become intractable if the space of relevant options and states is as large as it usually is.

As a first step towards answering these worries, I suggest that we replace the classical framework of "expected utility theory" with a version of Richard Jeffrey's [1965] "logic of decision".

4 The logic of decision

The Logic of Decision [Jeffrey 1965] introduced some key innovations to decision theory. One is to jettison the traditional distinction between acts, states, and outcomes. In Jeffrey's account, probabilities and utilities (or "desirabilities") are both defined over entities of the same kind: propositions. Among other things, this has the advantage that one can consider probabilities conditional on acts, or probabilities of complex propositions involving acts – something that isn't possible in earlier frameworks like [Savage 1954] and that proves important to give a clear treatment, for example, of Newcomb problems (see [Joyce 1999: 117]). Moreover the basic utilities in Jeffrey's account are assigned not to restricted "outcomes", but to maximally specific propositions, which allows us to easily capture agents who care not only about outcomes but also about how the outcomes were caused.

Jeffrey's theory is also beautifully simple. Let $\mathcal{A}$ be the set of propositions whose

truth-value is relevant to a given decision problem. Assume that $\mathcal{A}$ is closed under conjunction, disjunction and negation. Every element of $\mathcal{A}$ is then entailed by a *maximal* element: a proposition W that entails A or $\neg A$ for every A in $\mathcal{A}$. Let $\mathcal{W}$ be the set of these maximal elements. Intuitively, a proposition in $\mathcal{W}$ is a description of a possible way things could be that includes everything that matters for the evaluation of the options. Now assume we have a value function V over the members of $\mathcal{W}$, and a probability measure P over $\mathcal{A}$. Then we can define the *desirability* of any proposition $A \in \mathcal{A}$ as the (conditional) expected value

$$EV(A) = \sum_{W \in \mathcal{W}} P(W/A)V(W). \qquad (*)$$

Here $P(W/A)$ is the probability of W conditional on A, which equals the ratio $P(W\&A)/P(A)$ whenever $P(A)$ is greater than 0.[8]

Here is an example. Suppose you consider having a glass of cognac. The problem is that cognac sometimes gives you a headache, and you don't like headaches. For simplicity, let's assume that there is nothing else you care about to which your choice might make a difference. We can then take the atomic propositions in $\mathcal{A}$ to be the proposition that you have a glass of cognac (for short, C) and the proposition that you will get a headache (H). The set $\mathcal{W}$ of maximally specific propositions then has four elements: (1) $C\&H$, (2) $C\&\neg H$, (3) $\neg C\&H$, and (4) $\neg C\&\neg H$. Conditional on the hypothesis that you have cognac, (3) and (4) have probability zero: given that you do have cognac, it is certainly not the case that you don't have cognac. So the the desirability of having cognac is given by

$$EV(C) = P(C\&H/C)V(C\&H) + P(C\&\neg H/C)V(C\&\neg H),$$

which can be further simplified to

$$EV(C) = P(H/C)V(C\&H) + P(\neg H/C)V(C\&\neg H).$$

That is, to evaluate the desirability of having cognac, we have to consider two scenarios: that you have cognac and get a headache ($C\&H$), and that you have cognac and don't get a headache ($C\&\neg H$). The desirability of having cognac is the average of the values assigned to these scenarios, weighted not by their absolute probability, but by their probability conditional on a decision to have cognac. The desirability of not having cognac is similarly given by

$$EV(\neg C) = P(H/\neg C)V(\neg C\&H) + P(\neg H/\neg C)V(\neg C\&\neg H).$$

8 The present definition of desirability is unproblematic if the relevant set $\mathcal{W}$ is finite. Standard ways of extending the notion of a sum to infinite sets are often employed to deal with decision problems that require larger sets of possibilities, but it is not entirely unproblematic that these extensions are adequate, and they don't work for all cases. Fortunately, most real-life decision problems can be modelled with a finite set $\mathcal{W}$.

Jeffrey assumes that desirability is also a measure of choiceworthiness: rational agents should choose options with greatest desirability. The assumption is natural. However, it is widely (though not universally) taken to go wrong in unusual scenarios like Newcomb's problem where an act would be evidence for a desirable or undesirable proposition without having any causal bearing on that proposition (see e.g. [Gibbard and Harper 1978], [Joyce 1999: sec. 5.1]). According to proponents of "causal decision theory", an adequate standard of choiceworthiness must take into account the causal relationship between a possible choice and the relevant scenarios in $\mathcal{W}$. Different proposals have been made to render the idea precise – classical treatments include [Gibbard and Harper 1978], [Lewis 1981], [Skyrms 1984] and [Joyce 1999], but arguably none of them is fully satisfactory. I will not enter into these details here. In everyday decision problems, extant forms of causal decision theory and Jeffrey's "evidential" theory almost always agree, and give sensible results.

Jeffrey's approach has not been popular among economists and psychologists, perhaps because it makes no testable predictions about what rational agents will do, without assuming anything about their desires. The reason is that the ultimate bearers of value in Jeffrey's system are the fine-grained propositions in $\mathcal{W}$ that specify not only the consequences of an act, but also the act itself as well as anything else that might matter to the agent. Hence Jeffrey's theory does not predict that if you would choose an apple over a banana, and a banana over a carrot, then you should 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, the theory does not say that you should prefer a gamble between \$10 and \$20 to a sure outcome of \$1: if you hate gambles, you may well prefer the sure dollar. For analogous reasons, Jeffrey's theory makes no predictions about the scenarios of Allais [1953] and Ellsberg [1961] that are widely thought to provide counterexamples to traditional expected utility theory. Indeed, any pattern of choice dispositions whatsoever can be made to trivially conform to the standard of maximizing desirability: if the agent is disposed to choose option O in a decision problem D, simply assume that she assigns high basic value to elements of $\mathcal{W}$ according to which the agent chooses O in D. There is an elegant representation theorem for Jeffrey's theory, due to Bolker [1966], which shows how the value function V can be derived from a qualitative preference relation,[9] but that relation is not a relation between possible acts; rather, it is a relation of comparative desirability between arbitrary propositions in $\mathcal{A}$. The Bolker-Jeffrey axioms alone do not constrain an agent's choices.[10]

9 Like Ramsey [1931] and Savage [1954], Jeffrey and Bolker actually consider the harder task of deriving both a utility function and a probability measure from the agent's preference relation. For our purpose, we can take the probabilities as given, which simplifies the problem.

10 To illustrate, consider the following axiom of *averaging*:

> If two propositions X and Y are incompatible and X is preferred to Y, then X is preferred

In my view, all this is just as it should be if we're interested in a pure theory of *instrumental* rationality: in a theory that says what an agent should do to further her goals, without presupposing anything about these goals. Substantive predictions about what such an agent should do must always be based on assumptions about the agent's goals.

In any case, I want to suggest that a theory along Jeffrey's lines is an adequate framework to address moral decision problems. To that end, it will be useful to adopt a minor variation of Jeffrey's framework, due to Lewis ([1979], [1981]), on which probabilities and utilities are defined not over propositions but over *properties* – things like living in London, liking lasagna, or being friends with a plumber.

Assigning probabilities to properties may seem odd. What does it mean to say that there is a 50 percent probability of living in London? Well, recall that the probabilities we are interested in capture imperfect information about a decision problem. In this context, to say that some property F has probability x simply means that the agent in question has property F with probability x. Using properties as bearers of probability helps to model cases in which an agent is uncertain who they are or what time it is (see [Lewis 1979]). For our topic, the more important advantage of properties is that they provide a natural locus of agent-relative and time-relative value (compare [Lewis 1989: 73–76], [Dreier 1996a]).

Return to the above case where (according to some moral theory) Alice should dance with Bob while Bob shouldn't dance with Alice – perhaps because Alice has promised to dance with Bob and Bob has promised not to dance with Alice. The case is hard to explain if the bearer of moral value are propositions or states of affairs. Either the situation in which Alice and Bob dance with one another is morally better than the situation in which they don't, or it is not morally better. The problem dissolves if we take the bearers of moral value to be properties. After all, the properties that Alice would come to instantiate by dancing with Bob are quite different from those Bob would come to instantiate by dancing with Alice. In particular, Alice would *fulfil a promise* while Bob would *break a promise*. To be sure, Alice would also *contribute to Bob breaking his promise*, but agent-relative theories will insist that breaking a promise is worse than allowing others to break a promise. Consequently, the total profile Alice would come to instantiate by dancing with Bob can be assigned greater moral value than the profile Bob would come to instantiate in the same scenario.

---

to the disjunction $X \vee Y$, which in turn is preferred to Y.

The idea is that an unspecific proposition $X \vee Y$ cannot be better or worse than both of the more specific properties X and Y. Killing an innocent cannot be worse than killing an innocent man and also worse than killing an innocent woman. This is compatible with the assumption 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.

Having properties as bearers of value also allows us to apply decision theory without first consequentializing moral theories. Act types are properties. So we can directly assign moral value to things like intentionally killing an innocent, or resisting a strong desire to gossip, without any detour through finely individuated outcomes.[11]

If the algebra $\mathcal{A}$ over which probability and desirability are defined is an algebra of properties, then its maximal elements $\mathcal{W}$ are maximally specific properties: properties that entail, for every other property in $\mathcal{A}$, whether it is instantiated or not. I will call such maximal properties *profiles*. Intuitively, the profile brought about by a choice in a concrete decision situation is the totality of all properties the agent comes to instantiate as a result of the choice. If we can find a moral value function V that provides a cardinal ranking of possible profiles, we can use Jeffrey's formula (*) to evaluate the options in decision problems with imperfect information.

5 Moral value

I mentioned above that for any pattern of choices there is a trivial way to construct a value function V that makes the choices conform to the principle of maximizing expected value in the sense of (*): if the pattern involves choosing option O in decision problem D, simply assign high moral value to choosing O in D. In that sense, any ranking of options in decision problems with perfect and imperfect information can be made to fit the decision-theoretic picture. However, I want to defend a more substantive, non-trivial role of decision theory in moral theories. The basic idea is that we can use decision theory to turn the "objective" verdicts a moral theory gives for decision problems with perfect information into "subjective" verdicts about decision problems with imperfect information.

A theory of objective moral status in effect provides a (partial) ranking of what I called profiles. If the ranking of profiles has a suitable structure, it can be represented by a value function V that allows us to rank the options in ordinary decision problems by their expectation. What needs to be shown is, first, that the objective ranking of profiles has the right structure, and second, that ranking options by their expected value is appropriate for decision problems with imperfect information.

Before I turn to these matters, I should emphasize that the ranking of profiles, and the value function that represents it, is not supposed to be a moral primitive. I do not assume that promoting V is the goal of moral agency. Instead, a moral theory might attribute

11 An alternative response to the worry that basic value in non-consequentialist theories does not attach to states of affairs would follow the observations of [Pettit 1991] and argue that the utilities taken as basic in standard decision theory are actually derived from more basic utilities assigned to properties instantiated in the relevant outcomes. But that alone would not allow for agent-relative and time-relative values.

basic moral relevance to specific properties such as keeping promises, harming innocents, obeying God's commands, or acting in accordance with the categorical imperative. Even if a moral theory evaluates acts by their relation to some independently given measure of good, that measure may or may not match the value function V. For example, rule utilitarians might take overall welfare as the ultimate standard for evaluating acts, without ranking profiles by the amount of overall happiness. (Rather, they will base their ranking on whether a profile involves an act type whose general adoption would lead to greater welfare.)

Any theory that provides an objective ranking of options in fully specified decision problems thereby provides a partial ranking of profiles. At this stage, we can't assume that the ranking already covers the options in arbitrary decision problems with imperfect information, so we can't use the representation theorems of axiomatic decision theories to construe the moral value function. We have more work to do.

Decision theory requires that the ranking of profiles satisfies certain structural conditions. In particular, the ranking should be sufficiently comprehensive and sufficiently cardinal. Let's begin with the first.

Consider the trolley scenario in which your are uncertain whether the construction worker is in the tunnel. There are at least four relevant profiles, corresponding to the four combinations of flipping and not flipping with the two possibilities concerning the construction worker:

(1) you flip the switch, save the three people tied to the track, but kill the construction worker;
(2) you flip the switch, save the three, and kill no-one;
(3) you don't flip the switch, let the trolley run over the three, and avoid killing the construction worker;
(4) you don't flip the switch, let the trolley run over the three, while the construction worker was not in danger.

In order to apply decision theories, we must be able to compare these profiles. It should make sense to say, for example, that (1) is better than (2).

In decision theory, it is commonly assumed that the ranking of profiles (or outcomes) is in fact *complete* in the sense that for any two profiles W_1 and W_2, either W_1 is better than W_2, or W_2 is better than W_1, or W_1 and W_2 are equally good. It is not plausible that moral theories must provide a complete ranking in this sense. Incompleteness can arise in several ways.

First of all, moral theories may have little or nothing to say about the comparative status of profiles that concern different decision situations. Second, moral theories can be vague and indeterminate in various respects. For example, if we say that benevolence is good, without assuming a perfectly precise definition of benevolence, our theory might

fail to rank profiles involving borderline cases of benevolence. Third, moral theories might define different dimensions of right and wrong, but no precise and general rule for aggregating these into a single moral ranking. Fourth, moral theories might explicitly *allow for* a whole range of rankings. For example, a moral theory might say that if the only way to prevent a tragedy is to harm an innocent, then no choice is straightforwardly right, nor are the two options exactly equal in terms of rightness; a morally conscientious agent could go either way, depending on how she personally resolves that difficult choice.[12]

Incompleteness of the first type is irrelevant for our application, which only requires well-defined values for the profiles within any given decision problem. The other forms of incompleteness can all be modelled by assuming a whole class of moral value functions. If an option maximizes moral value relative to some legitimate value functions and not others, we can then say that there is no fact of the matter about whether you ought to choose the option. Alternatively, if the multitude of value functions represents indeterminacy of the fourth kind, we can say that the option is not required but permitted. As I will further explore in section 7, having a class of value functions nicely captures the intuition that morality often leaves open exactly what we should do.

So we don't need completeness. But we still need a certain degree of comprehensiveness: if a theory can say nothing definite about which of the profiles in our trolley problem are better or worse than others, decision theory can't tell the agent what to do.

Comprehensiveness is not enough. For one thing, we must also assume that the ranking is *consistent* in the sense that (i) if a profile W_1 is strictly better than W_2, then W_2 isn't strictly better than W_1, and (ii) if a profile W_1 is better than W_2 and W_2 is better than W_3, then W_1 is better than W_3. More substantively, the objective evaluation of profiles should settle not only whether one profile is better than another, but also whether it is *a lot* better, or only *a little* better, compared with the difference between other profiles. In our trolley problem, a theory that emphasizes the distinction between killing and letting die might say that (2) is better than (3), which is better than (4), which is better than (1). But that is not enough to figure out what you should do. We also need to know how the difference between (2) and (3) compares to that between (4) and (1).

Again, we don't need to assume that such ratios of differences are always well-defined and precise. We can allow for indeterminacy in numerical representation. But we do need at least *approximately cardinal* structure if we want to give non-trivial verdicts about decision problems with imperfect information. The cardinal structure can't be read off from a theory's "objective oughts". If we start with a theory that only provides an ordinal ranking of the options in decision problems with perfect information, we have to enrich the ranking before we can tackle problems with imperfect information.

However, it is plausible that our objective evaluation of profiles already has considerable

12 The non-obvious fourth possibility was brought to my attention by [Gert 2004] and [Rabinowicz 2008], where it is actually used to model a form of the third possibility, what Chang [2002] called *parity*.

cardinal structure. We know that the moral difference between murdering innocents and stealing $10 is greater than that between stealing $10 and stealing $5. Such judgements are needed, among other things, to connect our moral theory with our practice of blame and praise, punishments and rewards. The latter come in "cardinal" units. We tend to criticise people less for cheating on their taxes than for sexual assault. It is implausible that this does not track any moral difference between the acts. We also 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. 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 make little sense if we didn't have approximately cardinal moral values.

In addition, we can take a lesson from the preference-based approach and work backwards: consider the verdicts we want for decision problems with imperfect information and try to retrofit the cardinal structure of objective values. To illustrate, imagine a scenario in which an agent faces a choice of letting one person die or taking a 50 percent chance that either two people or no-one will die. At first glance, it might seem that all we can say from an objective point of view is that in the given scenario letting two people die would be worse than letting one die, which again would be worse than letting no-one die. This merely ordinal ranking is not enough to determine what the agent should do, if she doesn't know what would happen as a result of taking the risk. But suppose we also think that it would be wrong for the agent to take the risk. Assuming that acts are evaluated in terms of expected moral value, it follows that the difference in moral value between the profile in which two people die and the profile in which one dies is less than the difference between the latter and the profile in which no-one dies. Can we make systematic sense of these cardinal judgements? Arguably yes. For example, the judgements could be explained by the assumption that there is something intrinsically wrong not only with letting people die but also with imposing risks on people. This wrong is involved in the profiles with zero and two deaths, but not in the profile with one death, which explains why the latter gets ranked higher than it would if the number of deaths were all that mattered.[13]

There is no guarantee that the forwards and backwards route will always cohere. Consider an "absolutist" moral theory according to which it is never permissible to (say) kill an innocent, no matter the circumstances. This is a claim about objective ought, since it doesn't say what to do in situations where it is uncertain whether a given act

13 Personally, I think it is an open question whether the agent's degrees of belief should factor in a profile's moral value. We certainly criticise people for acting in a way that might have caused serious harm, even if no actual harm occurred, but those judgements might well be judgements about subjective moral status. See [Smith 2010: 87–89].

would involve a killing of an innocent. Absolutists might suggest that one should never act in a way that might possibly involve a killing of an innocent. However, with some imagination it is not difficult to see that almost every option in almost every real-life decision problem would thereby come out as impermissible. So consider a moderate absolutist who agrees that it is sometimes permissible to incur very low probabilities of killing an innocent. So far, this combination of objective and subjective claims is easy to model in the decision-theoretic framework. But suppose the absolutist's account of objective moral status includes more detailed, cardinal verdicts: it says that killing someone is always worse than letting people die, and that the moral difference between letting $n+1$ people die and letting n people die is the same for all n. Now we have a problem.

In fact, we have several problems. A first is how to model these verdicts about objective status by a numerical value function at all, even without turning to cases with imperfect information. If killing an innocent has finite disvalue and there is a fixed amount x by which letting $n+1$ people die is worse than letting n people die, then it looks like for some n, letting n people die will have to be worse than killing an innocent. A popular way to prevent this conclusion is to say that killing an innocent should have infinite disvalue. Alternatively, we could move to a non-Archimedian number system in which there are finite limits of $n \cdot x$ as n goes to infinity (see [Bernstein and Wattenberg 1969]); or we could use vectors of reals instead of single numbers to reflect the apparently "lexical" ordering of profiles. There is certainly no conclusive reason why legitimate value functions should only take real-numbered values. The basic decision-theoretic approach does not presuppose that they do.

Having dealt with this preliminary problem in some such fashion, we can now see the tension between the moderate absolutist's claims about objective and subjective status. If killing an innocent is "infinitely worse" than letting someone die, it looks like, by decision-theoretic rules, it is never permissible to, say, rescue a life if that comes with a non-zero probability of killing an innocent; and then it is hard to see what other acts could be permissible. Thus, if we work backwards from the verdicts our moderate absolutist wants to give for decision problems with imperfect information, it looks like we get objective moral rankings that contradict the absolutist's direct cardinal judgements about profiles.

There are many ways in which one might try to resolve this apparent decision-theoretic incoherence, apart from reconsidering the relevant judgements. Two obvious strategies are to identify further relevant features of profiles (following the above treatment of risk) or to revise the implicit assumption that the overall value of a profile is simply the sum of the value of its components. These possibilities would be worth exploring before we conclude that the moderate absolutist position is incompatible with the decision-theoretic

approach.[14]

Since I just mentioned non-standard numbers, it may be worth pointing out that the use of numbers to represent moral value is really just a matter of mathematical convenience. By suggesting that moral theories should use the framework of decision theory, I don't suggest that they should include a specific numerical value assignment to possible profiles (or a class of assignments). We don't need to decide whether flipping the switch in a trolley problem has moral value 12.7 or -4.5. The proposal is rather that (i) moral theories should provide a ranking of profiles that is approximately cardinal in the sense that one can say whether W_1 is a lot, or only a little, better than W_2, and that (ii) the evaluation of decision problems with imperfect information should be systematically determined by the ranking of the profiles, in the sense that an option is ranked as better than another iff it has greater expected moral value relative to every value function V that respects the cardinal ranking of profiles. The actual numbers cancel; they need not figure anywhere in the moral theory.

6 Expectations

Why should the subjective ranking of an agent's options be given by the options' expected moral value in the sense of (*)? The question is especially pressing for consequentialist theories in which the moral value function is identified with some prior measure of good. Why should the right way of promoting a good in the light of limited information be to maximize its expectation? A promising line of response is pursued in [Broome 1991]. Adapted to Jeffrey-type decision theories, Broome's reasoning is that once a probability measure is given, the consequentialist goodness ranking can be extended from complete profiles to arbitrary propositions. If the extended ranking satisfies the Bolker-Jeffrey axioms, then it follows by Bolker's representation theorem that the (extended) moral value assigned to any option equals the expectation of the value assigned to the corresponding profiles, as stated in (*). To conclude the argument, we would have to explain, first, why the objective ranking of propositions should satisfy the Bolker-Jeffrey axioms, and second, why subjective rightness should go with extended moral value.

The situation is easier for non-consequentialist theories. As I explained in the previous section, we can here make the construction of the value function sensitive to verdicts we want to give for decision problems with imperfect information. In contrast to decision-theoretic forms of utilitarianism, we can, if we want, punish risk or reward fairness. The substantive assumption is no longer the expectation standard itself, but the assumption that objective value often turns on features of which the decision maker is unaware, and

14 Some attempts at modelling absolutist theories in a decision-theoretic setting are discussed in [Jackson and Smith 2006] and [Colyvan et al. 2010].

that objective and subjective value judgements should cohere in the sense of the previous section. These are the assumptions that make the decision-theoretic approach non-trivial.

I have no knock-down argument that any credible moral theory should satisfy these assumptions. But it does seem plausible that the moral evaluation of our options often turns on different possibilities about which we are not fully informed. That is what makes decision-making with imperfect information hard. So we need some way of taking into account the various possibilities. The expectation rule provides a simple, unified and systematic way to do that. It can turn any value function for profiles into rankings for almost every decision problem an agent could face.

This systematicity is a great advantage. The space of possible decision problems is enormous. There are countless ways to 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 verdicts 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 alternatives to the expectation rule. But to the extent that they diverge from the expectation rule, these alternatives tend to be extremely implausible. Consider the *maximin* rule, which says that one should choose an option with the best worst-case profile. If a medical treatment has a 99.9 percent chance of curing a crippling condition, but a 0.1 percent chance of making it slightly worse, maximin suggests that it would be wrong to prescribe the treatment (because the worst-case outcome, the slightly worsened condition, is worse than the outcome of non-action). Or consider the popular, but unnamed strategy to do whatever is most likely to be best. This suggests that you should prescribe a treatment that has a 51 percent chance of slightly improving a minor condition and a 49 percent chance of killing the patient.

Another class of alternatives that are sometimes put forward by non-consequentialists are *threshold rules* according to which an act is right or wrong if it is sufficiently probable that certain states of affairs obtain – say, that you should prescribe a treatment only if it is more than 95 percent likely to cure the patient. Such rules involve hard to justify cut-offs, they don't generalize well, and tend to have implausible consequences for various edge cases. For example, the rule just outlined does not extend easily to cases where your choice is to prescribe the treatment to several people at once, and it wrongly assumes that it is irrelevant what happens in the 5 percent cases where the cure doesn't work.[15]

15 These and other problems for threshold rules have been extensively discussed elsewhere, see e.g. [Feldman 1995], [Jackson and Smith 2006] and [Lazar Unpublished].

An adequate rule should consider all things that might happen, and it should take them into account in proportion to their probability. This is just what the expectation rule does.

Here is another reason to think that the expectation rule should play a role in the moral evaluation of decisions. Above I mentioned that a Jeffrey-type decision theory is plausible as a theory of purely instrumental rationality. It says how to rationally pursue one's desires or goals in the face of limited information, without assuming anything about the content of these desires or goals. It is natural to think of objective moral judgements as putting forward moral goals – not to kill innocents, etc. Thus the corresponding application of 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.

Relatedly, imagine an agent who cares about nothing but doing what's morally right. Arguably, such an agent could still face non-trivial decision problems. And it would be odd if she would then have to make choices that violate the principles of a minimal, Jeffrey-style decision theory. Morality does not require practical irrationality.

As I said, I have no knock-down argument that any credible moral theory should have a non-trivial decision-theoretic structure of the kind I have advocated. Consider the view that the only right- or wrong-making features of acts are located in the agent's intentions, and suppose people generally know what intentions would go along with the choice of a given option (which is plausible if the options actually *are* intentions, as argued e.g. in [Weirich 1983] and [Sobel 1983]). Then there is (at least on the surface, and from the decision-maker's own perspective) never any problem about decision-making with imperfect information, since agents are always fully informed about the morally relevant features of their options. For another example, consider the view that the moral status of options is given by their evidential expected value as given by (*), relative to some value function, while the rational status of options is given by causal expected value as defined by some form of causal decision theory. There is obvious decision-theoretic structure in such a theory, but it is not quite the structure that I think is most plausible for moral theories.

7 Optionality and supererogation

In these final two sections I want to address the last two worries left from section 3: the problems of optionality and supererogation, and the problem of practical guidance.

The problem of supererogation is to explain how it can be morally permissible to choose an option even though another option would be morally better. It is hard to see how this could be true if (a) we are always morally obligated to choose an option

with greatest expected moral value, and (b) an option is morally better iff it has greater expected moral value.

More generally, the decision-theoretic approach seems to leave us little freedom to follow our personal goals and whims as long as these don't have serious moral repercussions. If morality demands that we choose options with greatest expected moral value, the only choices left to us would concern cases where several options are exactly tied for expected value (or cases where the expectations are undefined). It is implausible that such cases are very common, and they certainly don't sustain the kind of personal freedom many of us think we have.

However, remember from section 5 that the decision-theoretic approach doesn't assume a unique moral value function (or a value function that is unique up to positive linear transformations). Moral theories don't need to provide a complete ordering of all possible profiles. They may not fix exactly how different right-making and wrong-making features should be balanced; they may even explicitly delineate a range of permissible rankings. As a result, we have to evaluate an agent's options by a whole range of value functions. While ties relative to any single value function will be unlikely and rare, it is not unrealistic that different permissible value functions will often rank an agent's options differently. If we say that an option is obligatory only if it maximizes expected moral value relative to all permissible value functions, we get a considerable amount of personal freedom.

We still don't get supererogation. The central feature of supererogatory acts is that they are determinately better, yet not obligatory. A superficial way to allow for that is to dissociate obligatoriness from maximizing expected value. Thus we could follow the idea of "satisficing consequentialism" and say that one is obligated to choose an option only if it is (determinately) *a lot* better in terms of expected moral value than the alternatives, or if it is the only option whose expected value (determinately) exceeds a certain threshold, relative to some scale. This is still an essentially decision-theoretic account insofar as the moral status of an agent's options is still determined by their expected moral value.

Comparing an option's expected value to fixed points on the value scale might also provide a means to allow for moral dilemmas in which all options would be wrong. They would be wrong in the sense that their expected moral value lies below the fixed threshold that makes an option wrong.[16]

I have no general objections to these proposals, but I don't think they suffice to fully account for the phenomenon of supererogation. Consider an everyday scenario in which you could use a significant portion of your yearly income to rescue a friend from losing their home, or to save many distant strangers from starvation, and suppose we judge

16 Another type of moral dilemma in decision-theoretic accounts are so-called unstable decision problems, where any given choice makes a different choice preferable. Some forms of "ratificationism" entail that no act is permissible in these situations (see e.g. [Harper 1984]).

this act to be supererogatory: laudable, but not obligatory. Crucial for this judgement is clearly that rescuing the friend or the strangers would involve a great personal sacrifice. If you didn't care about your income and had no other use for it, helping would be obligatory. Thus in cases where you're allowed to keep the money, your personal reasons for keeping it somehow trump the moral reasons for giving it away. Moreover, the trumping is a *moral* kind of trumping: it's not that you are morally required to give away the money, but all-things-considered allowed to keep it. Whether or not you're allowed to keep the money is a genuinely moral question, something different moral theories will disagree about. On the other hand, it would be wrong to count your reasons for keeping the money as trumping *moral* reasons, for then giving away the money would be morally wrong.

Genuine supererogation therefore requires that non-moral reasons can morally override moral reasons (see [Portmore 2008a]). This may seem puzzling, and some might want to conclude that the idea of supererogatory acts should simply be given up. The decision-theoretic approach certainly doesn't require that there is supererogation. However, when we turn to decision problems with imperfect information, the phenomenon does seem to be fairly common. How can it be morally acceptable to go on a bike ride if there's a small chance that this will cause harm to innocents?[17] It looks like the non-moral goods probably involved in going on a bike ride can override the low risk of harming an innocent. Here again, the crucial point is that going on the bike ride isn't just "all things considered" permissible; it is *morally* permissible.

[Portmore 2008b] develops a promising approach to model genuine supererogation. His focus is on objective moral status and consequentialist theories, but the basic strategy also works in our more general context.

To adequately account for supererogation, a moral theory should recognize nonmoral values or reasons, and specify how these may be balanced against moral reasons. In essence, we need not one but two rankings of profiles. The first ranks profiles by their purely moral status, the second ranks them in terms of the agent's well-being, welfare, preference satisfaction, or whatever non-moral considerations are acknowledged by the theory. (What kinds of non-moral reasons can trump moral reasons is again a moral question that we can't fix once and for all, for every moral theory.) Finally, the theory must say how the two kinds of value may be balanced against each other to determine an overall ranking. A simple way to model this might define a range of permissible overall value functions as weighted averages of the moral and non-moral value functions, relative to some fixed scale. The agent's options can then be evaluated in terms of expected

17 In real life, not going on a bike ride also involves a small chance of harming innocents, and going on the bike ride can lead not only to harm but also to various desirable events, but presumably it would be OK to go even if for some reason we could rule out these possibilities.

overall value relative to all permissible overall value functions.[18] An option is morally best iff it has greatest expected moral value. But the agent is not always obligated to choose the morally best option because the option may no longer be best when non-moral values are taken into account. In that case, the option is merely supererogatory. If an option is morally and overall best relative to all permissible weightings of moral and non-moral values, then it is obligatory.[19]

In practice, the division between the two kinds of values may be a little blurry, because the agent's personal goals and intentions often matter for the purely moral evaluation. Torturing and killing animals for fun is widely regarded as worse than torturing and killing animals for profit. (In many countries, the former is illegal and the latter subsidised.) A moral value function can certainly be sensitive to such intentional aspects of profiles. However, one might also try to account for the relevant judgements in terms of the balancing of moral and non-moral values. That is, one might assign moral disvalue to torturing and killing animals (ignoring motivation), and count profit, but not fun, as a good enough nonmoral reason to outweigh the moral disvalue. The two accounts are not equivalent. For example, they give different verdicts about the case of a person who tortures and kills animals, sells their meat for profit, but is motivated largely by the joy of killing and torturing.

8 Guidance

I have assumed that moral theories need to evaluate the options in decision problems with imperfect information: they should answer questions about what an agent ought to do given such-and-such options and such-and-such information about the world. One motivation for this assumption was that moral theories should be action guiding in the sense that endorsing a moral theory should manifest itself in judgements about the choices people should make in real-life decision problems, where not all information is given.

I have suggested that the framework of decision theory provides an attractive method to address this need. The proposal concerns the internal structure of moral theories. I have argued that a theory's verdicts about decision problems with incomplete information should be related in a certain systematic way to its verdicts about decision problems with complete information.[20]

18 Alternatively, we might aggregate the moral and non-moral values only on the level of options, defining the permissible overall rankings of options as a weighted average of their expected moral value and their expected non-moral value. Mathematically, these two approaches are equivalent. Of course, further, more complicated approaches are also possible.

19 Note that even if the moral value function is unique, we can now get a range of morally permissible options in many decision problems. So our solution to the problem of supererogation further helps with the problem of optionality (as one might have expected).

20 This explains why I have paid no attention to moral uncertainty. What should you do, say, in a trolley

I have said nothing about the cognitive mechanisms involved in endorsing a moral theory or in applying it to a given decision situation – that is, about the mechanisms by which our actions might be guided by endorsing a given theory. In particular, my proposal is not that morally ideal agents should somehow store a moral value function and then compute the expected value of her options on the basis of that function and the subjective probabilities whenever she faces a decision. How to design a cognitive system so that it systematically follows the advice of a moral theory is an interesting question, but I don't think it is a question for ethics.

On the other hand, it is an open question whether it is physically and biologically possible for agents like us to consistently choose options that (say) maximize expected value relative to some non-trivial value function. In real-life decision problems, the relevant space of profiles and options is typically huge. It therefore becomes infeasible to explicitly store the answers to all decision problems, but also to compute the answers on the spot within the time and energy constraints typically imposed by a decision situation (see e.g. [Gershman and Daw 2012]). Thus it is not implausible that decision-theoretic moral theories provide a standard of subjective rightness that agents like us cannot systematically live up to. One might argue that if the best feasible implementation deviates from the decision-theoretic ideal in certain situations, then what should (subjectively) be done in those situations is what the best implementation would do. I don't think this is correct, but even if it were, the resulting notion of subjective moral status would still be approximately decision-theoretic, so it would still vindicate decision-theoretic considerations in moral theories.

Some authors interpret the guidance requirement in a very different way. On that interpretation, a theory is action guiding only if it specifies a concrete recipe for conscious deliberation that agents are supposed to follow when facing a choice. Decision theory is sometimes thought to put forward such a recipe, along something like the following lines: "First, create a mental list of all available options. For each option, consider all possible profiles that may come about by choosing the option. Multiply the moral value of each profile by its probability conditional on the option, and add up the products. The result is the expected value of the option. Then choose the option with greatest expected value." But this is a misunderstanding. Properly understood, decision theory is not about how agents should go about making choices. It does not prescribe a particular decision procedure, nor does it say that people should treat expected utility maximization as a goal. Decision theory merely specifies *which* options an agent should choose, given their

problem if you are unsure whether it would be right or wrong to flip the switch? From the internal perspective of moral theories, the answer is usually clear. If a theory says that it would be wrong to flip the switch, it will presumably say that it would be wrong to do so no matter if the agent knows that it is. Merely believing that something is right or wrong doesn't make it so. (I don't rule out that a moral theory might have special rules for cases of moral uncertainty, although that won't help agents who are uncertain whether the theory is true.)

goals and information, not *how* they should choose these options.[21]

To be sure, we often face a choice not only about how to act, but also about how to decide how to act: whether to trust our gut, to follow some heuristics, or to draw up a decision matrix. Decision theory can be applied to these choices. It will recommend whatever method has greatest expected utility. Drawing up a complete decision matrix will very rarely come out as the best procedure (even if it is an available option, as it usually isn't), and so decision theory will generally advise against it.

In the same way, I take it that moral theories deal in the first place with choices between acts – say, whether or not one should flip the switch in a given trolley problem with imperfect information. The answer will leave open what mechanisms or procedures the agent is supposed to use in order to make her choice. Decision-theoretic moral theories (of the kind I advocate) therefore do not imply that agents should be actively "guided" by the principle to maximize expected moral value, in the sense of [Smith 2012].

Again, we can also consider choices between decision procedures, but these don't get a special treatment. If an agent faces a choice between deciding by different decision methods, decision-theoretic moral theories will rank the methods by their expected moral value. The best method will depend on the situation: sometimes it may be to follow one's gut, other times to apply some heuristic, and occasionally it may involve setting up a simplified decision matrix.[22]

So decision theory does not prescribe any general, all-purpose procedure or recipe for deliberation. I don't think this is a flaw. There simply is no such procedure. The best procedure depends on contingent facts about the decision situation and the agent's cognitive system. If an agent happens to know that she has infallible moral intuitions which always favour acts with greatest expected value, she does well to follow those intuitions. If an agent is very good at maths and at introspecting her degrees of belief, computing expected values may be a useful procedure. If a decision has to be made within milliseconds, different procedures are advisable than if the agent has months to prepare. No single recipe could fit all decision-makers under all circumstances.[23]

21 This is implied by the orthodox approach to decision theory on which decision theoretic norms are ultimately norms on preferences, as well as by the functionalist accounts mentioned in note 5 above. It is also frequently emphasized by decision theorists; see, among many others, [Pettit 1991: 167–169], [Jackson 1991: 468–471], [Joyce 1999: 80].

22 The decision procedure that maximizes expected moral value can lead to a choice that doesn't maximize expected moral value. In that case, the agent did one thing right and another thing wrong: she chose the subjectively right decision procedure, but the subjectively wrong act.

23 Even a very complicated disjunctive recipe, such as the one outlined in [Smith 2010], is problematic. For example, applying it may often take too long. Smith's recipe is more plausible not as a single disjunctive procedure, but as a disjunction of simpler procedures, applicable to different decision situations. In effect, decision-theoretic moral theories also determine such a disjunctive assignment, insofar as that they also recommend different procedures for different situations. Moreover, they answer a question that becomes pressing in any account along Smith's lines: why exactly is procedure

You might object that if an agent doesn't know, and can't compute, the expected values, it is of little help to say that she ought to use whatever decision procedure maximizes expected value. But recall that when I say that the agent should use whatever procedure maximizes expected value, I do not mean that she ought to explicitly compute the expected moral value of the available procedures. As before, decision theory merely prescribes a choice, not a procedure to arrive at that choice. (It just happens that the choice here is a choice between procedures.) You may still press your point: how in practice is the agent supposed to figure out what procedure to use? The question makes sense as a question about cognitive implementation: how might the agent's cognitive system end up making the decision-theoretically optimal choice? As I said, this is a good question, on which artificial intelligence and cognitive science have made considerable progress. But it is not a question for ethics. Alternatively, the question could be meant as a question in applied ethics: what would be the morally right way for the agent to figure out a decision procedure? As such, the question presupposes that the agent could in principle use a variety of ways – that she faces a "third-order" choice between different ways to arrive at a (second-order) decision procedure to choose between (first-order) acts. Unsurprisingly, decision theory says that she ought to choose whatever second-order procedure maximizes expected moral value. And so on, for fourth-order choices and fifth-order choices. However, the hierarchy can't go on forever: at some point, the decision process must get started. The conscious mind is not an unmoved mover. At the highest level, there is no longer a moral question about how the agent is supposed to determine the right options.

References

Maurice Allais [1953]: "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine". *Econometrica: Journal of the Econometric Society*: 503–546

Allen R. Bernstein and Frank Wattenberg [1969]: "Non-Standard Measure Theory". In W. Luxemburg (Ed.) *Applications of Model Theory of Algebra, Analysis, and Probability,* Holt, Reinhart and Winston

Simon Blackburn [1998]: *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press

Ethan Bolker [1966]: "Functions Resembling Quotients of Measures". *Transactions of the American Mathematical Society*, 124: 293–312

X adequate under conditions Y? Is that a brute moral fact? In the decision-theoretic framework, we can answer that X is the right procedure under conditions Y because it then maximizes expected moral value.

John Broome [1991]: *Weighing Goods*. Oxford: Blackwell

Ruth Chang [2002]: "The Possibility of Parity". *Ethics*, 112(4): 659–688

Mark Colyvan, Damian Cox and Katie Steele [2010]: "Modelling the Moral Dimension of Decisions". *Nous*, 44(3): 503–529

James Dreier [1996a]: "Accepting Agent-Centered Norms: a Problem for Non-cognitivists and a Suggestion for Solving It". *Australasian Journal of Philosophy*, 74: 409–421

— [1996b]: "Rational Preference: Decision Theory as a Theory of Practical Rationality". *Theory and Decision*, 40: 249–276

Daniel Ellsberg [1961]: "Risk, Ambiguity and the Savage Axioms". *Quarterly Journal of Economics*, 75: 643–669

Heidi Li Feldman [1995]: "Science and uncertainty in mass exposure litigation". *Texas Law Review*, 74: 1–48

Samuel Gershman and Nathaniel D. Daw [2012]: "Perception, action and utility: the tangled skein". In M. Rabinowich, K. Friston and P. Varona (Eds.) *Principles of Brain Dynamics: Global State Interactions,* Cambridge (MA): MIT Press, 293–312

Joshua Gert [2004]: "Value and Parity". *Ethics*, 114: 492–520

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 [1978]: "Bayesian Decision Theory and Utilitarian Ethics". *The American Economic Review*, 68(2): 223–228

Frank Jackson [1991]: "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection". *Ethics*, 101(3): 461–482

Frank Jackson and Michael Smith [2006]: "Absolutist Moral Theories and Uncertainty". *The Journal of Philosophy*, 103: 267–283

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

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

Seth Lazar [Unpublished]: "In Dubious Battle: Uncertainty and the Ethics of Killing"

David Lewis [1974]: "Radical Interpretation". *Synthese*, 23: 331–344

— [1979]: "Attitudes *De Dicto* and *De Se*". *The Philosophical Review*, 88: 513–543

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

— [1989]: "Dispositional Theories of Value". *Proceedings of the Aristotelian Society*, 63: 113–137. Reprinted in Lewis's *Papers in Ethics and Social Philosophy*, Cambridge: Cambridge University Press, 2000

Christopher J.G. Meacham and Jonathan Weisberg [2011]: "Representation theorems and the foundations of decision theory". *Australasian Journal of Philosophy*, 89(4): 641–663

Philip Pettit [1991]: "Decision Theory and Folk Psychology". In M. Bacharach and S. Hurely (Eds.) *Foundations of Decision Theory: Issues and Advances,* Cambridge (MA): Blackwell, 147–175

Douglas W. Portmore [2008a]: "Are Moral Reasons Morally Overriding?" *Ethical Theory and Moral Practice*, 11(4): 369–388

— [2008b]: "Dual-ranking act-consequentialism". *Philosophical Studies*, 138(3): 409–427

Wlodek Rabinowicz [2008]: "Value Relations". *Theoria*, 74(1): 18–49

Frank Ramsey [1931]: "Truth and Probability (1926)". In R.B. Braithwaite (Ed.) *Foundations of Mathematics and other Essays,* London: Routledge & P. Kegan, 156–198

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

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

Holly Smith [2010]: "Subjective Rightness". *Social Philosophy and Policy*, 27(02): 64–110

— [2012]: "Using Moral Principles to Guide Decisions". *Philosophical Issues*, 22(1): 369–386

Jordan Howard Sobel [1983]: "Expected utilities and rational actions and choices". *Theoria*, 49: 159–183. Reprinted with revisions in [Sobel 1994: 197–217]

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

Robert Stalnaker [1984]: *Inquiry*. Cambridge (Mass.): MIT Press

Amos Tversky [1975]: "A critique of expected utility theory: Descriptive and normative considerations". *Erkenntnis*, 9(2): 163–173

John von Neumann and Oskar Morgenstern [1944]: *Theory of games and economic behavior*. Priceton: Princeton University Press

Paul Weirich [1983]: "A decision maker's options". *Philosophical Studies*, 44(2): 175–186