CHAPTER 33

PROBABILITY IN ETHICS

DAVID MCCARTHY

Ethics is mainly about what we ought to do, and about when one situation is better than another. But facing uncertainty about the consequences of our actions, and about how situations will evolve, is an all-pervasive feature of our condition. Should this not be a central topic in ethical theory?

Probability is by far the best-known tool for thinking about uncertainty, a well-known aphorism telling us that it is the very guide to life. But despite important exceptions, it is easy to get the impression that mainstream moral philosophy has not been much concerned with probability.

This reflects what seems to be a natural division of labour. The most fundamental questions for ethical theory seem to arise in the absence of uncertainty. For example, it seems hard to believe that the questions of whether it is better to give priority to the worse off, and of whether we ought to favour our nearest and dearest, have anything to do with uncertainty. Many influential discussions of these topics never mention uncertainty.

Of course, once answers to these fundamental questions are in, we can try to extend them to cases involving uncertainty. But ethical theorists may seem well advised to hand this task over to others, given how mathematical the various disciplines concerned with probability have become. Technically and philosophically interesting as it may be, the extension of central ethical ideas to problems involving probability seems to be outside the main business of ethical theory.

This chapter will argue for the opposite view. The major ethical problems to do with probability involve very little mathematics to appreciate; many topics which do not seem to have anything to do with probability are arguably all about probability; and thinking about various problems to do with probability can help us solve analogous problems which do not involve probability, sometimes even revealing that popular positions about such problems are incoherent.

Almost every topic discussed here could easily be given its own survey article, and an adequate bibliography would exceed the space allotted for the whole chapter. Positive positions are often argued for sketchily, many important positions on each topic are neglected, and some major topics are not discussed at all.

Instead, the goal is to offer enough breadth to illustrate some ways in which questions about probability run systematically throughout ethical theory, while in places going into

enough depth to articulate some surprising and potentially important applications. In brief, what follows is much less a survey of or an argument for particular positions than a plea for ethical theory to take probability more seriously.

I said that ethics is largely about what we ought to do, and when one situation is better than another. Some say that rationality is about these things as well. Given that theories of rationality in the face of uncertainty are highly developed, it might be thought that an appeal to these theories of rationality straightforwardly solves ethical problems about probability.

This line of thought is importantly mistaken. First, Hume famously claimed that it is not irrational for an agent to prefer the destruction of the whole world to the scratching of his finger. Nor would it be irrational for the agent to bring about the destruction to avoid the scratching. But the destruction is neither better than the scratching, nor better for the agent. And the agent surely ought not to bring about the destruction. On at least one widely-held view, therefore, ethics and rationality are not about the same things.

Secondly, it is undeniable that contemporary theories of rationality are an indispensable resource for thinking about ethics and probability. However, whether and how to apply these theories to ethics is far from straightforward, and will be one of the principal concerns of this chapter. Furthermore, in my view, at least, appeals to rationality are almost always epiphenomenal. For example, suppose we have a convincing argument for the claim that rational preferences have such and such a structure. We could then try to claim that an evaluative relation like betterness has to have that structure on the grounds that a rational agent can surely prefer what's better to what's worse. However, it is almost always less committal and more direct just to modify the original argument to make it apply directly to the structure of the evaluative relation. Claims about rationality often have historical priority over parallel claims about ethics, but I believe they do not have any kind of important conceptual priority.

The chapter starts with four sections which discuss which probabilities are relevant to ethics, establish terminology, and rehearse expected utility theory. It then turns to the evaluative question of when one situation is better than another, focusing on the question of when one distribution of goods is better than another. Sections 33.5 and 33.6 discuss popular but I think inadequate approaches to this question. These serve as a backdrop to a hugely important theorem due to Harsanyi (1955) introduced in section 33.7. Sections 33.8 to 33.16 discuss such things as the relationship between Harsanyi's theorem and utilitarianism; criticisms of Harsanyi's premises and the relationship of these criticisms to other distributive views such as egalitarianism, the priority view, and concerns with fairness; the extension of Harsanyi's theorem to problems of population size; incommensurability; continuity; non-expected utility theory; evaluative measurement; and the question of what Harsanyi's theorem really shows about aggregation. These sections also list various open problems and directions for further work. All of these topics have to do with probability.

One of the benefits of thinking about Harsanyi's theorem is the way it helps us organize our thinking about all sorts of fundamental evaluative questions. Section 33.17 will suggest that thinking about decision theory can have the same value in thinking about fundamental normative questions, questions about what we ought to do. With particular focus on probability, the remaining sections illustrate by discussing what are arguably the three most important kinds of normative theories: act consequentialism, rule consequentialism or contractualism, and deontology (these will be defined in section 33.3).

The discussion aims to be self-contained. For those with a background in ethics who would like to know more about how probability is involved, the chapter keeps technicalities

to a minimum. But the topic just cannot be addressed without a certain amount of rigor, and passing acquaintance with expected utility theory and decision theory will be helpful, though not strictly necessary. For those who know about probability and would like to see how it applies to ethics, the chapter gives brief guides to the relevant ethical debates. Such readers will recognize occasional allusions to relatively sophisticated ideas to do with probability. For one thing is clear: the questions about probability which ethics raises are profound, and are surely best addressed by combining expertise.

33.1 Probabilities

One difficulty in thinking about probability in ethics is assessing when ethicists need to be involved. Suppose we are told that some action will benefit many but involves a small probability of harming a few. We might think it the job of epistemologists, metaphysicians or philosophers of science to tell us what kind of judgment 'the probability is small' expresses, what laws probabilities obey, and what makes such a judgment correct. Ethicists need only ask whether we ought to perform the action given that the probability of harm is small, and need not be involved any earlier. However, the division of labour is unlikely to be so neat.

There are many conceptions of probability (see e.g. Hájek, 2012, for a survey). This raises the question of which conception is most relevant to ethics, or whether different conceptions are appropriate in different ethical contexts. One of the most basic distinctions is between subjective and objective conceptions of probability, and this distinction will enable us to illustrate many of the issues.

The best-known subjective conception claims that the preferences of an ideally rational agent between uncertain prospects must satisfy various structural conditions (Ramsey, 1931; Savage, 1954). Suppose the agent also has a rich set of preferences. Then Ramsey and Savage showed that there exists a unique function on events satisfying the usual probability axioms (call it her subjective probability function) and a function on outcomes (her utility function) such that: the agent weakly prefers one prospect to another if and only if the former has at least as great expected utility, as calculated by those functions.

Perhaps the most prominent objective conception of probability in the contemporary debate is the best-system analysis pioneered by Lewis. The original best-system analysis of the laws of nature of Lewis (1973) says that the laws are the theorems of the best systematization of the world: the true theory which does best in terms of simplicity and strength (or informativeness). To allow probabilistic laws in, Lewis (1994) introduced the idea of fit. The more likely the actual world is by the lights of the theory, the better the fit of that theory. Theories are now judged according to how well they do in terms of simplicity, strength, and fit. If some of the laws of the best theory are probabilistic, those are what determine the objective probabilities.

Suppose we have to choose between subjective and objective conceptions for use in ethics, understood along the lines just sketched.[1] Which conception should it be?

[1] Neither the Ramsey-Savage story about subjective probabilities nor Lewis's version of best-system analysis has a hegemony. For surveys of alternative views about subjective probability, see Gilboa (2009), and for alternative best-system analyses, see Schwarz in this volume (2016).

Perhaps it depends on context: for example, subjective probabilities may be appropriate for agent-evaluation (blame, responsibility etc.), but inappropriate in other contexts. But let us fix the context by focussing on the most basic normative question of what we ought to do.

Each conception has features we might find appealing. Objective probabilities seem in some important sense to trump subjective probabilities. This is reflected in the popular view that when an agent has beliefs about objective probabilities, rationality requires her to conform her subjective probabilities to those beliefs. This is the basic idea behind the so-called principle principal of Lewis (1980). But if objective probabilities do indeed trump subjective probabilities, it may seem that what we ought to do depends on the objective probabilities, not our subjective probabilities.

On the other hand, objective probabilities may be disappointingly sparse or epistemically inaccessible. For example, best-systems analyses may make good sense of the objective probability of radium atoms decaying or coins landing on heads. But it is much less clear what best-systems analyses have to say about the objective probability of events like a run on a particular bank next year, one-off macro events involving chaotic systems. Such events may fail to have reasonably determinate objective probabilities (compare Hoefer, 2007), and even if they do, the epistemology may be too difficult for the objective probabilities to be usefully action guiding.

So perhaps we should instead say that what we ought to do depends at least in part on our subjective probabilities. One option is to use subjective probabilities exclusively; another is to use objective probabilities where available, and subjective probabilities to fill in the gaps. But every view which makes significant use of subjective probabilities faces at least two major problems.

First, the Ramsey-Savage story about subjective probabilities is a chapter in the Humean story about rationality. But just as the Humean story refuses to condemn the preference for the destruction of the whole world over the scratching of a finger, the Ramsey-Savage story does not condemn subjective probabilities which, to most people, are just as crazy. For example, provided her preferences are appropriately structured, there is nothing in the Ramsey-Savage story to condemn someone who thinks it highly likely the world will come to an end before teatime. Such subjective probabilities will seem to many too irrational to have any bearing on what we ought to do. But it is a major challenge to articulate a principled account of which subjective probabilities should be excluded.

Secondly, as soon as we allow in subjective probabilities, we face questions of *whose* and *how*. Whose subjective probabilities count in determining whether an agent ought to perform some action – the agent's, those of her potential victims or beneficiaries, everyone's? If the subjective probabilities of at least two people are relevant, how should they be used? At least if we switch to the problem of evaluating the uncertain prospects which actions result in, this is a long-standing problem in welfare economics. The so-called *ex post* approach recommends first aggregating the separate subjective probability functions into a single social probability function, then using this social probability function to evaluate uncertain prospects. The *ex ante* approach gives the separate subjective probability functions a direct evaluative role, at least in a special case. Just to give one version, *ex ante Pareto* says: if for each individual i, an uncertain prospect P is better for i than another uncertain prospect P' relative to i's own subjective probability function, then P is better than P'. Both the *ex post* and *ex ante* approaches look appealing, but they are extremely difficult to combine consistently. For example, given weak assumptions, there will be prospects P and P' such

that *ex ante* Pareto has the apparent pathology of implying that P is better than P' despite the fact that P' is guaranteed to produce a better outcome than P. But *ex post* approaches will adopt principles which from the outset say that in such cases, P' is better than P.[2]

Now it is not my goal to try to answer any of the large questions raised in this section. My claim is rather that they are questions with which ethicists must engage, and that one's answers to these questions may depend on one's more general ethical views. To illustrate, suppose one sees ethics as being primarily about coordinating action to achieve good outcomes, and one is prepared to tolerate a significant amount of indeterminacy in one's normative theory. Then one may be tempted to claim that the probabilities which are relevant to ethics are the objective probabilities alone. By contrast, suppose one instead sees ethics as being about trying to achieve some sort of fair compromise between agents with diverse beliefs and goals. Then it may seem tempting to allow in subjective probabilities no matter how irrational, and to follow the *ex ante* approach. On this picture, individual autonomy is central, and it may seem more important to respect the notion of unanimity built into *ex ante* Pareto than to try to avoid the apparent pathology which comes with it. There are, of course, many other options, but the important point is that which probabilities are relevant to ethics, and how, is itself a fundamental ethical question.

33.2 Outcomes

Some writers, however, think that probabilities are never relevant to what we ought to do. A parallel view applies to the question of when one uncertain prospect is better than another. Jackson (1991) illustrates with the following. A doctor has to choose between three treatments for a patient with a minor complaint. Drug A would partially cure the complaint. One of drugs B and C would completely cure the patient while the other would kill him, but the doctor cannot tell which is which.

The obvious view, as Jackson notes, is that the doctor ought to give the patient drug A. This verdict would be delivered by any broadly decision-theoretic account. Along similar lines, the prospect associated with giving the patient drug A is better than the prospects associated with drugs B and C. Call any view which assesses actions and prospects involving uncertainty along broadly decision-theoretic lines *probability-based*.

But there is a different view: if drug B would cure the patient, the doctor ought to give the patient drug B; similarly for drug C. Likewise, if drug B would cure, the prospect associated with giving drug B is better than the other prospects. Call such views, positions which assess actions and prospects in terms of what their consequences would be, *outcome-based*.

As the drug example shows, an objection to outcome-based views is that they make the truth about what we ought to do too epistemically inaccessible, or provide poor guides to action. But there are at least two interesting arguments for outcome-based views.

First, transposing an argument due to Thomson (1986) to the present example, suppose the pharmacist walks in and knowing full-well that drug B would cure the patient, says to

[2] The large literature on this topic is rather technical, but Broome (1991, ch. 7) provides a good introduction and philosophical discussion. Mongin (1995) contains a very general set of results.

the doctor: "You ought to use drug *B*". The pharmacist seems right. But doesn't that imply an outcome-based view?

In response, consider the case where the pharmacist says: "Drug *B* would cure. So you ought to use drug *B*". By the time the pharmacist has finished the first sentence, the doctor has new evidence, and should upgrade her probabilities accordingly. There is then no clash between a probability-based view and the truth of the pharmacist's second sentence. Likewise, I think that in the actual case something like "Drug *B* would cure" is implied when the pharmacist just says: "You ought to use drug *B*". What is implied by the pharmacist's normative assertion impacts upon the probabilities the doctor should have, making the literal construal of the normative assertion true (McCarthy, 1998).

Secondly, advocates of probability-based views have to say which probability functions are relevant to what we ought to do. But there are many candidates, e.g. the probabilities of this agent or that agent, at this time or that time. Jackson concludes that we have to recognize the existence of "an annoying profusion ... of a whole range of oughts" (Jackson, 1991, p. 471).

But this seems dissatisfying. When we ask ourselves or others what we ought to do, we don't want to learn that some oughts recommend this while others recommend that. We want to know what we ought to do *fullstop*. But if there is only one ought, we need to privilege one probability function. The function of an omniscient agent may seem to be the only distinguished choice, so we end up with an outcome-based view.

In response, just because it is not obvious which probability function is privileged, it does not follow that no function (or reasonably narrow class of functions) is privileged. In the previous section we saw that if we adopt a probability-based view, a variety of fairly fundamental ethical factors and disputes bears upon the question of which probabilities are relevant to ethics. The complexity of this topic explains why it is not obvious which probability function is privileged, but the fact that the problem is complex hardly entails that some outcome-based view wins by default. Outcome-based views have to be assessed in terms of various ethical desiderata just as much as probability-based views do, and they do quite badly in terms of desiderata such as the idea that an ethical theory should be suitably action-guiding.

It is also worth noting that outcome-based views may result in large-scale indeterminacy. The drug example stipulated that various counterfactuals relating actions to outcomes are true. But an increasingly popular view claims that most counterfactuals are false (see e.g. Hájek, 2014). In particular, it will often be the case that for some potential action *A* there is no outcome *O* such that the counterfactual: "If *A* were performed, *O* would result", is true. On this view about counterfactuals, the facts on which outcome-based views have to call are much sparser than might have appeared, with the result that there is a lot more evaluative and normative indeterminacy on outcome-based views than we might have hoped. This may further undercut the appeal of outcome-based views.

In what follows, I will assume that some probability-based view is correct. But it is a major question which conception of probability is relevant to ethics, so ethicists need to be involved with questions about probability early on. In light of the difficulties of aggregating probability functions alluded to in the previous section, ethicists also need to be prepared for the possibility that the eventual input into ethics is going to be messier than a single probability function which satisfies the usual axioms.

33.3 Terminology

However, to simplify I henceforth assume that probabilities are supplied and satisfy the usual axioms. To reflect this I will often speak of *risk* rather than probability or uncertainty. A *lottery* over a nonempty set of world histories (past, present and future) assigns positive probabilities to finitely many of the histories with the probabilities all summing to one (these are sometimes known as lotteries with *finite support*). I will often write lotteries in the form $[p_1,h_1;\ldots;p_m,h_m]$ where the h_j's are the histories which could result from the lottery and the p_j's their probabilities.

The *betterness relation* holds between two lotteries just in case the first is at least as good as the second. An individual *i*'s *individual betterness relation* holds between two lotteries L_1 and L_2 just in case: *i* exists in every history which could result from the lotteries, and L_1 is at least as good for *i* as L_2. By identifying histories with lotteries in which the history gets probability one, and restricting the betterness and individual betterness relations to such lotteries, we obtain relations between histories. I will refer to these relations as *risk-free* versions of the originals. For example, the *risk-free betterness relation* holds between two histories just in case the first is at least as good as the second.

There are many views about when one history is better for someone than another, or in a more suggestive phrase, about what makes someone's life go best (Parfit, 1984, Appendix I). On one popular classification, the three main views are that having a good life is a matter of: (i) having good-quality experiences; (ii) satisfying one's preferences or desires; or (iii) attaining what are said to be objective goods, such as deep knowledge or close personal relationships. However, some philosophers think that when doing ethics, we should not be in the business of making fine-grained comparisons between different people's lives, but should make interpersonal comparisons only in terms of such things as the resources, freedoms, or opportunities people enjoy (see e.g. Rawls, 1982; Sen, 1985). Which of these views is correct will not matter in what follows, but it will be important that the discussion can accommodate any of them.

We will be talking a lot about the betterness relation. Not everyone thinks that this is a useful way of looking at ethics (see e.g. Foot, 1985; Thomson, 2001). But in response, talking about betterness can be seen as a harmless organizing tool (see e.g. Broome, 1991), and is popular enough for us to be able to cover many major positions. For example, *consequentialism* (on a probability-based interpretation) is the view that lotteries can be ranked in terms of betterness, and that betterness somehow determines normativity.[3] For example, *act consequentialism* says that we always ought to bring about the best available lottery, whereas *rule consequentialism* says that we always ought to act according to the rule such that, if everyone acted in accord with it (or on a different version, accepted it), the best available lottery would be realized. *Contractualism* tends to be framed not in terms of betterness, but in terms of an ideal social contract. However, when it comes to the assessment of different social contracts, contractualists are concerned with

[3] As far as I can see, there is no universally accepted account of consequentialism, so I am only trying to convey the rough idea rather than provide a precise definition. In addition, the way moral philosophers use the term 'consequentialism' should not be confused with an important decision-theoretic idea which also goes by the name of 'consequentialism' (see e.g. Hammond, 1998).

competing sets of principles or rules (see e.g. Scanlon, 1982), so at the concrete level of normative theorizing, it is often hard to tell the difference between contractualism and rule consequentialism. Finally, *deontology* is often characterized as the position that some acts are wrong even when they would have the best available consequences, such as killing one innocent person to prevent five innocent people from being killed.

33.4 Expected utility theory

This chapter expresses the view that whatever one ultimately makes of expected utility theory and decision theory, looking at basic evaluative and normative questions through the frameworks they provide is extremely useful. This section therefore provides a quick rehearsal, first of the terminology of expected utility theory, and then of its most basic result. It takes X to be some fixed nonempty set. In applications, X will usually be a set of histories, or more colloquially, outcomes.

A *preorder* on X is a binary relation R on X which is reflexive ($\forall x \in X$, xRx) and transitive ($\forall x,y,z \in X$, xRy & $yRz \implies xRz$). It is *complete* if for all $x,y \in X$, either xRy or yRx. It is *incomplete* just in case it is not complete. An *ordering* of X is just a complete preorder of X. If L and M are lotteries over X, then for all $\alpha \in (0,1)$, $\alpha L + (1-\alpha)M$ is the so-called *compound lottery* in which each member x of X has probability $\alpha p + (1-\alpha)q$ where p is x's probability under L and q is its probability under M. Suppose that $\succsim$ is an ordering on X. Then a real-valued function f is said to *represent* the ordering just in case: for every x and y in X, $x \succsim y$ if and only if $f(x) \geq f(y)$.

Suppose that $\succsim$ is a binary relation on lotteries over X. Here are the three expected utility axioms.

Ordering $\succsim$ is a complete preorder.

Strong Independence For all lotteries L, M and N, and $\alpha \in (0,1)$: $L \succsim M$ if and only if $\alpha L + (1-\alpha)N \succsim \alpha M + (1-\alpha)N$.

The rough idea of Strong Independence is that the "addition" of the same lottery N to either side of $L \succsim M$ should make no difference: the added N's will cancel out. Strong Independence is sometimes explained by imagining that the compound lotteries will be realized by first tossing a biased coin, where heads has a probability of α and tails a probability of $1-\alpha$, then running whichever lottery results. For example, suppose you strictly prefer L to M, and you now have to decide between $\alpha L + (1-\alpha)N$ and $\alpha M + (1-\alpha)N$. If the coin lands on tails, you will face N in either case, so in that scenario there is nothing to choose between the two compound lotteries. But if the coin lands on heads, you will face L or M, and will therefore prefer to have chosen $\alpha L + (1-\alpha)N$ to $\alpha M + (1-\alpha)N$. Since heads has a positive probability, you should therefore strictly prefer $\alpha L + (1-\alpha)N$ to $\alpha M + (1-\alpha)N$ prior to the coin being tossed. Or at least that is one of the typical ways of motivating Strong Independence. The example has focused on preference relations, but it can clearly be applied directly and without any discussion of rationality to a variety of evaluative comparatives, such as betterness and individual betterness relations.

Continuity For all lotteries L, M and M such that $L \succ M \succ N$,[4] there exist $\alpha, \beta \in (0,1)$ such that $M \succ \alpha L + (1-\alpha)N$ and $\beta L + (1-\beta)N \succ M$.

To illustrate, suppose you strictly prefer \$1000 to \$100, and strictly prefer \$100 to \$10. Then if your preferences are continuous, there will be some lottery which almost guarantees you \$1000 with a tiny chance of \$10 (one in a billion, say) which you will strictly prefer to getting \$100 for certain. And you will strictly prefer \$100 for certain to some lottery which almost guarantees you \$10 with a tiny chance of \$1000. As the example is meant to suggest, many people think that Continuity is a plausible requirement on various evaluative comparatives.

A binary relation $\succsim$ on lotteries over X satisfies the expected utility axioms just in case it satisfies Ordering, Strong Independence, and Continuity. Here is the most basic result of expected utility theory, due to von Neumann and Morgenstern (1944), but anticipated in a deeper way by Ramsey (1931).

Theorem 1 *(von Neumann and Morgenstern) Let X be a nonempty set, and $\succsim$ be a binary relation on lotteries on X which satisfies the expected utility axioms. Then there exists a real-valued function u on X such that*

1. *For all lotteries $L_1 = [p_1, x_1; \ldots; p_m, x_m]$ and $L_2 = [q_1, y_1; \ldots; q_n, y_n]$,*

$$L_1 \succsim L_2 \iff p_1 u(x_1) + \cdots + p_m u(x_m) \geq q_1 u(y_1) + \cdots + q_n u(y_n)$$

2. *Any function v satisfies (i) when substituted for u if and only if there exist real numbers $a > 0$ and b such that $v = au + b$.*

Roughly speaking, (i) says that there is a function u (often referred to as a "vNM utility function") such that $L_1 \succsim L_2$ if and only if the expected value of u associated with L_1 is at least as great as the expected value of u associated with L_1. The expected value of u associated with a lottery is obtained by applying u to each of the lottery's possible outcomes, weighting the result by the probability of those outcomes, then adding all those numbers up. In such circumstances, I will say that the ordering $\succsim$ is *represented by the expected value of u*. (ii) says that the function u is unique up to choice of zero and unit, or in fancier terminology, unique up to positive affine transformation. For an analogy, Fahrenheit and Centigrade measure temperature in essentially the same way, except that they use different zeros and units. Overall, the main message is that if an ordering of lotteries satisfies the expected utility axioms, it can be represented by the expected value of some function which is more or less unique.

The literature on expected utility theory is vast. It has been applied to all sorts of topics, and has received a great deal of defense, criticism, and mathematical elaboration.[5] Beyond a few remarks, this chapter will assume some sort of familiarity with the defense, but will rehearse many of the criticisms, particularly as they apply to ethics. We now need to ask: When is one lottery better than another? Which lotteries ought we to bring about? We begin with the first question.

[4] $L \succ M$ is defined as $L \succsim M$ and not $M \succsim L$. $L \sim M$ is defined as $L \succsim M$ and $M \succsim L$.

[5] At varying levels of philosophical and mathematical ambition, personal favourites include Fishburn (1970), Resnik (1987), Kreps (1988), Broome (1991), Hammond (1998), Ok (2007) and Gilboa (2009). In this volume, see Buchak (2016).

33.5 Expected goodness

Some philosophers imply that that if we know when one history is better than another, the question of when one lottery is better than another is straightforward. For example, Parfit (1984, p. 25) and Scheffler (1982, p. 1, note 2) start their discussions of consequentialism only by assuming

(1) The risk-free betterness relation is an ordering.

To cover risky cases, they think that we need to appeal only to expected utility theory. In particular, they think we just need to add

(2) One lottery is at least as good as another if and only if its expected goodness is at least as great.

In other words, the betterness relation is represented by the expected value of goodness. Parfit and Scheffler are not claiming that it is obvious when one history is better than another. Rather, they are claiming that once we have an ordering of histories in terms of betterness, (2) then tells us how to order lotteries in terms of betterness.

Now Parfit and Scheffler are quite brief about this and their real concerns lie elsewhere. But this sort of claim is commonly made, and it is important to realize that it contains a serious mistake. The basic difficulty is that (2) presupposes the existence of goodness measures, measures of how good histories are, and various problems arise depending on where we think these measures are coming from.

First, provided certain technical conditions are met, (1) guarantees that the risk-free betterness relation can be represented by some function.[6] To deal with the possibility that there may be more than one such function, we might treat the set of all goodness measures as the set of all of the functions which represent the risk-free betterness relation. It would then be natural to interpret (2) as saying: $L_1 \succsim L_2$ if and only if the expected goodness of L_1 is at least as great as the expected goodness of L_2 according to every goodness measure. Unfortunately, however, this approach leads to massive indeterminacy. An example will illustrate. Suppose there are exactly three histories x, y and z, ordered $x \succ y \succ z$ by the risk-free betterness relation. Let L be the lottery $[\frac{1}{2},x;\frac{1}{2}z]$ and let us consider how it compares with y. Consider the two functions u and v defined by $u(x) = v(x) = 1$, $u(y) = 0.9$, $v(y) = 0.1$, and $u(z) = v(z) = 0$. Both of these functions represent the risk-free betterness relation, and therefore count as goodness measures on the current proposal. But according to u, the expected goodness of L is less than that of y, and according to v, the expected goodness of L is greater than that of y. The current proposal therefore leaves L and y unranked, and it only takes a bit more work to show that this will be true of almost every pair of lotteries. So interpreting (2) along these lines does almost nothing to cover risky cases.

Secondly, to get around this problem we might hope to narrow down all of the functions which represent the risk-free betterness relation to (essentially) a single function to be used

[6] The result goes back to Cantor; for details, see any reasonably advanced book on utility theory, such as Kreps (1988) or Ok (2007).

as a goodness measure.[7] This line of thought is tacitly quite common, and what tends to happen is that one of the functions which represents the risk-free betterness relation seems quite simple or natural, and it is taken to be the goodness measure.[8] An old idea will illustrate. According to this idea, each "just noticeable difference" between outcomes is given the same magnitude of goodness, so that the difference in goodness between the best outcome and the second best outcome is equal to the difference in goodness between the second best outcome and the third best outcome, and so on.[9] In the toy example of the previous paragraph, this would be done by a function w where $w(x) = 1$, $w(y) = 0.5$, and $w(z) = 0$. Using (2) would then provide a ranking of all lotteries in terms of betterness. For example, L and y would turn out to be equally good. However, this proposal is ethically entirely arbitrary, and it is easy to invent circumstances in which the method delivers implausible conclusions. To illustrate, let us apply the same idea to individual betterness relations. Consider a wine connoisseur who is able to discriminate among a vast number of wines, and let us take her ordering of wines as given. Let a^+ be the outcome in which she gets the best possible wine, a the next wine down, r some rough house wine, and r^+ the next one up. The current method would regard the two lotteries $[\frac{1}{2}, a^+; \frac{1}{2}, r]$ and $[\frac{1}{2}, a; \frac{1}{2}, r^+]$ as equally good. But our connoisseur might regard experiencing the best possible wine as worth risking a lot for, and improving a rough house wine as hardly worth anything, leading her to conclude that the first lottery is better. But the current method woodenly regards the two lotteries as equally good.

Thirdly, one might approach the problem from a different direction. Suppose we start with a claim which is presupposed by (2), namely

Social EUT The betterness relation satisfies the expected utility axioms.

Now by the vNM theorem, Social EUT implies

(3) For some real-valued function on histories f, the betterness relation is represented by the expected value of f.

We might then define f as a goodness measure (along with its positive affine transformations). It follows that (2) now gives us the right results: one lottery is better than another just in case its expected goodness is greater. Unfortunately, however, just as the first method yielded almost complete indeterminacy, this method is almost completely uninformative. In almost all cases, it provides us with no concrete method of ranking lotteries. For example, in the toy example used to show why the first method leads to indeterminacy, it is consistent with the present method that L is better than y, that L and y are equally good, and that L is worse than y.

We have now looked at three ways of trying to fill in the story gestured towards by Parfit, Scheffler, and many others, the story which thinks that once we are given the risk-free

[7] More precisely, to a set of functions which are all related by positive affine transformation. The vNM theorem tells us that these will all be equivalent when it comes to ordering lotteries in terms of expected goodness.

[8] For example, McCarthy (2013) argues that this approach is common in accounts of the priority view and leads to unsatisfactory definitions of it.

[9] The basic idea goes back to Edgeworth (1881). For criticism and defense see e.g. Vickrey (1960) and Ng (1975) respectively.

betterness relation, we need only to appeal to expected utility theory to cover risky cases. Each attempt to say where goodness measures are coming from leads to a problem. The first leads to indeterminacy, the second to arbitrariness, and the third to uninformativeness. Now expected utility theory does indeed turn out to be a powerful tool for thinking about evaluative questions about risk, and even questions which do not seem to be about risk. But the story has to be more sophisticated than anything we have so far seen.

33.6 Veils of ignorance

To simplify, I will from now on assume that in evaluating lotteries, we are only concerned with the ethics of distribution, and in addition, not concerned with rights or responsibilities. In particular, I will assume: if h_1 and h_2 contain the same population and for each member i, h_1 is exactly as good for i as h_2, then h_1 and h_2 are equally good.

The best-known strategy for augmenting an appeal to expected utility theory is to use a so-called veil of ignorance, made famous but used in different ways by Harsanyi (1953) and Rawls (1971).

Assume a fixed population $1,\ldots,n$. Harsanyi's presentation of his argument tacitly identifies individual betterness relations with individual preference relations. But there are objections to that identification, and following Broome (1991) we can avoid them by restating Harsanyi's argument in terms of individual betterness relations. This enables us to leave it open whether the content of individual betterness relations has to do with preference satisfaction, the quality of experience, achievements, or some other account. Harsanyi's argument then begins with

Individual EUT Individual betterness relations satisfy the expected utility axioms.

Assume also that interpersonal comparisons are unproblematic in that

Interpersonal Completeness For all individuals i and j and histories h_1 and h_2, either h_1 is at least as good for i as h_2 is for j, or vice versa.

Together Individual EUT and Interpersonal Completeness imply that there are real-valued functions $u_1,\ldots,u_n$ on histories such that (i) for each individual i, i's individual betterness relation is represented by the expected value of u_i, and (ii) for all individuals i and j, h_1 is at least as good for i as h_2 is for j if and only if $u_i(h_1) \geq u_j(h_2)$. From now on, $u_1,\ldots,u_n$ will always be such functions, but their existence presupposes Individual EUT and Interpersonal Completeness. I will sometimes call them utility functions.

Harsanyi (1953) took ethics to be impartial.[10] But how should this be modeled, or made more concrete? This is where Harsanyi appeals to a veil of ignorance. Choosing under the

[10] Some of the arguments which follow make slightly stronger assumptions about interpersonal comparisons than I have made explicit. The point of these is to make various impartiality assumptions have an effect, and also to guarantee that the functions $u_1,\ldots,u_n$ are essentially unique, in that if some other set of functions $v_1,\ldots,v_n$ plays their role, there are real numbers $a > 0$ and b such that for all i, $v_i = au_i + b$. But I will suppress this slightly technical issue. For full details, see e.g. Broome (2004, p. 96).

equiprobability assumption is understood as choosing between two social situations on the assumption that one is equally likely to turn out to be each member of the population. Then Harsanyi took the idea that ethics is impartial to be well-modeled by

Veil of Harsanyi One lottery is at least as good as another if and only if it would be weakly preferred by every self-interested and rational person choosing under the equiprobability assumption.

I will skip the formal details, but from Individual EUT, Interpersonal Completeness and Veil of Harsanyi, Harsanyi gave a simple argument for

Sum The betterness relation is represented by the expected value of the function $u_1 + \cdots + u_n$.

Rawls (1971) agrees with Harsanyi that ethics is impartial, and that a veil of ignorance is a good way of modeling impartiality. To focus on their treatment of veils, we will ignore other differences, such as the different ways in which they understand interpersonal comparisons. With those aside, Rawls can be taken as agreeing with Individual EUT and Interpersonal Completeness. But his interpretation of the veil differs. Choosing under the *uncertainty assumption* is understood as choosing between two social situations on the assumption that one will turn out to be one of the members of the population, but with complete uncertainty about who that will be. Then Rawls took the idea that ethics is impartial to be well-modeled by

Veil of Rawls One history is at least as good as another if and only if it would be weakly preferred by every self-interested and rational person choosing under the uncertainty assumption.

Rawls then argued that Individual EUT, Interpersonal Completeness, and Veil of Rawls would result in

Maximin One history is better than another if and only if the former is better for the worst off.

Many commentators have thought Rawls should instead have concluded with

Leximin One history is better than another if and only if it is better for the worst off, or equally good for the worst off and better for the second worst off, and so on.

These arguments raise three basic questions: (i) What does rational choice under the uncertainty assumption really require? (ii) Given that one is going to model impartiality via some sort of veil of ignorance, is the uncertainty assumption a better way of doing it than the equiprobability assumption? (iii) Is modeling impartiality via a veil of ignorance a good idea anyway?

Briefly, (i) seems to be unclear. For example, suppose the Ramsey-Savage story is right about rational choice under conditions of uncertainty. For the agent behind the veil to lack implicit subjective probabilities of any degree of determinateness – and thus to model complete uncertainty – that story implies that her preferences are incomplete. At best, maximin (or leximin) would then seem to be but one rationally permissible

choice among many, whereas Rawls needs it to be rationally required (see Angner (2004) for further discussion). For (ii), the equiprobability assumption seems at first glance a reasonable attempt at giving impartiality a concrete and reasonably clear interpretation. Moreover, given the difficulties in understanding what rationality in conditions of complete uncertainty requires, it is hard to see what motivates shifting to the uncertainty assumption, aside from a question-begging attempt to avoid Sum. I will return to some of these issues, but the most fundamental question is (iii), and a later result of Harsanyi's seems to show that the use of veils of ignorance was never a good idea in the first place.

33.7 Harsanyi's theorem

To present Harsanyi's result we need to state two more premises. We continue to assume a fixed population. The first premise expresses a kind of impartiality.

Impartiality For all histories h_1 and h_2, if there is some permutation π of the population such that for each individual i, h_1 is exactly as good for i as h_2 is for $\pi(i)$, then h_1 and h_2 are equally good

The second premise is a so-called Pareto assumption.

Pareto (i) If two lotteries are equally good for each member of the population, they are equally good. (ii) If one lottery is at least as good for every member of the population and better for some members, then it is better.

This is Harsanyi's theorem. For an accessible proof, see e.g. Resnik (1987).

Theorem 2 *(Harsanyi) Assume a constant population. Then Individual EUT, Interpersonal Completeness, Social EUT, Impartiality, and Pareto jointly imply Sum.*

To recap what Sum says, the conclusion of the theorem says that one lottery is better than another just in case it has a greater sum of individual expected utilities. This implies that one history is better than another just in case it has a greater sum of individual utilities. However, in its classical form, utilitarianism is usually defined as the claim that one history is better than another just in case it has a greater sum of individual goodness. This raises the disputed question of what Sum has to do with utilitarianism, and thus whether Harsanyi's premises imply utilitarianism. Roughly speaking, Harsanyi's premises imply the classical version of utilitarianism just in case individual utilities are measures of individual goodness. Simplifying somewhat, Sen (1976) and Weymark (1991) denied that the two should be identified, whereas along with e.g. Harsanyi (1977b), Broome (1991), and Hammond (1991), I believe that they should be identified. I will say more about this in section 33.14, but the most important claim is that it does not really matter who is right. The conclusion of Harsanyi's theorem appears to tell us exactly what the content of the betterness relation is, and what name we should give to that conclusion is of much less importance.

In my view, it is hard to exaggerate the importance of Harsanyi's result. I will assume enough familiarity with expected utility theory, references to which were provided earlier,

to see the prima facie case for Individual EUT and Social EUT. The rough idea is that the prima facie case for rational preference relations satisfying the expected utility axioms can be modified to apply directly to evaluative relations like individual betterness relations and the betterness relation. The prima facie case for the other premises is fairly natural as well. The best way to explore this further will be to look at criticisms of the premises. We will do that shortly, but first I want to consider how Harsanyi's theorem improves on what we have seen so far.

The popular appeal to expected utility theory sketched in section 33.5 suffered from telling us little of any use about the betterness relation. But if we take individual betterness relations as given, and accept the premises of Harsanyi's theorem, the theorem shows that the content of the betterness relation is completely determined.

Consider now veil of ignorance arguments. Both Harsanyi's and Rawls's accept Individual EUT and Interpersonal Completeness. That leaves Harsanyi's veil argument with Veil of Harsanyi and Rawls's with Veil of Rawls, while Harsanyi's theorem is left with Social EUT, Impartiality, and Pareto.

Harsanyi's veil argument works by assuming that the person behind the veil is rational, and therefore has preferences which satisfy the expected utility axioms. Given that, Veil of Harsanyi yields Social EUT, and also, obviously, Impartiality and Pareto. So Harsanyi's veil argument enjoys no advantage over his theorem, and the theorem simply bypasses worries about veil arguments expressed by e.g. Scanlon (1982).

The comparison with Rawls is less clear. When discussing the veil, Rawls usually considers only the problem of ranking different histories. But someone behind the veil could also try to rank different lotteries (thus facing two forms of ignorance: uncertainty behind the veil, and risk beyond the veil). So we can ask what she thinks about Social EUT, Impartiality, and Pareto. It would be surprising if the uncertainty assumption led her to reject any of these claims, and thence Sum. But since Rawls is so plainly opposed to Sum, I think this suggests that aspects of his informal reasoning have not been fully captured in what seems to be his formal model. Sections 33.11 and 33.12 will discuss two major Rawlsian worries about some of Harsanyi's premises. But to foreshadow, these worries can be expressed directly as criticisms of the premises of Harsanyi's theorem, and appealing to the veil does not seem to add anything.

Finally, we will see in section 33.9 that there is at least one major view about the ethics of distribution which is impartial but is immediately ruled out by the adoption of a veil of ignorance, whether Harsanyi's or Rawls's. So much the worse for the veil as a model of impartiality. Thus in my view, the veil turns out to be just an unhelpful distraction, and the proper focus of attention for the ethics of distribution should be Harsanyi's theorem.

33.8 Variable populations

Before looking at various worries about and alternatives to the premises of Harsanyi's theorem, it is worth mentioning a way in which it can be extended. Problems where the population can vary are difficult. But we do not need to add much to the premises of Harsanyi's theorem to make progress.

The following says that only the kinds of lives people are living matters, not the identities of those people.

Anonymity For all histories h_1 and h_2 containing finite populations of the same size, if there is a mapping ρ from the population of h_1 onto the population of h_2 such that for every member i of the population of h_1, h_1 is exactly as good for i as h_2 is for $\rho(i)$, then h_1 and h_2 are equally good.

This premise makes the nonidentity problem discussed by Parfit (1984) rather trivial: if no one else will be affected, and a woman has to choose between having one of two different children, Anonymity plus Pareto implies that it would be better if she had the child whose life would be better.

Let U be the function defined on histories such that for any history h with population $1,\ldots,n$,

$$U(h) \coloneqq u_1(h) + u_2(h) + \cdots + u_n(h).$$

Then the premises of Harsanyi's theorem, but with Impartiality replaced by the stronger Anonymity, jointly imply

Same Number Claim Assume that all histories contain populations of the same size. Then the risk-free betterness relation is represented by U.

Turning to comparisons between populations of different sizes, I will outline an approach due to Broome (2004) and Blackorby, Bossert, and Donaldson (1995). I lack the space to discuss the details, but the crucial step is to argue for the

Neutral existence claim There exists a life l such that in every situation, provided no one already existing is affected, (i) it is better to create an extra life which is better than l; (ii) it is worse to create an extra life which is worse than l; (iii) it is a matter of indifference to create an extra life which is exactly as good as l.

Call such a life a *neutral existence*. Given a parameter v, let V be the function defined on histories such that for each history h with population $1,\ldots,n$

$$V(h) \coloneqq (u_1(h) - v) + (u_2(h) - v) + \cdots + (u_n(h) - v)$$

Some simple algebra shows that the same number and neutral existence claims together imply the

Variable number claim Assume that all histories contain finite populations. Then the risk-free betterness relation is represented by V, where v is the utility level of a neutral existence.

The value of v makes no difference to same number problems. For when comparing two histories with populations of the same size using V, the subtracted v's cancel out. In variable number problems, the presence of v in the definition of V means that ignoring effects on other people's lives, someone's existence makes a positive contribution towards goodness if and only if her life is better than a neutral existence.

Nothing so far said tells us what the value of v is, however. Setting it will involve further ethical issues, and is difficult to do in a way which respects common intuitions (Broome, 2004). For example, setting it low leads to the conclusion that a large number of people (e.g. a billion) all with extremely good lives is worse than an extremely large number (e.g. a billion billion) all with lives which may seem hardly worth living. Parfit (1984) evidently did not think much of this idea when he famously called it "the repugnant conclusion." On the other hand, setting the value of v high makes it bad to create someone who would have an intuitively good life, and that may seem implausible too.

When we ethicists first start to think seriously about probability, it may seem like a bane for us, vastly expanding the complexity of questions we have to address. But it may now look like a blessing. The problem of aggregating individual well-being to form an overall judgment about when one history is better than another seems difficult. Yet without appearing to make any assumptions about aggregation, and instead by largely appealing to expected utility theory, which is all about probability, Harsanyi's theorem seems to provide a solution. Section 33.15 will provide a closer look at the question of whether the theorem really does solve the "problem of aggregation." But we first examine criticisms of and alternatives to Harsanyi's premises which are also about probability.

33.9 Equality and fairness

The additive form of the conclusion of Harsanyi's theorem will make some suspect that its premises conflict with the idea that in the distribution of goods, equality and fairness matter. But where, if anywhere, is the tension? Assume a population of two people, A and B, and consider the following lotteries, which combine examples due to Diamond (1967) and Myerson (1981).

L_E	*heads*	*tails*
A	1	0
B	1	0

L_F	*heads*	*tails*
A	1	0
B	0	1

L_U	*heads*	*tails*
A	1	1
B	0	0

Anyone who thinks that equality is valuable should think that L_E is better than L_F. For while L_E and L_F are equally good for each person, L_E has in its favour that it guarantees equality of outcome while L_F guarantees inequality (Myerson, 1981). But Pareto implies that L_E and L_F are equally good, so it is inconsistent with the idea that equality is valuable.

Anyone who thinks that fairness is valuable should think that L_F is better than L_U. For while Impartiality implies that the outcomes under L_F and L_U are equally good, L_F has in its favour that it distributes the chances fairly (Diamond, 1967).

Diamond's example leads to the first of a series of challenges to the assumptions about expected utility in Harsanyi's premises. By Impartiality, all of the outcomes under L_F and L_U are equally good. Strong Independence of the betterness relation then implies that L_F and L_U are equally good.[11] Hence the assumption that the betterness relation satisfies the expected

[11] Proof: for all lotteries L and M, write $L \succsim M$ for "L is at least as good as M". By Impartiality, $[1,0] \sim [0,1]$. Strong Independence for $\succsim$ then implies $L_U = \frac{1}{2}[1,0] + \frac{1}{2}[1,0] \sim \frac{1}{2}[0,1] + \frac{1}{2}[1,0] = L_F$ as required.

utility axioms, in particular Strong Independence, clashes with the idea that fairness is valuable.

I think that Myerson's and Diamond's examples lie at the heart of concerns with equality and fairness.[12] It is difficult to argue for this in a short space, though section 33.16 will say more. But suppose it is correct. How could the examples be generalized into full-blown theories about what it is for equality or fairness to be valuable?

I will just illustrate an approach for the case of equality. Suppose we are given a preorder $\succsim_e$ on histories such that $h_1 \succsim_e h_2$ if and only if h_1 is uncontroversially (among egalitarians) at least as good in terms of equality as h_2. My own account of the extension of $\succsim_e$ is in McCarthy (2015). But to give two simple cases, every equal distribution is going to be uncontroversially better in terms of equality than every unequal distribution, and all equal distributions are going to be uncontroversially equally good in terms of equality. Consider

Equality-neutral Pareto Assume a fixed population. For all lotteries $L_1 = [p_1, h_1; \ldots p_m, h_m]$ and $L_2 = [p_1, k_1; \ldots; p_m, k_m]$: (i) *if* L_1 is exactly as good as L_2 for all individuals, and $h_j \sim_e k_j$ for all j, *then* L_1 and L_2 are equally good; and (ii) *if* L_1 is at least as good as L_2 for all individuals and better for some individual, and $h_j \succsim_e k_j$ for all j, *then* L_1 is better than L_2.

Equality principle Assume a fixed population. For all lotteries $L_1 = [p_1, h_1; \ldots; p_m, h_m]$ and $L_2 = [p_1, k_1; \ldots; p_m, k_m]$: *if* L_1 is at least as good as L_2 for all individuals, $h_j \succsim_e k_j$ for all j and $h_j \succ_e k_j$ for some j, *then* L_1 is better than L_2.

McCarthy (2015) argues that together, these principles are the core of egalitarianism. Equality-neutral Pareto is a weakening of Pareto, designed to avoid clashes with examples like Myerson's. The equality principle is designed to generalize the idea that equality is valuable, as illustrated by Myerson's example. Thus we obtain a very general egalitarian theory by starting with Harsanyi's premises, weakening Pareto to its equality-neutral cousin, then adding the equality principle.

Notice that the equality principle is inconsistent with the adoption of either Harsanyi's or Rawls's veil of ignorance. But it can easily be shown to be consistent with the notion of impartiality captured by Impartiality. So if it was meant only to model impartiality, the adoption of a veil of ignorance is too strong.

The characterization of the idea that equality is valuable via the equality principle exploits natural dominance ideas. Roughly speaking, suppose that each part of some object x is at least as good with respect to some value V as the corresponding part of object y. Then x is said to *weakly dominate* y in terms of the value V. If x weakly dominates y, but y does not weakly dominate x, then x *strictly dominates* y. Thus the equality principle says that if L_1 weakly dominates L_2 in terms of well-being, and strictly dominates L_2 in terms of equality, then L_1 is better than L_2. I lack the space to discuss the details, but I believe that the way to characterize the idea that fairness is valuable is to develop dominance ideas in a way suggested by Diamond's example. However, while the apparent similarities between Diamond's and Myerson's examples suggest parallels, it appears that there are subtle asymmetries between concerns with equality and concerns with fairness (McCarthy and Thomas, 2016).

[12] This is not quite right. In my view it is better to say that Myerson's example is about equality of outcome, and Diamond's is about equality of prospects, not fairness. But here I stick with the more usual terminology. For reasons for not talking about fairness, see McCarthy (2015).

33.10 Priority

Parfit (2000) argued that what he called the priority view is an important alternative to egalitarianism, sharing many of its apparent virtues but avoiding what he called the leveling-down objection. He summarized it via the slogan that "benefiting the worse off matters more", but commentators have been divided over whether he managed to articulate a genuine alternative to egalitarianism.

A puzzle about making sense of the priority view is that its distinctive feature is advertised as an intrapersonal phenomenon: what is bad about people being worse off is that they are worse off than they might have been (Parfit, 2000, p. 104). This has suggested to commentators that according to the priority view, it matters more to more to benefit someone the worse off she is even when no others are around at all (Rabinowicz, 2002). But in cases where only one person is around and risk is not involved, the priority view, like any other sane view, will accept that one history is better than another if and only if it is better for the sole person.

Matters are different, however, when risk is involved. Several commentators have thought that the priority view should be formulated in a way which makes it have distinctive consequences in one-person cases involving risk (Rabinowicz, 2002; McCarthy, 2008; Otsuka and Voorhoeve, 2009). I am inclined to go further and say that the key idea behind the priority view receives its clearest and most fundamental expression in such cases.

To illustrate, suppose A is the only person around, and compare the history $h = [1]$ with the lottery $L = [\frac{1}{2}, 2; \frac{1}{2}, 0]$, with the numbers supplied by u_A. Because L and h are equally good for A, Pareto implies that they are equally good. But I believe that the priority view should be understood as saying that h is better than L.

More generally, I believe that the key idea of the priority view is what I call the

Priority principle Assume a fixed population. Suppose histories h_1, h_2 and h_3 each contain perfect equality. Then (i) h_1 is at least as good as h_2 if and only if h_1 is at least as good for each individual as h_2; and (ii) *if* for each individual i, h_1 is better for i than h_2, h_2 is better for i than h_3, and h_2 is exactly as good for i as $L = [\frac{1}{2}, h_1; \frac{1}{2}, h_3]$, *then* h_2 is better than L.

Notice that this is inconsistent with equality-neutral Pareto. Some writers find it absurd that in one-person worlds, the betterness relation and the sole person's individual betterness relation could diverge (e.g. Otsuka and Voorhoeve, 2009), as the priority principle implies. Rabinowicz (2002) regards this claim as acceptable, while Parfit (2012), for example, offers a defense.

But rather than discuss possible defenses of the priority principle, I will note a less discussed objection to the priority view. The priority view can be formulated by starting with Harsanyi's premises, weakening Pareto far enough to accommodate the priority principle, then adding the priority principle (McCarthy, forthcoming a). But when this is done, any account of the extension of the betterness relation which is consistent with the Harsanyi premises turns out to be consistent with the priority view premises, and vice versa. But the priority view has a more complicated way of describing the betterness relation, because of the less simple relationship it posits between betterness and individual betterness in one-person worlds. So the objection is that the priority view fails to provide a reasonable alternative to the Harsanyi premises, not because of any ethically absurd implications, but

because of the theoretical vice of needless complexity (cf. Harsanyi, 1977b; Broome, 1991; McCarthy, 2008, forthcoming a).[13]

33.11 Continuity

Continuity is seldom discussed. When it is mentioned, it is often said just to be a technical assumption. But when the claim is that the betterness relation or individual betterness relations satisfy Continuity, this is a clear mistake.

To illustrate, let a be a very good life, a^+ a slightly better life, and z an extremely bad life, such as being in severe pain or enslaved for a long time. The claim that individual betterness relations satisfy Continuity implies that there is a gamble which would almost guarantee an individual a^+ with a small chance of z which is better for the individual than having a for certain. But regardless of what one thinks about this case, it is not a technical assumption to claim that the risk is worth it. It is a substantive evaluative judgment, and different views about it are reasonable. For what it is worth, I believe that many of Rawls's informal remarks about his veil of ignorance would have been more naturally modeled by denying that individual betterness relations satisfy Continuity because of this kind of case than by his actual model.

It is clear that Continuity is something ethicists should pay attention to. The good news is that the result of weakening the expected utility axioms by dropping Continuity is formally well understood, thanks to results by Hausner (1954) and others.

But there are several pieces of bad news. First, the general statement of Hausner's result is quite mathematically complex and not easy to speak about informally. Secondly, it is time to stop speaking of *the* continuity axiom. There are several EUT-style continuity axioms (see e.g. Hammond, 1998), and it is far from clear what the ethical grounds for adopting one but not another might be. Thirdly, speaking loosely, Continuity failures occur when one lottery in some sense has "infinitesimal" value compared with another. But such cases pose a challenge to standard treatments of probability as well, and this needs to be incorporated into the analysis.[14] In summary, perhaps in the end ethicists can safely ignore Continuity. But it would be better to know that than to hope for it, and the work needed to arrive at such a conclusion appears to be substantial.[15]

33.12 Incommensurability

One of the major contributions of the contractualist literature has been to force us to take seriously difficulties with evaluative comparisons of different kinds of goods. But part of

[13] As an analogy, consider again the best-system analysis of laws. Suppose someone offers some account of the laws of the world which captures all relevant facts. But this account is more complex than some other account which also captures all relevant facts. On the best-system analysis, the more complex account is mistaken about what the laws are, despite getting the relevant facts right. McCarthy (forthcoming a) argues that the priority view is mistaken on similar grounds.

[14] For an accessible account of how the challenge applies to Savage's treatment of subjective probability, and a sketch of mathematically sophisticated responses, see Gilboa (2009) pp. 99–100.

[15] For recent work in this direction, see Jensen (2012).

the assumption that the betterness relation and individual betterness relations satisfy the expected utility axioms is that these relations are complete. But from the perspective of difficulties with evaluative comparisons, such completeness assumptions look far from obvious. They may seem particularly implausible if we adopt the popular view that the basis for such things as interpersonal comparisons should be as neutral as possible between competing substantive views about what a good life is, as argued, for example, in Rawls (1982).

One response would be to adopt something like resources, freedoms, or opportunities as the basis for interpersonal and intrapersonal comparisons (see e.g. Rawls, 1982; Sen, 1985). However, the premises of Harsanyi's theorem are silent on the content of individual betterness relations, so there is no obvious reason why the theorem cannot be run when their content is understood in terms of resources and so on. Nevertheless, even resources have their own problems to do with comparability because of the different nature of different kinds of resources. So this response is a diversion, and we should turn directly to Harsanyi's premises to see what can be done about difficulties with comparability.

The most immediately tempting response is simply to drop the completeness assumptions. This means that the various evaluative relations featuring in the theorem become preorders which are not assumed to be complete. A large advantage of working with preorders is that mathematically speaking, they are relatively tractable. For example, a corollary of Szpilrajn's theorem is that a preorder is identical to the intersection of all of the complete preorders which extend it. This has the advantage that in thinking about preorders one can often work with complete preorders anyway.

This corollary is strikingly parallel to the supervaluationist treatment of vague predicates: a sentence involving a vague predicate is true if it is true on all admissible sharpenings of the predicate, false if it is false on all admissible sharpenings, and neither true nor false otherwise.

But this should suggest caution: if a natural response to difficulties to do with comparability is to shift to preorders, the response looks like one of the classic candidates for a solution to the problem of vagueness. But supervaluationist approaches have been heavily criticized (see e.g. Williamson, 1994). Furthermore, perhaps the parallel suggests that the basic problem with comparing different kinds of goods is one of vagueness. In fact, cases in which evaluative comparisons look extremely difficult seem to lend themselves to sorites paradoxes, one of the hallmarks of vagueness.

In one way this is good news: there is a vast amount of work on vagueness, so ethicists have plenty of material to borrow from. Since the topic is probability, it is worth mentioning that some treatments of vagueness are probabilistic, and that an extensive literature takes this approach to vague comparatives; see e.g. Fishburn (1998) for a survey. In another way it's bad news: perhaps the main reason why there is so much literature on vagueness is the almost complete lack of consensus.

Perhaps we ethicists should just shelve the problem of how best to model difficulties to do with evaluative comparisons until there is more convergence in the literature on vagueness. However, in the absence of such convergence, it may still be possible to achieve some kind of stability result: show that the solutions to a class of interesting ethical problems which involve goods which are difficult to compare are insensitive to the resolution of more general problems about vagueness. For example, Broome (2004) takes this approach in his discussion of the neutral level for existence. In section 33.16 I will suggest that the same can be done for the question of what Harsanyi's theorem really shows.

33.13 Non-expected utility theory

The backbone of Harsanyi's theorem is expected utility theory, but we have seen a number of ways in which the claim that various evaluative relations satisfy the expected utility axioms can be criticized. The axioms so far criticized are Strong Independence, Ordering (insofar as completeness was criticized), and Continuity. Some writers even go so far as to criticize transitivity (see e.g. Temkin, 2012).

These criticisms are directly based either on distributive intuitions (Strong Independence, Continuity), or on the nature of goods being distributed (Ordering). But a serious question about the expected utility axioms arises from a different direction.

Since Allais (1953) and Ellsberg (1961), it has appeared to many that individual preference relations violate the expected utility axioms in fairly systematic ways. The attempt to describe these violations has led to a huge body of work developing alternatives to the expected utility axioms (for surveys see e.g. Schmidt, 2004; Sugden, 2004; Gilboa, 2009; Wakker, 2010).

This project has been accompanied by two broad views. One is that the alternative axioms simply help us catalogue human irrationality, which might of course be very important in various descriptive and explanatory contexts. The other, often prompted by the fact that the violations are often stable under criticism, is that the support the alternative axioms tacitly enjoy genuinely threatens the picture of rationality provided by expected utility theory.

Now these are views about rationality, whereas we have been interested in such things as betterness and betterness for people. But the development of non-expected utility theory suggests that it would be interesting to modify distributive theories which to varying extents involve the expected utility axioms by weakening those axioms and then adding some of the non-expected utility axioms.

If the application of the non-expected utility axioms to such things as individual betterness relations turns out to be reasonably well motivated, the result should be an expanded account of reasonable distributive theories. But even if those axioms are not well motivated when applied to evaluative relations, this project would still be worth pursuing. If a class of popular distributive intuitions turns out to be generated by such an application of non-expected utility theory, we would in effect have an important error theory.

33.14 Evaluative measures

Discussions of the ethics of distribution commonly assume the existence of quantitative measures of various evaluative properties, then use these measures to formulate various apparently natural ideas. For example, individual goodness measures, quantitative measures of how good histories are for individuals, are often taken to exist. Then assuming a constant population $1, \ldots, n$, it is often claimed that

(**U**) According to utilitarianism, two histories are equally good if they contain the same sum of individual goodness.

(**E**) According to egalitarianism, an equal distribution is better than an unequal distribution of the same sum of individual goodness.

(**P**) According to the priority view, it is better to give a unit of individual goodness to a worse-off person than to a better-off person.

These claims tacitly assume that talk of units of individual goodness is well-defined. They are often taken to be (at least partial) definitions of the distributive theories in question, making what seems natural or appealing about the theories in question transparent. For more detail, McCarthy (2013) examines the role of evaluative measurement in common understandings of the priority view.

However, there are serious difficulties with this kind of approach to the ethics of distribution. I will mention just one specific problem.

The only obvious fact about individual goodness measures is that they have to represent risk-free individual betterness relations. But this only makes individual goodness measures unique up to increasing transformation.[16] But for units of individual goodness to be well-defined, individual goodness measures must be unique up to positive affine transformation. So to make them well-defined it looks as if we need to make an arbitrary choice of measure (Broome, 2004, p. 90). But this will make the theories partially defined by (U), (E), and (P) rest on an arbitrary choice, and fail to vindicate the idea that they are the fundamental theories about the ethics of distribution we take them to be.

More generally, taking the existence of quantitative evaluative measures as given, then using them to theorize about the ethics of distribution, is strongly at variance with standard views about measurement in the physical and social sciences. There, quantitative measures are seen as emerging as canonical descriptions of qualitatively described prior structures (see e.g. Krantz, Luce, Suppes and Tversky, 1971; Narens, 2007; Roberts, 2009). My own view is that we should treat evaluative measurement along the same lines.

By itself, this does not begin to settle what we should say about individual goodness measures. But individual goodness measures turn out to be well-enough defined for talk of units, sums, and so onto make sense, at least given certain background assumptions.

I can only sketch this view, but in more detail, sections 33.9 and 33.10 point to a characterization of egalitarianism and the priority view in terms of primitive qualitative relations (betterness, individual betterness). Similarly, I think the premises of Harsanyi's theorem should be understood as characterizing utilitarianism. Now (U), (E), and (P) are close to platitudinous. But given these characterizations of utilitarianism, egalitarianism and prioritarianism, this means that we can treat (U), (E), and (P) as implicit definitions of individual goodness measures. The result is that individual goodness measures turn out to be the positive affine transformations of $u_1, \ldots, u_n$, or what Broome (1991) calls Bernoulli's hypothesis. For details, see, for example, McCarthy (2015).

The background assumptions are that individual betterness relations satisfy the expected utility axioms and that interpersonal comparisons are unproblematic. But what if these fail? I will not pursue this, for I think the most important lesson about evaluative measures is not that they are arguably well-defined, but that it does not much matter. We can and

[16] I.e. if some function f represents the risk-free betterness relation, and g is some strictly increasing function on the reals ($x < y \implies g(x) < g(y)$), then $g \circ f$ also represents the risk-free betterness relation.

should theorize about the questions which really matter in the ethics of distribution without using evaluative measures. By focusing instead on comparatives and various claims about probability, none of the distributive views we have been discussing presuppose the existence of evaluative measures, the preeminent example, of course, being Harsanyi's.

33.15 Aggregation

But this raises the question of what Harsanyi's theorem really shows. Ethicists often talk about the "problem of aggregation". What they typically have in mind is the task of somehow combining an assessment of what things are like for each individual in a particular situation to form some sort of overall judgment of the situation which enables us to make an evaluative comparison with other situations.

Supposing the premises of Harsanyi's theorem are correct, it is tempting to think that Harsanyi's theorem solves the problem of aggregation. I believe this was Harsanyi's view, and I think it is popular among welfare economists. Harsanyi did not use the terms 'individual betterness relation' and 'betterness relation', and I stress that the following passage is mine, not his. But I think the following captures the spirit of his view (see especially Harsanyi, 1977a).

> Determining the content of individual preferences relations (despite filtering out various irrationalities, excluding such things as sadistic preferences, and requiring preferences to be rich enough to enable interpersonal comparisons) is basically a psychological matter (Harsanyi, 1977a). It does not involve any significant evaluative or aggregative assumptions. But we should identify individual betterness relations with individual preference relations. Given the truth of Harsanyi's premises, Harsanyi's theorem then explicitly determines the extension of the betterness relation. Problem of aggregation solved.

This position underplays the role of evaluative assumptions in determining the content of individual betterness relations in at least two ways. First, determining the content of individual preference relations may well involve prior evaluative assumptions because of the role of such assumptions in popular accounts of radical interpretation (see e.g. Lewis, 1974). Secondly, even when they are restricted to histories, identifying individual betterness relations with individual preferences relations is highly controversial. It is a major evaluative question whether to understand the content of risk-free individual betterness relations in terms of preferences, the quality of the individual's experiences, her achievements, or some combination thereof.

But suppose that evaluative question has been settled, and that Harsanyi's premises are true. The theorem certainly shows that figuring out the content of the betterness relation is no harder than determining the content of individual betterness relations. But what exactly does it show about the problem of aggregation?

First, it is a vast exaggeration to say that the theorem solves the problem of aggregation. Problems of aggregation arise whenever we have to make some sort of assessment of a whole based on an assessment of its parts. But figuring out the content of individual betterness relations involves major questions of aggregation. Even in the case in which all outcomes are equally likely, to assess whether facing some lottery is better for someone than some particular outcome, we will have to assess what each of the possible outcomes of the lottery

are like for her, then somehow aggregate to reach an overall assessment of the lottery. This problem is complicated and is, in my view, much neglected. Like many economists, Harsanyi's own account tacitly appeals to the individual's preferences. But this should not seem very appealing to those of us who think that preference satisfaction accounts are mistaken even for the question of when one outcome is better for an individual than another.

Secondly, there is no logical reason why we cannot use the theorem to deduce the content of individual betterness relations from the content of the betterness relation, in particular from judgments about when one history is better than another. In cases where we are very confident about the latter, this will even seem appealing. I am afraid I lack the space to discuss this, but I think this idea provides a natural way of interpreting various contractualist comments about veil of ignorance arguments (see e.g. Scanlon, 1982; Nagel, 1970), in particular leading to an interesting case for rejecting the claim that individual betterness relations satisfy Continuity.

More generally, if its premises are true, Harsanyi's theorem teaches us that determining the content of the betterness relation is easier than we may have thought. But the flipside is that determining the content of individual betterness relations is harder than many of us have assumed.

33.16 Summary on evaluation

When thinking about the ethics of distribution, it may seem that the real evaluative questions are about when one history is better than another, or better for some individual. Factoring in probability may then seem like a basically technical exercise, not one ethicists need be much concerned with.

Almost every topic discussed could easily have its own survey article. I have had to omit many important positions, and give only sketchy defenses of positive positions. Nevertheless, I have tried to make the case for the opposite view. Not only are there very important ethical issues about how to rank lotteries, but these issues directly bear on questions about when one history is better than another. I will end the evaluative discussion with two opinions.

First, if I am right, almost every major position on the ethics of distribution is essentially to do with probability. For example, assuming a constant population $1 \ldots n$, concerns with fairness, equality, and giving priority to the worse-off as characterized in sections 33.9 and 33.10 can each be shown to be consistent with the popular idea that the risk-free betterness relation is represented by $w \circ u_1 + \cdots + w \circ u_n$ for some strictly increasing and strictly concave function w. These views come apart only when probability is introduced. So one aspect of the importance of probability is the increase in expressive power its introduction provides: it allows us to draw distinctions which are difficult or impossible to draw in a risk-free framework.

Secondly, I think the various challenges to Harsanyi's premises stemming from appeals to equality, fairness, priority, and non-expected utility theory fail. To be sure, there is at least a reasonable case for rejecting Continuity, and Ordering (at least, the completeness part of it) is under serious threat. Nevertheless, we can drop Continuity and under many ways of modeling difficulties to do with comparability, what I take to be the core lesson of Harsanyi's

theorem remains stable:[17] determining the content of individual betterness relations and determining the content of the betterness relation are just different descriptions of the same problem. This may help. Our initial judgments about individual betterness and about betterness may be in tension with each other, and we may be more confident about some judgments than others. Harmonizing these judgments in an attempt to achieve reflective equilibrium may increase our confidence in the result.

33.17 Ought

Expected utility theory has turned out to be hugely important for developing a taxonomy of answers to the fundamental evaluative question: when is one history or lottery better than another? I have not emphasized this, but I also think that the clarity of this taxonomy is also extremely helpful for assessing which answer is correct. In the remaining space I have room for only one suggestion which, though hardly very original, is that the same turns out to be true for decision theory and the fundamental normative question: what ought we to do?

One immediate disclaimer is needed. Expected utility theory is usually understood as a theory about the structure of the preferences of ideally rational agents. But this chapter has discussed the application of expected utility theory to understanding evaluative comparatives without having to say anything about rationality. Rather, many of the ideas and criticisms of expected utility theory are directly applicable to questions about evaluative comparatives.

Similarly, decision theory is usually understood as an account of ideally rational action, and it is typically assumed that the rationality of an action depends in some way upon the agent's preferences. However, we can apply many ideas from decision theory directly to questions about the fundamental normative question without having to presuppose some grand connection between rationality and ethics. For example, it is a serious mistake to think that decision theory is going to be important to ethics only if ethics is somehow about preference satisfaction, or if we hitch ourselves to the unlikely project of deriving ethics from rationality. Thus the discussion of decision theory in what follows is only meant to draw parallels between questions about ethics and questions about rationality. Because the debates about rationality are often better developed, these parallels may be illuminating. With no attempt at exhaustiveness, the sequel will look briefly at three examples, with particular emphasis on probability.

33.18 Act consequentialism

Given some account of betterness, the most obvious ethical theory is act consequentialism: what we ought to do is to bring about the best available lottery. If we assume for simplicity that the betterness relation satisfies the expected utility axioms, act consequentialism then

[17] In fact, this is true even if we weaken some of the EUT ideas in Harsanyi's framework and add various well-known nonEUT ideas. This is further pursued in McCarthy, Mikkola, and Thomas (2016) and McCarthy (forthcoming b).

implies that there is some value function such that we ought to perform the action with the greatest expected value. Thus act consequentialism is the ethical theory which most obviously parallels decision theory.

Act consequentialism is also one of the most criticized theories, one standard criticism being that it has implausible implications. For example, assuming an impartial method of valuation, Williams (1973) argued that act consequentialism undermines the partiality which for many people makes life worth living: devotion to personal projects and particular people, often friends and family. But this raises the question of what act consequentialism really requires in the first place.

Taking for granted a probability-based view which uses subjective probabilities, or at least, probabilities which are relative to the evidence available to the agent, Jackson (1991) famously argued that because of facts about each individual's probabilities, act consequentialism will typically not require each agent to promote general well-being and pursue whichever projects are the most impartially valuable. Rather, it will require a typical agent – Alice, let's call her – to promote the well-being of the relatively small group of people Alice knows and cares about, and to adopt and then pursue projects in which Alice takes a natural interest. This does not amount to a rejection of impartial valuation, but instead reflects facts about each agent's limited information, the costs of deliberation and of acquiring new information, the complexity of the interpersonal and intrapersonal coordination problems she faces, the effects her actions will have on the expectations others will have of her future behaviour, her motivational strengths, and so on. Such facts will be encoded in the agent's probabilities, and will therefore affect which of her acts will maximize expected value. Very often, Jackson argued, such acts will favour her nearest and dearest.

Jackson's argument was offered as a response to Williams, but it offers a much more general lesson. Understanding what act consequentialism implies is going to require sophisticated thinking about probability. The huge complexity of this problem stands in sharp contrast to the occasional complaint that act consequentialism is simple-minded.

33.19 Rule consequentialism

Many writers, however, prefer rule consequentialism (or contractualism: at the normative level, these views are often very similar). On the one hand, rule consequentialism seems to fit better with common opinion about what we ought to do than act consequentialism (it is said to secure rights etc.). On the other, it seems to avoid the obscurities of deontology by resting its account of what we ought to do on an appeal to what is good for people. But how is this achieved?

Harsanyi's writings on rule utilitarianism offer a relatively clear answer. Simplifying slightly, Harsanyi (1980) claims that each member of a society of act utilitarians will always maximize the sum of expected individual utilities where the calculation is based on her subjective probabilities of what the other members are going to do. Each member of a society of rule utilitarians is committed to and thus will always act upon the rule *R* which is such that if everyone acts according to *R* expected utility will be greater than if everyone acts according to some other rule (I ignore the possibility that two rules could be tied).

Harsanyi claims that rule utilitarianism will lead to "incomparably superior" overall results in comparison with act utilitarianism because of its superiority in two kinds of scenarios: (i) in certain simultaneous coordination games (e.g. choosing whether to vote), and (ii) in certain sequential games (typically involving choices about respecting rights, keeping promises etc.). This superiority is despite the fact that *R* will sometimes tell agents to perform actions which they are certain will produce suboptimal results, where optimality is understood in terms of maximizing the sum of expected utilities. This last feature leads many to suspect that there is something unstable about rule utilitarianism, but Harsanyi claims that these superior overall results imply that rule utilitarianism is correct.

It would take a separate article even to outline the important issues here, and I merely want to make three points to illustrate the potential value of looking at this style of argument through the lens of contemporary debates about decision theory. To do that, I will assume for the sake of argument (though this is far from obvious) that Harsanyi is right about the superior overall results of rule utilitarianism in comparison with act utilitarianism.

First, Harsanyi stresses that the rule utilitarians take themselves to be facing a problem involving complete probabilistic dependence: each will commit to (and thus act on) rule *R* if and only if all commit to *R*. In this respect, rule utilitarians are like clones in the well-known case of clones playing a prisoner's dilemma. It is this probabilistic dependence which leads to rule utilitarianism's superior performance in the coordination games. However, in these coordination games, there is causal independence between the actions of each player. But "probabilistic dependence yet causal independence" takes us to a crucial issue in decision theory. Very roughly, so-called *evidential decision theory* assesses (the rationality of) actions in terms of how likely good outcomes are conditional upon the actions being performed. By contrast, *causal decision theory* assesses actions in terms of their causal tendency to produce good outcomes. The classic case in which the two come apart is Newcomb's problem. However, for those of us who think that Newcomb's problem teaches us to be causal decision theorists (see e.g. Joyce, 1999), probabilistic dependence is a red herring when there is causal independence, as there plainly is in Harsanyi's simultaneous coordination games. So we may think that Harsanyi has tacitly built something like evidential decision theory into rule utilitarianism, and so much the worse for rule utilitarianism.

Secondly, the success of rule utilitarianism in various sequential games stems from the rule utilitarians' commitment to the rule *R* even in contexts in which acting on *R* leads to suboptimal results. The conclusion that in virtue of this success, rule utilitarianism is right about what we ought to do is parallel to a revision to standard decision theory later urged by Gauthier (1994) and McClennan (1990). This revision claims that if it is rational at time t to become committed to performing some action at a later time t' which is obviously irrational when considered in isolation, it is rational to commit to the action and then later perform that action. But those of us who take the toxin puzzle of Kavka (1983) to dramatize why this revision is mistaken may think that rule utilitarianism is making the same kind of mistake.

Thirdly, Harsanyi's characterization of act versus rule utilitarianism parallels the influential distinction in von Neumann and Morgenstern (1944) between games against nature and games against other people. Each act utilitarian will have probabilities about a number of relevant variables, and will maximize expected value accordingly. The fact that some of these variables are the behaviour of other people who like herself are act utilitarians is neither here nor there; the decision theoretic model still applies. But when an agent is in a situation in which the outcome depends in part on the behaviour of agents just

like her, von Neumann and Morgenstern argued that decision theory is inappropriate. The problem of self-reference embedded into such situations requires the different tools of game theory, and Harsanyi's rule utilitarians reason along similar lines. Perhaps von Neumann and Morgenstern's argument could be used to bolster Harsanyi's approach. Alternatively, those of us who are convinced by Skyrms (1990) in thinking that problems of self-reference can and should be handled without having to abandon decision theory may think this points to a further difficulty for rule utilitarianism.

Of course, the fact that Harsanyi focussed on rule utilitarianism rather than rule consequentialism has been inessential to the discussion. These crude and preliminary remarks are meant only to suggest the value of looking at the foundations of rule consequentialism through the lens of parallel and often much more extensive debates about decision theory.

33.20 Deontology

Those with strong deontological intuitions may reject rule consequentialism, either because they are not convinced that it is a stable alternative to act consequentialism, or because its conclusions are not deontological enough. But we may now seem to have reached the limits of the usefulness of thinking about decision theory. Very roughly, anything like a decision theoretic approach to deontology looks like the wrong model: the former is all about weighing goods against evils, and the latter thinks there are circumstances in which such weighing is illegitimate, or counts for nothing. Nevertheless, one lesson from thinking about probability is that weighing is not so easy to avoid.

In trying to characterize a deontological view, there seem to be two basic options. What I will call *agent-centered* views typically prohibit actions which would involve the agent's mental states bearing some kind of inappropriate relation to the outcome. The most obvious example is the so-called *principle of double effect*, which in its simplest form prohibits bringing about intended harm, but permits certain otherwise identical cases of bringing about merely foreseen harm. What I will call *causal structure* views typically prohibit actions which stand in some kind of inappropriate causal relation to the outcome. For example, in the famous trolley problem, an out-of-control trolley is going to kill five people who are stuck on the track, but a bystander can switch the trolley to a sidetrack where it will kill one person. Many people who have strong deontological intuitions think it is permissible to switch the trolley. But in most cases, they think that killing one to save five is impermissible, as in the variant where the bystander can push a fat man off a bridge to stop the trolley (Thomson, 1976), killing him but saving the five. Causal structure theorists think the intentions of the bystander are irrelevant, and search for differences in the causal structure of the cases to explain the difference in permissibility.

Many deontologists have not had much sympathy for agent-centered views, and have preferred some kind of causal structure view (e.g. Kamm, 1996). But here is what I believe is a relatively neglected problem about views. If the inappropriate causal relation is between the action and the outcome – as in, e.g., the fat man variant but not the trolley problem itself – then prima facie, there are going to be actions which bring about the following lotteries:

some benefit occurs with nonzero probability p, some inappropriate causal structure obtains with probability $1 - p$. For example, driving a truck across the bridge will either miss the fat man and deliver aid elsewhere, or else hit him and topple him off the bridge, stopping the trolley and saving the five.

What should causal structure deontologists say about such actions? There are at least five responses: (i) All such actions are impermissible. Objection: this leads to an intolerably restrictive view. (ii) Such actions are impermissible if and only if they turn out to result in the inappropriate causal structure. Objection: similar to the objections to outcome-based views in section 33.2. (iii) Actions which lead to the inappropriate causal structure with probability one are impermissible, all others are permissible. Objection: it is not credible that there should be such a gulf between probability one and probabilities just less than one. (iv) Actions performed by agents whose reasons for performing them include the benefits resulting from the inappropriate structure are impermissible. Objection: this collapses causal structure views into agent-centered views. (v) Actions are impermissible if and only if p exceeds some intermediate probability threshold. Objection: this seems to be the most principled response for a causal structure view, but it suggests the acceptability of weighing the alleged badness of the causal structure against the production of benefits. This seems to fit poorly with the guiding deontological image of the inappropriateness of weighing when inappropriate causal structures are concerned.

Perhaps this kind of case points towards a serious problem for causal structure views; see further Jackson and Smith (2006). Or it may provide an opportunity for causal structure theorists to refine their views. Either way, thinking about probability and deontology seems helpful.

Acknowledgments

Thanks to Alan Hájek and Kalle Mikkola for very helpful comments. Support was partially provided by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (HKU 750012H).

References

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque, critique des postulates et axiomes de l'ecole américaine. *Econometrica*. 21. pp. 503–46.

Angner, E. (2004) Revisiting Rawls: *A Theory of Justice* in the light of Levi's theory of decision. *Theoria*. 70. pp. 3–21.

Blackorby, C., Bossert, W., and Donaldson, D. (1995) Intertemporal population ethics: critical-level utilitarian principles. *Econometrica*. 63. pp. 1303–20.

Broome, J. (1991) *Weighing Goods*. Cambridge, MA: Blackwell.

Broome, J. (2004) *Weighing Lives*. Oxford: Oxford University Press.

Buchak, L. (2016) Decision theory. In Hájek, A. and Hitchcock, C. (eds.). *The Oxford Handbook of Philosophy and Probability*. Oxford: Oxford University Press.

Diamond, P. (1967) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: comment. *Journal of Political Economy*. 75. pp. 765–6.

Edgeworth, F. (1881) *Mathematical Psychics*. London: Kegan Paul.
Ellsberg, M. (1961) Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*. 75. pp. 643–69.
Fishburn, P. (1970) *Utility Theory for Decision Making*. New York, NY: Wiley.
Fishburn, P. (1998) Stochastic utility. In Barberá, S., Hammond, P., and Seidl, C. (eds.) *Handbook of Utility Theory*. Vol. 1. Dordrecht: Kluwer.
Foot, P. (1985) Utilitarianism and the virtues. *Mind*. 94. pp. 196–209.
Gauthier, D. (1994) Assure and threaten. *Ethics*. 104. pp. 690–716.
Gilboa, I. (2009) *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.
Hájek, A. (2012) Interpretations of probability. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. (Winter) [Online] Available from: http://plato.stanford.edu/archives/win2012/entries/probability-interpret.
Hájek, A. (2014) *Most Counterfactuals are False*. Manuscript.
Hammond, P. (1991) Interpersonal comparisons of utility: why and how they are and should be made. In Elster, J. and Roemer, J. (eds.). *Interpersonal Comparisons of Well-Being*. pp. 200–54. Cambridge: Cambridge University Press.
Hammond, P. (1998) Objective expected utility. In Barberá, S., Hammond, P. and Seidl, C. (eds.) *Handbook of Utility Theory*. Vol. 1. pp. 143–211. Dordrecht: Kluwer.
Harsanyi, J. (1953) Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*. 61. pp. 434–5.
Harsanyi, J. (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*. 63. pp. 309–21.
Harsanyi, J. (1977a) Morality and the theory of rational behavior. *Social Research*. 44. pp. 623–56.
Harsanyi, J. (1977b) Nonlinear social welfare functions: a rejoinder to Professor Sen. In Butts, R. and Hintikka, J. (eds.). *Foundational Issues in the Special Sciences*. pp. 293–96. Dordrecht: Reidel.
Harsanyi, J. (1980) Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory and Decision*. 12. pp. 115–33.
Hausner, M. (1954) Multidimensional utilities. In Thrall, R., Coombs, C., and Davis, R. (eds.) *Decision Processes*. pp. 167–80. New York, NY: John Wiley & Sons.
Hoefer, C. (2007) The third way on objective probability: a sceptic's guide to objective chance. *Mind*. 116. pp. 549–96.
Jackson, F. (1991) Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*. 101. pp. 461–82.
Jackson, F. and Smith, M. (2006) Absolutist moral theories and uncertainty. *Journal of Philosophy*. 103. pp. 267–83.
Jensen, K. (2012) Unacceptable risks and the continuity axiom. *Economics and Philosophy*. 28. pp. 31–42.
Joyce, J. (1999) *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
Kamm, F. (1996) *Morality, Mortality*. Vol. 2. New York, NY: Oxford University Press.
Kavka, G. (1983) The toxin puzzle. *Analysis*. 43. pp. 43–6.
Krantz, D., Luce, R. D., Suppes, P., and Tversky, A. (1971) *Foundations of Measurement*. Vol. 1. New York, NY: Academic Press.
Kreps, D. (1988) *Notes on the Theory of Choice. Underground Classics in Economics*. Boulder, CO: Westview Press.

Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1974) Radical interpretation. *Synthese*. 27. pp. 331–44.
Lewis, D. (1980) A subjectivist's guide to objective chance. In Jeffrey, R. (ed.). *Studies in Inductive Logic and Probability*. Vol. 2. pp. 83–132. Berkeley, CA: University of California Press.
Lewis, D. (1994) Humean supervenience debugged. *Mind*. 103. pp. 473–90.
McCarthy, D. (1998) Actions, beliefs and consequences. *Philosophical Studies*. 90. pp. 57–77.
McCarthy, D. (2008) Utilitarianism and prioritarianism II. *Economics and Philosophy*. 24. pp. 1–33.
McCarthy, D. (2013) Risk-free approaches to the priority view. *Erkenntnis*. 78. pp. 421–49.
McCarthy, D. (2015) Distributive equality. *Mind*. 124. pp. 1045–109.
McCarthy, D. (forthcoming a) The priority view. *Economics and Philosophy*.
McCarthy, D. (forthcoming b) *The Structure of Good*. Oxford: Oxford University Press.
McCarthy, D., Mikkola, K., and Thomas, T. (2016) Utilitarianism with and without expected utility. MPRA Paper No. 72578 https://mpra.ub.uni-muenchen.de/72578/.
McCarthy, D. and Thomas, T. (2016) Egalitarianism with risk. Manuscript.
McClennen, E. (1990) *Rationality and Dynamic Choice*. Cambridge University Press.
Mongin, P. (1995) Consistent Bayesian aggregation. *Journal of Economic Theory*. 66. pp. 313–51.
Myerson, R. (1981) Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica*. 49. pp. 883–97.
Nagel, T. (1970) *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.
Narens, L. (2007) *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science*. Mahwah, NJ: Lawrence Erlbaum Associates.
Ng, Y. (1975) Bentham or Bergson? Finite sensibility, utility functions, and social welfare functions. *Review of Economic Studies*. 42. pp. 545–70.
Ok, E. (2007) *Real Analysis with Economic Applications*. Princeton, NJ: Princeton University Press.
Otsuka, M. and Voorhoeve, A. (2009) Why it matters that some are worse than others: an argument against the priority view. *Philosophy and Public Affairs*. 37. pp. 171–99.
Parfit, D. (1984) *Reasons and Persons*. Oxford: Clarendon Press.
Parfit, D. (2000) Equality or priority? In Clayton, M. and Williams, A. (eds.). *The Ideal of Equality*. pp. 81–125. Basingstoke: Macmillan.
Parfit, D. (2012) Another defense of the priority view. *Utilitas*. 24. pp. 399–440.
Rabinowicz, W. (2002) Prioritarianism for prospects. *Utilitas*. 14. pp. 2–21.
Ramsey, F. (1931) Truth and probability. In Ramsey, F. and Braithwaite, R. (ed.). *Foundations of Mathematics and other Essays*. pp. 83–132. London: Kegan, Paul, Trench, Trubner, & Co.
Rawls, J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, J. (1982) Social unity and primary goods. In Sen, A. and Williams, B. (eds.) *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.
Resnik, M. (1987) *Choices: An Introduction to Decision Theory*. Minneapolis, MN: University of Minnesota Press.
Roberts, F. (2009) *Measurement Theory*. Cambridge: Cambridge University Press.
Savage, L. (1954) *The Foundations of Statistics*. New York, NY: John Wiley.
Scanlon, T. (1982) Contractualism and utilitarianism. In Sen, A. and Williams, B. (eds.) *Utilitarianism and Beyond*. Cambridge, MA: Cambridge University Press.
Scheffler, S. (1982) *The Rejection of Consequentialism*. Oxford: Oxford University Press.

Schmidt, U. (2004) Alternatives to expected utility: formal theories. In Barberá, S., Hammond, P., and Seidl, C. (eds.) *Handbook of Utility Theory*. Vol. 2. pp. 757–837. Dordrecht: Kluwer.
Schwarz, W. (2016) Best system approaches to chance. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Philosophy and Probability*. Oxford: Oxford University Press.
Sen, A. (1976) Welfare inequalities and Rawlsian axiomatics. *Theory and Decision*. 7. pp. 243–62.
Sen, A. (1985) Well-being, agency and freedom. *Journal of Philosophy*. 82. pp. 169–221.
Skyrms, B. (1990) *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.
Sugden, R. (2004) Alternatives to expected utility: foundations. In Barberá, S., Hammond, P. and Seidl, C. (eds.). *Handbook of Utility Theory*. Vol. 2. pp. 685–755. Dordrecht: Kluwer.
Temkin, L. (2012) *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.
Thomson, J. (1976) Killing, letting die, and the trolley problem. *The Monist*. 59. pp. 204–17.
Thomson, J. (1986) Imposing risks. In Parent, W. (ed.) *Rights, Restitution, and Risk*. Cambridge, MA: Harvard University Press.
Thomson, J. (2001) *Goodness and Advice*. Princeton, NJ: Princeton University Press.
Vickrey, W. (1960) Utility, strategy, and social decision rules. *The Quarterly Journal of Economics*. 74. pp. 507–35.
von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
Wakker, P. (2010) *Prospect Theory: For Risk and Ambiguity*. Cambridge: Cambridge University Press.
Weymark, J. (1991) A reconsideration of the Harsanyi-Sen debate on utilitarianism. In Elster, J. and Roemer, J. (eds.). *Interpersonal Comparisons of Well-Being*. Cambridge: Cambridge University Press.
Williams, B. (1973) A critique of utilitarianism. In Smart, J. and Williams, B. (eds.) *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.
Williamson, T. (1994) *Vagueness*. New York, NY: Routledge.