PDF version of the entry
Moral Decision-Making Under Uncertainty
https://plato.stanford.edu/archives/spr2024/entries/moral-decision-uncertainty/

from the Spring 2024 Edition of the

Stanford Encyclopedia of Philosophy

Co-Principal Editors: Edward N. Zalta & Uri Nodelman

Associate Editors: Colin Allen, Hannah Kim, & Paul Oppenheimer

Faculty Sponsors: R. Lanier Anderson & Thomas Icard
Editorial Board: https://plato.stanford.edu/board.html

Library of Congress ISSN: 1095-5054

Notice: This PDF version was distributed by request to members of the Friends of the SEP Society and by courtesy to SEP content contributors. It is solely for their fair use. Unauthorized distribution is prohibited. To learn how to join the Friends of the SEP Society and obtain authorized PDF versions of SEP entries, please visit https://leibniz.stanford.edu/friends/ .

Stanford Encyclopedia of Philosophy
Copyright © 2024 by the publisher
The Metaphysics Research Lab
Department of Philosophy
Stanford University, Stanford, CA 94305

Moral Decision-Making Under Uncertainty
Copyright © 2024 by the authors
Christian Tarsney, Teruji Thomas, and William MacAskill

All rights reserved.

Copyright policy: https://leibniz.stanford.edu/friends/info/copyright/

Moral Decision-Making Under Uncertainty

First published Wed Mar 13, 2024

Many important debates in contemporary ethics centre on idealized thought experiments in which agents are assumed to have perfect information about the effects of their actions and other morally relevant features of the choices they face. If Abe turns the trolley, one person will certainly be killed; if he does not, five people will certainly be killed (Foot 1967); how the one and the five got into that situation, whether blamelessly or recklessly (Thomson 1976: 210–11), is also a matter of certainty. If Betty conceives a child now, it will certainly have a life that is hard but worth living, while if she waits, her child's life will certainly be better—and the two choices will certainly result in *different* children being born (Parfit 1984: 358).

These debates, focused on conditions of certainty, often suggest principles that are hard to generalize to conditions of *un*certainty. If it is always wrong to kill an innocent person (even to save more innocent people), how do we evaluate actions that carry a *risk* (perhaps minuscule) of killing an innocent person? If there is a moral obligation (all else being equal) not to bring *bad* lives into existence, but no obligation to bring *good* lives into existence, what do we say in situations where it is uncertain whether some potential future life will be bad or good? And what if we are uncertain about the relevant moral principles themselves?

While uncertainty creates challenges for ethicists, ethical considerations also raise distinctive challenges for standard principles of decision-making under uncertainty. Orthodox decision theory advises *expected utility maximization* as the rational response to uncertainty. But expected utility maximization may seem infeasible in ethical contexts due to the complexity of the morally significant effects of even ordinary choices

(Lenman 2000; Greaves 2016). It may give implausible weight to small probabilities of astronomically good or bad outcomes, like human extinction (Balfour 2021). It may struggle to accommodate certain aspects of non-consequentialist moral theories, like absolute moral constraints, permissions to act suboptimally, and the moral significance of unrealized risks (Hansson 2003; Lazar 2017a; Tenenbaum 2017). And it is unclear how, if at all, to extend expected utility theory to accommodate uncertainty about morality itself (Gracely 1996; Hedden 2016).

These questions are not just theoretically interesting but practically important. For morally motivated agents who care about making the world a better place, treating others fairly and respecting their rights, and so on, different principles for moral decision-making under uncertainty can yield very different advice. Whether such agents ought to prioritize the near-term or long-term effects of their actions might depend on how much weight they give to small probabilities of extreme outcomes, and how they deal with the unpredictability of long-term, indirect effects. Their dietary choices may depend on how they deal with uncertainty about the moral status of animals. And the problems of uncertainty may also be epistemically relevant to first-order moral questions—both consequentialism and deontology, for instance, have been accused of having no plausible way of dealing with uncertainty.

This entry will survey some of the central questions concerning moral decision-making under uncertainty. There are many active debates in this area, and we won't be able to cover them all. Instead, we have tried to focus on problems that have both general philosophical significance (affecting a wide range of ethical views) and practical significance (making a difference to important real-world choices).

We begin in section 1 by distinguishing several kinds of normative assessment for actions under uncertainty: "fact-relative" vs. "belief-relative" vs. "evidence-relative", and moral vs. rational. Section 2 then describes the orthodox normative theory of decision-making under uncertainty outside of ethics, namely, expected utility theory, and considers its potential applications to ethics. The subsequent sections each consider distinctive features of *morally motivated* decision-making under uncertainty that challenge straightforward applications of expected utility theory, and suggest the need for either modifications or entirely different approaches in the ethical context. Section 3 considers the apparent impossibility of applying expected-utility-like criteria to real-world ethical decision-making given the many long-term, indirect effects of even mundane choices. This difficulty, sometimes called the "problem of cluelessness", has been seen as posing a particular challenge to (impartial, agent-neutral) consequentialism, which gives full weight to these indirect effects. In contrast, section 4 considers some of the difficulties of incorporating *non*-consequentialist considerations into a theory of moral decision-making under uncertainty—in particular, agent-centred constraints, the moral significance of risk imposition, and the aim of respecting other agents' preferences with respect to risk and uncertainty. Section 5 considers how (if at all) agents should take account of their uncertainty about morality itself. Finally, section 6 considers a problem that afflicts consequentialists, non-consequentialists, and the morally uncertain alike: the possibility that moral decision-making under uncertainty comes to be implausibly dominated by low-probability, high-stakes moral considerations (like tiny probabilities of securing an astronomically good outcome or violating an extremely weighty moral constraint).

1. Objective, Subjective, and Prospective "oughts"

The following classic example illustrates both the need for a theory of moral choice under uncertainty, and some of its desirable features.

> Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug *A*, drug *B*, and drug *C*. Careful consideration of the literature has led her to the following opinions. Drug *A* is very likely to relieve the condition but will not completely cure it. One of drugs *B* and *C* will completely cure the skin condition; the other though will kill the patient, and there is no way that she can tell which of the two is the perfect cure and which the killer drug. What should Jill do? (Jackson 1991: 462–3)[1]

Here is a natural answer: In one sense, Jill ought to prescribe whichever drug is in fact the perfect cure—either *B* or *C*. This is what would have the best consequences for John (which seems to be all that is morally at stake); it is also what a fully informed and morally ideal version of Jill would do. Call this the "fact-relative" or "objective" sense of "ought". But this objective standard cannot be the *only* morally relevant standard in this case. For one thing, it's not the standard that determines praise or blame: We would not blame Jill for prescribing drug *A*, and we *would* blame her for prescribing either *B* or *C*, even if she got lucky and prescribed the curative rather than the fatal drug. For another thing, a moral injunction to

prescribe the drug that in fact has the best consequences does not seem *useful* from Jill's perspective—she cannot act on that advice, given her ignorance. And it seems that morality should, among other things, give agents usable guidance in morally important choice situations.[2] So there must also be at least one importance sense in which Jill morally ought to prescribe drug *A*—even though this is the one option that is certain *not* to effect a perfect cure. Given her uncertainty, it would be wrong (in this sense) to risk John's life when there is an imperfect but safe alternative in drug *A*. This article is concerned with this general type of verdict about what one ought to do, taking uncertainty into account.

There are some important questions about what uncertainty we are talking about here. We could focus on Jill's *beliefs* about the efficacy of the drugs; insofar as uncertainty is to be understood in probabilistic terms, the decision-relevant probabilities would be Jill's credences (or "degrees of belief" or "subjective probabilities"). We would then be asking what Jill ought to do in a "belief-relative" or "subjective" sense of "ought". One common alternative is to focus on Jill's *evidence* about the efficacy of the drugs; the decision-relevant probabilities will then be *evidential probabilities* representing how likely various possibilities are on Jill's evidence. We would then be asking what Jill ought to do in an "evidence-relative" or "prospective" sense of "ought". (The term "prospective" is due to Zimmerman [2008].) These two questions could come apart if Jill failed to proportion her credences to her evidence. (In certain circumstances, too, one might wish to focus on the *evaluator's* beliefs or evidence, or on the *available* evidence, rather than on features of the agent.)

Some philosophers think that "ought" is univocal, and that we must decide between objective, subjective, and prospective accounts of it.[3] Others think that there are multiple legitimate senses of "ought", including two or more of the objective, subjective, and prospective senses. (In a useful

discussion of these questions, Sepielli (2018b) calls the former group "debaters" and the latter group "dividers".) A similar range of positions is possible concerning other normative terms and concepts besides *ought*, like *rightness* or *reasons*.[4] In this article, we mostly steer clear of these debates. But we do unavoidably take for granted that there is *some* normative standard for evaluating morally significant choices that is sensitive to an agent's beliefs and/or evidence.[5]

A further question is whether the standards in play are *moral* standards, *rational* standards, or something else. We can ask, for instance, what Jill morally ought to do, given her uncertainty. But we can instead ask what it would be rational for her to do, given her uncertainty together with her moral commitments; or, we can ask what it would be rational for a morally conscientious person to do in her position. Such questions are still relevant to moral decision-making. (For a recent discussion, see Bykvist 2020.) Most of the discussion in this entry can be interpreted as addressing any one of these questions, although the distinction between them is especially relevant when it comes to *moral uncertainty*, the topic of section 5. Thus, in the rest of the entry, we will frame various substantive questions about moral choice under uncertainty in terms of what an agent *ought* to do or what she is *permitted* to do, while intending to be flexible about whether the relevant standards are belief-relative or evidence-relative, and rational or moral.

Though we can distinguish between objective and non-objective assessments of actions under uncertainty, it is natural to think that there is some tight relationship between the two. For instance, if you are *sure* that you objectively ought to φ, then you subjectively ought to φ. And if it is certain on your evidence that you objectively ought to φ, then you prospectively ought to φ. A natural generalization of these weak claims is that, in light of your uncertainty, you ought to do whatever is most likely to be the thing you objectively ought to do.

The central point of Jackson's case, however, is that this view is wrong. Given her uncertainty, Jill ought to prescribe drug *A*, the one option that is certain to be objectively wrong. In fact, even the weak claims may require qualification. If we allow "φ" to stand for "not prescribe *A*", then, given her uncertainty, it is not true that Jill ought to φ, even though it is certain that she objectively ought to ϕ.

What should we say instead? When thinking about Jill's situation, we note that although there is a good chance that prescribing drug *B* (say) is objectively right (because it would do a little better than drug *A*), there is also a good chance that it would be objectively wrong, and to a grievous extent (because it would kill John). What Jill ought to do seems to be a matter of weighing these positive and negative considerations against each other, both in terms of their probability and in terms of their moral seriousness. One simple approach to this sort of weighing is given by expected utility theory, a topic to which we now turn.

2. Expected Utility Theory in Ethics

Expected utility theory is the orthodox normative theory of choice under uncertainty. (There is really a family of slightly different theories, but we will gloss over the differences here.) It is typically understood as a theory about the structure of a rational agent's preferences. The same formal theory can, however, be repurposed as a theory about the structure of permissibility, or of what an agent ought to do, and that is how we will present it here. Perhaps there is a connection between these two applications of the formal theory: perhaps what you ought to do corresponds, more or less, to what you would prefer to do if you were rational and suitably motivated. However, it is a substantive question whether any such correspondence holds, and how to spell it out; we circumvent this issue by thinking directly in terms of permissibility.

Expected utility theory can then be understood as making the following claims. Each option leads to various possible outcomes with various probabilities. (Here, an "outcome" might include the act itself and such details as the identity and motives of the agent.) For each possible outcome O, there is a number $U(O)$, its *utility* (about which we will have much more to say below). The higher $U(O)$ is, the more reason there is to bring about O rather than any alternative.[6] So, in a choice where the outcome of each option is certain, one ought to choose an option that leads to the outcome with the highest utility. Moreover, in cases where the outcomes of some options are uncertain, one ought to choose an option with the highest *expected* utility, where the expected utility $EU(A)$ of an option A is the probability-weighted sum of the utilities of its possible outcomes:

$EU(A)$ the sum, over possible outcomes O, of the utility of O times the probability that O is the outcome of A.

(For simplicity of formulation, we assume that A has only finitely many possible outcomes.)

Before proceeding to some interpretive issues, let us illustrate this view by applying it to Jackson's case:

> Suppose that *B* and *C* are equally likely to be the perfect cure. Suppose the utility of John's death is -100 (there is strong reason to avoid it), the utility of a perfect cure is $+100$ (there is strong reason to achieve it), and the utility of a partial cure is $+90$ (there is slightly weaker reason to achieve it). Then Pill *B* and Pill *C* each have an expected utility of 0 $(100 \times 0.50 + (-100) \times 0.5)$, while Pill *A* has an expected utility of 90 $(90 \times 0.5 + 90 \times 0.5)$. Therefore Jill ought to treat John with Pill *A*.

As one can see in this example, expected utility maximization provides a "criterion of right": it says what the agent ought to do. But it can also, at least in simple cases, provide a decision procedure: the agent can figure out what to do by applying the expected utility formula (or else by reasoning in terms of the axioms that we will mention below). Some criticisms of expected utility theory (and of similar theories) in the moral context can best be understood as objections to it as a criterion of rightness, and others as objections to it as a decision procedure. (The distinction between "criteria of right" and "decision procedures" is due to R. Eugene Bales [1971].)

We have seen how to reason with utilities. But what *are* utilities? And why think that the permissible options are exactly the ones that maximize their expectation? These questions are arguably best treated together. Foundational treatments of expected utility theory typically proceed by proving a "representation theorem": if permissibility has certain *prima facie* plausible structural properties (and given certain background assumptions of a more technical nature), then there exists a (more or less unique) utility function such that the permissible options are precisely those that maximize expected utility.[7] According to this story, the utilities are functionally characterized by their role in the expected utility formula, and the existence of utilities fulfilling this role is guaranteed by the more basic structural axioms. Notably, then, expected utility theory does not require a moral theory to *explicitly* assign a utility to each outcome, and the claim that *there exists* an appropriate utility function is a mathematical claim about the structure of permissibility, not one that signals a further metaphysical commitment.

This is not the place for a detailed discussion of the relevant axioms, but let us informally state a few important ones to give the flavour:

Contraction consistency

If it is permissible to choose A from some set of options, then it is permissible to choose A from any subset of those options that includes A.

Expansion consistency
If A and B are both permissible choices from some set of options, then, from any larger set of options, they are either both permissible or both impermissible.[8]

Stochasticism
Which options are permissible in a given choice situation depends only on which possible outcomes the options lead to, and with what probabilities.

Continuity
If it is impermissible to choose A from some set of options, then it would still be impermissible if the probabilities were very slightly different.

The sure thing principle
If, in a given choice situation, one has no influence on whether E occurs, and one ought to choose A on the supposition that E occurs, and one ought to choose A on the supposition that E does not occur, then one ought to choose A.[9]

Although axioms of this kind are, at least at first sight, fairly ecumenical, it will be obvious that expected utility theory has a special affinity with maximizing act consequentialism. It is tempting to think of the utility of an outcome as representing the *value* or *goodness* of that outcome, in the sense dear to consequentialists; then, expected utility theory provides a natural way to understand consequentialism in the face of uncertainty. And it is true that moral theorists whose views depart widely from consequentialism will be especially motivated to modify or supplement expected utility theory. For just one example, expected utility theory, being

a maximizing theory, does not seem to allow for an interesting notion of moral supererogation. (At an axiomatic level, the maximizing flavour of the theory follows from contraction and expansion consistency; see note 8.) We will return to subtler problems along these lines in section 4.

However, the connection between expected utility theory and consequentialism is not a tight one. Consequentialists as well as others may wish to consider alternatives to (or generalizations of) expected utility theory, including ones that allow for "global" forms of risk aversion (typically violating the sure thing principle[10]); infinite values (typically violating continuity[11]); or incommensurability and similar phenomena (typically violating expansion consistency and perhaps stochasticism[12]). Moreover, the types of worries that we raise below about expected utility theory (and which apply to many variations of it) will move some consequentialists as well as non-consequentialists.

Conversely, nothing requires us to claim that the utility of an outcome represents its goodness. In general, the utility of O must represent something like the "choiceworthiness" or strength of reasons for choosing O, which need not be cashed out in consequentialist terms. For example, we noted above that things like the identity of the agent can be treated as part of each outcome. So, at a minimum, the utility of O can reflect some agent-relative considerations that standard forms of consequentialism exclude (cf. Brown 2011: 760–763). We will say more about non-consequentialist applications of expected utility theory in section 4.

One further subtlety is worth mentioning. We have said that $U(O)$ represents the choiceworthiness of O; all this means, however, is that more choiceworthy outcomes have higher utility. It doesn't follow that an outcome that is, say, *twice* as choiceworthy (by whatever measure) must have twice the utility. For example, a total utilitarian might identify the choiceworthiness of an outcome with its total welfare, but in combining

total utilitarianism with expected utility theory it is not logically required that an outcome with twice the total welfare must have twice the utility: in principle, U could be *any* increasing function of total welfare. Another way to put this point is that, even if permissibility is governed by expected utility theory, we are not necessarily required to be *risk neutral* with respect to choiceworthiness; perhaps we ought to be risk averse, risk seeking, or some mix.[13] Still, when utility does agree with some independently given measure of choiceworthiness or goodness we may refer to "expected choiceworthiness" or "expected value" instead of "expected utility".

Finally, classic work related to expected utility theory illustrates the following general lesson: views about moral decision-making under uncertainty can inform and constrain our views about moral decision-making even in the absence of uncertainty. Key examples are provided by Harsanyi's aggregation theorem (Harsanyi 1955, 1977) and the subsequent literature. Harsanyi derived a form of utilitarianism, using expected utility theory, impartiality, and the *ex ante* Pareto principle as essentially the only premises. (The *ex ante* Pareto principle says, roughly, that one option ought to be chosen over another if it is better for each person separately, taking uncertainty into account. Importantly, Harsanyi's utilitarianism is restricted to fixed populations, i.e., it does not tell us how to compare populations of different size or with different membership.) See Broome (1991) for a classic philosophical development of Harsanyi's theorem, and Greaves (2017) for further discussion of the connection between Harsanyi's result and classical utilitarianism. This result has been extremely influential in distributive ethics and population ethics, including the economics-based literature on "social choice". As a small sample from a vast literature: Harsanyi's approach has been extended to treat variable-population cases, whether in a utilitarian vein (e.g., Broome 2004; Blackorby, Bossert, & Donaldson 2005) or a non-consequentialist, person-affecting one (e.g., Thomas 2023); and to allow for more general forms of

utilitarianism (e.g., McCarthy, Mikkola, & Thomas 2020). Moreover, his theorem and associated ideas have been used to clarify what is at stake in debates between different distributive views such as egalitarianism and prioritarianism (e.g., Diamond 1967; Myerson 1981; Otsuka & Voorhoeve 2009; McCarthy 2015, 2017).

3. Cluelessness and Deep Uncertainty

Now we turn to a fairly general problem about decision-making under uncertainty. The problem is arguably most severe for maximizing act consequentialism and similar theories; insofar as this is true, it further illustrates how thinking about choice under uncertainty can inform our moral theorizing. A broader lesson is that there are arguably a number of forms of "deep uncertainty" that are highly relevant to moral decision-making but seem to evade simple probabilistic models like the one implicit in standard forms of expected utility theory.

The problem, in brief, is that we often feel clueless about the long-term (or otherwise non-immediate) consequences of our actions; insofar as what we ought to do depends on those consequences, this suggests that we are often clueless about what we ought to do. Such "cluelessness" has been seen by some as an objectionable feature of moral theories—including, but not limited to, standard forms of act consequentialism—that give significant weight to such long-term consequences. Why objectionable? Perhaps a desideratum for a moral theory is that it usually provides actionable advice. More specifically, though, insofar as we *are* clueless about the long-term effects of our actions, it seems plausible that we should be able to reliably figure out what to do by simply ignoring those effects. But it is at least unclear how act consequentialism (and theories that worry about consequences in a similar way) can license such a move. This objection was developed at length by Lenman (2000), starting from a line of thought sketched but not endorsed by Kagan (1998: 64).

One natural interpretation of "cluelessness" is that one has *no evidence* concerning the relevant consequences. Suppose right now I can pick up a pen either with my left hand or with my right. For the sake of illustration let us grant that it is predictable that, through the "butterfly effect", doing one of these actions rather than the other will lead to a greater number of destructive typhoons over the next millennium. Nonetheless, I seem to have no evidence whatsoever *which* action would do so. To put it another way, when it comes to typhoons, the evidence in favour of using my left hand seems to be perfectly symmetrical with the evidence in favour of using my right. Surely, then, no matter how bad the additional typhoons would be, I can simply set this consideration aside. However, this conclusion appears to rely on a "principle of indifference" to the effect that, given my lack of evidence, the probability that my left hand leads to more typhoons is the same as the probability that my right hand does. Tempting as this may be, indifference principles are "notoriously vexed" (Lenman 2000: 354; see also section 4.2 of the entry on Bayesian epistemology) and part of Lenman's argument is scepticism that a strong enough indifference principle is available. Greaves (2016), in contrast, defends indifference reasoning in cases like this of "simple cluelessness" involving evidential symmetry. However, she thinks that there is still a problem with cases of "complex" cluelessness, cases in which there is different evidence pointing in each direction but it is unclear how to weigh it up. Even if Greaves is right that simple cluelessness is unproblematic, this may not help much, if genuine evidential symmetry is not the normal case (Yim 2019; Greaves 2016, VII).

Diagnosing cluelessness in terms of a lack of evidence or in terms of complex evidence still does not tell us why cluelessness about consequences leads to cluelessness about what one ought to do. Indeed, as we have seen, a moral theory's verdicts about what we ought to do often take into account uncertainty about the consequences of our actions. So, uncertainty about the consequences of our actions does not imply

uncertainty about what we ought to do. For example, Jackson's doctor is uncertain what pill will have the best consequences for her patient, but nonetheless knows, in light of that uncertainty, which pill she ought to prescribe. To generate a problem, the sense in which we are clueless about the long-term consequences of our actions must have some upshot that makes it harder to deal with than other sources of uncertainty.[14]

One possible upshot of cluelessness is that the decision-relevant probabilities are hard to know or even to estimate precisely. The issue of precision is important, as is the idea that the long-term consequences of one's actions can have large (perhaps even infinite!) value. One might have thought that most of our actions will turn out to have little net impact on the far future: the effects die out (or cancel out) over time "like ripples in a pond" (Smart 1973: 33). Lenman argues against this claim, based in part on the common view that the identities of future people depend sensitively on what we do now. Instead, at least some possible consequences of our actions have very high value, systematically affecting many people over a long span of time. But if some possible consequence of our choice is large in value, then a small change in the probability of that consequence will make for a large change in expected value. If you cannot estimate the probabilities very precisely, then, you could be clueless about the expected values of your options and about what you ought to do. [15] When thinking about long-term consequences, such precise estimates seem hard to come by. (We've phrased this and much of the rest of the discussion in terms of expected value maximization, but it should be clear that the issues are more general.)

This kind of problem can obviously arise if the decision-relevant probabilities are objective: for example, if the decision relevant probabilities are the ones objectively mandated by my evidence, then I will have difficulty knowing what to do, if I do not know what my evidence is or what it supports. And it seems especially difficult to

scrutinize how my evidence bears on the long-term consequences of my actions. This thought fits well with Greaves's notion of complex cluelessness mentioned above. But a version of this problem might arise even if the relevant probabilities are subjective, corresponding simply to the agent's credences: there may be cases where these credences are difficult to introspect to the required precision. (More generally, the view that mental states are not "luminous" [Williamson 2000] is relevant here.) At any rate, on this first reading, the problem with cluelessness is that not only is there uncertainty about the consequences of one's actions, but one is also unable to access the decision-relevant probabilities to sufficient precision.

A different possible upshot of cluelessness is that the decision-relevant probabilities are *themselves* imprecise or in some sense indeterminate (Greaves 2016; Mogensen 2021). A common way to represent this sort of uncertainty is, not by a probability measure p over outcomes, but by a set S of such probability measures; see the entry on imprecise probabilities for further discussion. For each option, we will end up with a range of expected values, one for each p in S, and there is nothing more precise or determinate to be said about where the "true" expected value lies within that range. So, the thought is, cluelessness with respect to the long-term consequences of our actions means that the probabilities of those consequences must be somewhat indeterminate; moreover, insofar as the values of some of those consequences are very large, the expected values of our options will tend to be highly indeterminate as well.

To see why this might be problematic, we have to ask what one ought to do in light of all this indeterminacy or imprecision. Let us indicate three types of views, two of which lead to a kind of cluelessness objection. (See the entries on decision theory and on rivals to expected utility theory for further views and references.)

1. Insofar as it is indeterminate which options have higher expected value than which others, it is indeterminate what one ought to do. If it is indeterminate what one ought to do, then one can't (at least normally) know what one ought to do, and this might seem objectionable if it is sufficiently widespread. (In the context of cluelessness, Greaves [2016] tentatively endorses this interpretation, following Rinard [2015].)
2. Some views of decision-making with imprecise probabilities are highly permissive: if option x has higher expected value than option y with respect to even just one p in S, then it is permissible to choose x over y. In this case, our cluelessness about the non-proximate consequences of our actions may lead, not to cluelessness about what we ought to do, but to an implausibly nihilistic theory on which, in practice, everything is permitted. (Mogensen [2021] develops this worry.)
3. On the other hand, some theories of decision-making with imprecise probabilities, discussed in the context of "ambiguity aversion" (see section 4.3 of the entry on rivals to expected utility theory), lead to determinate verdicts about what one ought to do, while avoiding nihilism. (Of course, cluelessness may still arise in other ways, e.g., because the relevant probabilities or other decision parameters cannot be precisely estimated.)

These views raise interesting questions about the role of ambiguity attitudes in moral choice (see, e.g., Bradley 2022; Buchak 2023).

A different form of "cluelessness" might arise from *unawareness* of the possible consequences of our actions, rather than from sparse or complicated evidence. See Bradley (2017) and Steele & Stefánsson (2021) for introductions to unawareness, and Roussos (2021—see Other Internet Resources) for cluelessness as unawareness. The general idea is that, before we can properly take into account any uncertainty about some

possibility *P*, we must be aware of *P*; *P* must be on our radar. We might not be aware of a possibility because we lack the proper conceptual repertoire. More simply, it might just be a possibility which has never occurred to us, or even if it is something to which we have given long hard thought in the past, we may, being bounded agents, simply fail to bring it to mind when making a decision. All three of these types of unawareness might be particularly pervasive when it comes to thinking through the manifold long-term consequences of our actions, and might leave us with no reliable method of determining what we ought to do. Following Nozick (1974: 313–4), it may help to imagine neolithic people trying to anticipate their effects on twenty-first century society—it seems they would not and arguably could not entertain the key questions. On the other hand, as Steele & Stefánsson (2021, 2022) effectively argue, perhaps there are awareness-relative norms (in analogy to belief- or evidence-relative norms) that we can still try to follow in many cases of unawareness.

So far, we have written as if, in the relevant cases, one cannot get much of an idea of what one ought to do without first doing something along the lines of an explicit expected value calculation, eliciting the probabilities of various outcomes with adequate precision. But even if expected value maximization is the criterion of rightness, it does not follow that calculating expected value is the only, the best, or even a viable decision-procedure (Railton 1984; Jackson 1991; Feldman 2006). The literature on heuristics (see, e.g., Gigerenzer & Gaissmaier 2011) and on "decision-making under deep uncertainty" (see, e.g., Helgeson 2020; Marchau, et al. 2019) can be interpreted as proposing methods for decision-making that are more tractable and that bypass various kinds of cluelessness (Thorstad & Mogensen 2020; Mogensen & Thorstad 2022; Steele & Stefánsson 2021: 8.4). But, at least at a first glance, this only changes the target of the cluelessness objection: aren't we also clueless about which decision-procedures will do well in any given case (cf. Mogensen & Thorstad 2022: 3.2)? While environmental feedback and empirical study can help identify

procedures that tend to perform well with respect to relatively short-run and familiar sorts of consequences, it is at least unclear how to identify procedures that would tend to perform well with respect to very long-term consequences, about which feedback is much harder to obtain.

Things are arguably rosier if we ask, not which decision-procedure will lead to the right act in a given situation, but which decision-procedure it would be best to adopt *repeatedly*, to cover a range of future decisions. Similarly, we can ask, in the spirit of rule consequentialism, what decision-procedures it would be best for everyone to adopt. In some cases, at least, the long-term effects of universally and/or repeatedly applying a decision-procedure may be easier to predict than those of applying it once-off (Burch-Brown 2014). If so, rule consequentialism (and, more generally, moral theories that worry about the consequences of widespread rule-adoption, rather than the consequences of individual actions) may be in a better position than act consequentialism when it comes to cluelessness, although how much better is difficult to tell.

4. Moral Constraints Under Risk

The problem of cluelessness suggests that the standard decision-theoretic framework of expected utility maximization may be difficult or impossible to apply in practice for any agent who gives moral weight to long-term consequences, since these consequences can be enormously complex and unpredictable. In contrast, when we turn our attention to non-consequentialist moral considerations in the context of uncertainty, a central worry is that the expected utility framework is inapplicable *in principle*—it cannot do justice to all the things non-consequentialists care about—while at the same time plausible alternatives to that framework are very hard to come by. We now turn our attention to this challenge, with a particular focus on how non-consequentialists should evaluate actions that carry risks of harm to others.

A distinctive feature of many non-consequentialist moral theories is *agent-centred constraints*. An agent-centred constraint is, roughly, a moral reason against taking a particular type of action (e.g., killing, stealing, lying) that is not matched by an equally strong reason to *prevent* other actions of that type. Thus, for instance, one should not kill even to prevent multiple killing, or lie even to prevent multiple lies.[16] Some non-consequentialists (e.g., Kant 1797 [1996]; Anscombe 1958; Gewirth 1981, 1982; Finnis 2011: 223–226) have famously held that some agent-centred constraints are *absolute*, i.e., ought never be violated no matter the consequences.

Just as an agent can be uncertain about the consequences of her actions, however, she can be uncertain whether a given action would violate an agent-centred constraint. Sometimes these uncertainties are one and the same, since the constraint in question prohibits acts with certain consequences. For instance, I may be unsure whether a given action would result in someone's death, and therefore uncertain whether it violates an agent-centred constraint against killing. But there are other ways of being uncertain whether an action violates a constraint. For instance, perhaps it is permissible to kill a wrongful aggressor in self-defence, but impermissible to kill an "innocent threat" (someone who endangers you through no fault of their own, e.g., out of non-culpable ignorance), and you find yourself endangered by someone without knowing whether they are a wrongful aggressor or an innocent threat. Perhaps there is a constraint against breaking your promises, but you can't remember the details of some past promise and are therefore uncertain whether a particular action would violate it. Perhaps there is a constraint against acting with certain motives (e.g., to deceive) and you are unsure of your own motives.

The non-consequentialist, even the absolutist, cannot plausibly claim that any action that carries a *risk* of violating a constraint is impermissible. For

one thing, it's plausible that *all* our actions carry such a risk—for instance, anything you do *might* cause the death of an innocent person. But even if it's possible to be certain that you aren't violating a constraint (e.g., by simply doing nothing at all), the demand for such certainty is clearly excessive. It's permissible to drive your friend to the airport even though this entails *some* risk of hitting and killing an innocent pedestrian. It's permissible to take mundane actions that have some non-zero but remote probability of violating a long-forgotten promise.

A natural suggestion for the constraint-theorist to make is that there is some *probability threshold t* such that an action is permissible only if its probability of violating an agent-centred constraint is less than *t*. (Or, more plausibly, perhaps there are different thresholds for different constraints.) For instance, perhaps an action is permissible only if its probability of killing an innocent person is less than 0.1%.

One obvious worry is that the threshold proposal seems arbitrary—what principled basis could we give for claiming that the maximum acceptable risk of violating a deontological constraint is, say, 0.1% as opposed to 0.2%, or 0.01%, or…? (This point is made by, for instance, Portmore 2017: 293–4 and Jackson & Smith 2016: 284.) In addition, the threshold approach faces an "agglomeration" problem: it is hard to extend in a plausible way to sequences of risky choices and to choices involving multiple risks. Jackson and Smith (2006: 276) illustrate this problem with the following case.

Two Skiers

Two skiers are headed down a mountain slope, along different paths. Each, if allowed to continue, will trigger an avalanche that will kill different groups of ten innocent people. The only way to save each group is to shoot the corresponding skier dead with your sniper rifle. The moral theory you accept tells you that you ought to kill culpable

aggressors in other-defence, but are absolutely prohibited from killing innocent threats. Unfortunately, you are uncertain whether the skiers know what they're doing—they may be trying to kill their respective groups, or they may just be oblivious. Specifically, you assign each skier the same probability p of acting innocently, and the probabilities for the two skiers are independent.

Suppose that the threshold for permissible risk of violating the constraint against killing an innocent threat is t, so that, considering each skier in isolation, it would be permissible to shoot if and only if the probability p that they are acting innocently is less than t. And suppose that, while p is indeed less than t, the probability that *at least one skier* is acting innocently $(1 - (1 - p)^2 = 2p - p^2)$ is greater than t. Then, it seems, although you are permitted—perhaps even obligated—to shoot Skier 1, and to shoot Skier 2, you are prohibited from taking the combined action *shoot Skier 1 and shoot Skier 2*!

The deontologist might reply that we should evaluate each action in isolation, so that you are permitted to take the combined action because each of its component simple actions carries only an acceptable level of moral risk. But now suppose that it is *only* possible to kill both skiers with a single action—for instance, you only have one bullet, but are such a marksman that you can kill both skiers with a single shot. Then this single action would carry a risk greater than t of violating an absolute moral constraint, and so be prohibited. But it does not seem plausible that it is permissible to shoot both skiers if you use separate bullets, but impermissible to shoot them both with one bullet.

In reply to Jackson and Smith's argument, Aboodi, Borer, and Enoch (2008) suggest that deontologists should adopt an individualistic, patient-centred approach on which the threshold for acceptable risk of committing a rights violation is indexed to particular rightsholders; i.e., what matters is

not that an agent runs a total risk less than t of violating anyone's rights but rather that, for each individual rightsholder, she runs a risk less than t of violating their rights. Thus, in Two Skiers, you ought to shoot both the skiers because even though the total risk of wronging someone is greater than t, the risk of wronging either skier individually is not. But, as both Huemer (2010) and Jackson & Smith (2016) point out, the same agglomeration problem can reappear in cases involving multiple possible infringements on the same rightsholder. Further, as Huemer (2010) notes, the patient-centred approach will implausibly prefer to subject many individuals to risks just below the probability threshold—perhaps guaranteeing that *someone* will suffer serious harm—rather than subject a single individual to a risk just above the threshold.

Given the difficulties of the threshold approach, we might try instead to accommodate deontological considerations within the framework of expected utility theory. It is commonly recognized that many apparently non-consequentialist moral theories can be "consequentialized"—that is, we can state an extensionally equivalent moral theory in consequentialist language (see the entry on consequentializing). For instance, a theory that says it is always wrong to kill an innocent person even to save a greater number of innocent people can be consequentialized as a theory according to which any outcome where the agent has killed an innocent person is worse than any outcome where the agent has merely let innocent people die. In a similar manner, we may be able to represent a non-consequentialist theory's ranking of risky options as maximizing the expectation of a utility function, even if the theory's actual explanation or justification for that ranking makes no reference to "utility" or its expectation (Colyvan, Cox, & Steele 2010). (Indeed, as noted in section 2, expected utility theory on at least one orthodox interpretation does not claim that an option is permissible *because* it has maximal expected utility, but only that the facts about permissibility are such that there is an

extensionally accurate representation of them in terms of maximizing the expectation of a utility function.)

One possibility is to represent absolute constraints by means of lexicographic utilities. For instance, we might represent the objective choiceworthiness of acts with two-component vectors of real numbers, lexicographically ordered (meaning that $(x_1, y_1) \geq (x_2, y_2)$ if and only if (i) $x_1 > x_2$ or (ii) $x_1 = x_2$ and $y_1 \geq y_2$), with the first component taking value -1 if the act violates a constraint and 0 otherwise. An agent is then supposed to maximize expected utility (which is also a vector, found by taking the expectation of each component separately). This approach, however, seems implausibly fanatical, since it requires us to to minimize the probability of constraint violations at any cost.[17] See however Lee-Stronach (2018), who combines the lexicographic approach with a form of small-probability neglect in order to avoid this conclusion.

Another possibility is to represent constraint theories using ordinary, real-valued utilities. This is a natural approach for non-absolutist theories: If, for instance, a theory says that there is a constraint against killing an innocent person that can be overridden only to save 1000 or more innocent people, it is natural to model the theory as assigning a disutility to killing as is 1000 times greater than the disutility of letting die. But perhaps surprisingly, even absolutist theories can arguably be represented in the framework of ordinary, non-lexicographic expected utility theory. The trick is to assign lower-priority considerations (e.g., saving lives) diminishing marginal utility and bounded total utility, while assigning higher-priority considerations (e.g., not killing innocent people) utilities whose magnitudes exceed those bounds (Lazar & Lee-Stronach 2019; Black 2020). For instance, perhaps the utility of saving n lives is bounded above at 100, while the disutility of killing an innocent person is -1000. In this case, it is never permissible to kill an innocent person with certainty to save any number of lives. More generally, we might claim that

differences in the utility of acts arising from the impersonal value of their consequences are bounded above at some finite value b, and that the disutility of violating an absolute constraint is greater than b.

This view has the attractive feature of allowing small enough risks of constraint violations to be outweighed. (For instance, in the example above, we may run up to a 10% risk of killing an innocent person to save a large enough number of lives.) It also lets the deontologist represent more or less stringent constraints by larger or smaller disutilities, so that a small probability of violating a constraint is harder to justify if the constraint is more stringent. On the other hand, it inherits the various drawbacks of bounded expected utility as an approach to moral decision-making—for instance, extreme risk-aversion near the upper bound and risk-seeking near the lower bound (Beckstead & Thomas forthcoming), and potential violations of *ex ante* Pareto principles (Kosonen 2022: Ch. 1). We discuss these drawbacks in section 6 below.

Even if it can provide an extensionally plausible treatment of agent-centred constraints, it is up for debate whether non-consequentialists should embrace expected utility theory. Seth Lazar, who defends an expected utility-based approach to deontological ethics under risk, nevertheless argues that the expected utility framework requires substantial modifications to accommodate features of non-consequentialist morality like agent-centred options (Lazar 2017b) and the moral significance of sunk costs (Lazar 2017a). But other philosophers (e.g., Hansson 2003; Tenenbaum 2017) are sceptical of any expected utility-based approach. Here is one reason for such scepticism: A basic feature of orthodox expected utility theory is that the utility of taking a given act in a given state of nature does not depend on the probability of that state, on what other states were possible, or on their probabilities. But deontologists may claim that the *objective* (as well as the subjective/prospective) moral character of an act depends on the agent's beliefs and evidence in ways

that clash with this principle. For instance, suppose that *A* takes an action that recklessly endangers *B*, but luckily results in no harm. A deontologist might judge that this act is not, even *ex post*, equivalent to an act that was certain to be harmless all along. For instance, by acting with reckless disregard for *B*'s welfare, *A* may have violated *B*'s rights, or failed to treat her as an end in herself. If so, deontologists cannot treat risky actions simply as probability distributions over risk-free actions; more formally, they cannot evaluate the utility of an outcome independent of its probability and the probabilities of alternative outcomes, as expected utility theory requires.[18]

The debate we have surveyed thus far is focused on questions of individual morality—how individuals should act when they face risks of violating constraints. But there is also a parallel, closely related literature on "risk imposition" that is primarily focused on how governments and other social institutions should manage and regulate risk.[19] In this context, it is even more obvious that simple rules like an absolute prohibition on imposing *any* risk of grave harm on an innocent person, or a requirement to keep the probability of ever harming an innocent person below some fixed threshold, are unworkable. Any criminal justice system will run some risk of convicting and imprisoning innocent people, and in any sufficiently large system, it is all but inevitable that this risk will be realized at least once. Similarly, in permitting a new technology that causes even the smallest amount of pollution or carries even the smallest risk of accidents that harm non-consenting parties, governments allow many people to be subjected to risk, often with near-certainty that some will suffer serious harm.

But in the context of social and political institutions, a new set of considerations also becomes salient: rather than simply considering the *prospects* of risky actions (e.g., the amount of good that can be achieved by a risky act, the probability that the act would violate a constraint, and

the stringency of the constraint it would violate), it is natural to ask whether the act in question comports with a system of general norms that is just, fair, or otherwise legitimate. This might suggest principles like the following:

> Exposure of a person to a risk is acceptable if and only if this exposure is part of an equitable social system of risk-taking that works to her advantage. (Hansson 2003: 305)

Contractualists about individual morality, for whom the correct moral norms are those to which ideally reasonable individuals would consent or which no individual could reasonably reject (see the entry on contractualism), might be inclined to take a similar approach to cases of interpersonal (rather than institutional) risk imposition. But contractualists face an important tension here: On the one hand, reasonable individuals might—indeed, often do—accept small risks of severe harm in exchange for relatively minor benefits. (For instance, biking to the store might carry a slightly greater risk of fatal accident than walking, but this risk may be an acceptable price to save half an hour.) So it seems that contractualists should endorse norms that accept this sort of tradeoff—in particular, that permit every individual to impose very small risks of catastrophic harm on others (e.g., by driving on residential streets) for the sake of relatively minor benefits. On the other hand, the aggregate effect of everyone accepting and acting on such a norm will almost certainly be that a few people suffer catastrophic harms, while a much larger number of people receive minor benefits. And this *result* seems hard to justify to the few who are catastrophically harmed, given that contractualists characteristically wish to disallow the aggregation of many minor claims to outweigh individually much weightier claims. This dilemma is the fault line between "*ex ante*" contractualists, who focus on whether any individual could reasonably reject a given norm *in advance* (on the basis of the *prospect* that norm confers on them), and *ex post* contractualists,

who focus on whether any individual could reasonably reject a given norm *after the fact* (on the basis of the *outcome* they receive as a result of that norm). This debate has generated a substantial literature, with influential contributions including Ashford (2003), Lenman (2008), Fried (2012), Frick (2015), and Kumar (2015), among others. We do not explore this literature here, however, because it is well covered in section 11 of the entry on contractualism.

A final related question is whether, when we take actions that expose another person to risk (and perhaps also actions that confer uncertain benefits), we should take account of that person's particular attitudes and preferences with respect to risk. For instance, suppose that a government with a limited budget can focus either on reducing air pollution (which will have a fairly uniform effect on everyone's health, increasing average lifespan by 6 months) or on reducing fatal traffic accidents (which will reduce the probability of a rare but individually catastrophic outcome, increasing average lifespan by 3 months). From a self-interested standpoint, different members of the public might have different preferences between these interventions: a risk-neutral individual will prefer the former, but a sufficiently risk-averse individual might prefer the latter. How much weight, if any, should the government give to these individual preferences when deciding between the two projects? Similar questions arise where one individual must make risky choices on behalf of another, e.g., as a medical proxy. (Consequentialists confront these questions as well, but they seem more difficult for non-consequentialists, for whom weighing harms and benefits to others is not simply a matter of maximizing the impartial good but is influenced by ideals like respect and autonomy.)

This question has only recently begun to receive substantial attention. On the side of deference to individual risk preferences, Buchak (2017) has argued that, when making decisions that affect an individual with known

risk attitudes, the potential effects on that individual should be evaluated according to their own risk attitude. Thus, for instance, if the population is sufficiently risk-averse, the government in the preceding example might be required to focus on reducing traffic accidents even if reducing air pollution would almost certainly yield greater total benefit. On the other hand, Bovens (2015) has argued that it is reasonable for an agent making decisions on behalf of another to be *more* risk-averse than the patient herself. For a survey of this debate, see Thoma (2023).

5. Moral Uncertainty

So far we have considered some of the distinctive problems facing consequentialists and non-consequentialists respectively in dealing with uncertainty. But what if you are uncertain whether consequentialism or non-consequentialism is true? More generally, what if you're uncertain about any basic question of morality, and face a choice where different moral principles recommend different actions? How should you decide what to do in light of that uncertainty? This is often referred to as the problem of *decision-making under moral uncertainty*, where "moral uncertainty" means uncertainty about fundamental moral principles (as opposed to uncertainty about morally relevant empirical facts).

Aspects of this question were discussed by moral theologians in the seventeenth century, in a debate that pitted morally cautious Jansenists (who held that you ought to avoid actions that have even a small probability of being morally wrong) against more permissive Jesuits (who held that, for instance, one may take an action if the principle that permits it is at least a probable as the principle that forbids it). For a summary of these debates, see Sepielli (2010: 48–53). Toward the end of the twentieth century, these questions began to reemerge (e.g., in Hudson 1989; Gracely 1996), with a particular interest in the ethics of abortion (Greenwell 1977; Pfeiffer 1985). But since the publication of Ted Lockhart's *Moral*

Uncertainty and Its Consequences (2000), the problem of decision-making under moral uncertainty has received much more sustained and general attention.

One prominent view in this debate holds that there is no problem, because one's beliefs and uncertainties about basic moral principles have no bearing on what one ought to do, in any interesting sense of *ought*. Following Weatherson, we will call this view *externalism*, since it claims that what an agent ought to do is determined by moral truths external to (or at any rate, independent of) her mental states, rather than internal features of the agent like her beliefs or evidence concerning moral principles. Versions of this view are defended by Weatherson (2014, 2019), Harman (2015), and Hedden (2016), among others. Weatherson argues that acting in a way that is sensitive to one's beliefs about basic moral principles requires "*de dicto*" moral motivation (motivation *to do the right thing, whatever that turns out to be*), and that *de dicto* moral motivation manifests an objectionably "fetishistic" concern for rightness as such, as opposed to the things in the world that actually matter morally. Harman argues that, if what one subjectively ought to do depended on one's moral beliefs, then ignorance of basic moral principles would be exculpatory, which (she claims) it isn't. (For replies to Weatherson and Harman respectively, see Sepielli 2016, 2018a.) Hedden argues that decision rules that are sensitive to an agent's moral beliefs would have to overcome the "problem of intertheoretic value comparisons" (discussed below), and that this problem is insuperable. All these philosophers then conclude, with minor variations, that what an agent ought to do (in the relevant, subjective sense of "ought") depends on the *true* moral principles as well as the agent's beliefs/evidence about non-moral matters, but not on her beliefs/evidence concerning moral principles.

Many participants in the debate, however, have found this view implausible. We don't know, and cannot easily find out, what the true

moral principles are, and yet we must decide what to do *somehow*, with only our uncertain beliefs to guide us. And it certainly seems like there's *something* significant to be said for an agent who does what she believes to be right instead of what she believes to be wrong, independent of whether her beliefs are in fact correct—for instance, that she is *morally conscientious* or is responding *rationally* to her beliefs (MacAskill, Bykvist, & Ord 2020: Ch. 1; Bykvist 2020). Moreover, Podgorski (2020) gives a clever argument that, in cases where an agent's moral and empirical beliefs are not probabilistically independent, views that ignore an agent's moral beliefs will sometimes recommend dominated acts (acts that are certain to be morally worse than some available alternative).

If what an agent ought to do does depend on her moral beliefs, then we face the question of how agents should act when they are morally uncertain. One simple proposal, which Lockhart (2000) dubs "My Favorite Theory" (MFT), is that she should simply follow the moral theory in which she has greatest credence, disregarding all other moral theories. This view has been defended by Gracely (1996) and Gustafsson and Torpman (2014). But it faces substantial difficulties. First, it simply seems intuitively implausible to ignore all moral possibilities except one. For instance, suppose you have 49% credence that members of some animal species S have moral standing, and can take some action that would very slightly benefit a human being while imposing massive harm on members of S. Even if the most probable theory says that you ought to take this action, it seems reckless to disregard the substantial probability that it would be deeply morally wrong. Second, the implications of MFT are sensitive to how we individuate moral theories, i.e., how we decide whether to treat two moral views as versions of the same theory or as different theories. The most natural approach, adopted by Gustafsson and Torpman, is to individuate theories finely, treating T_1 and T_2 as versions of the same theory only if it is impossible for them ever to disagree about what one ought to do. But then you may well find that you have credence

in a vast number of moral theories (for instance, many different prioritarian theories with slightly different priority weighting functions), and that your "favorite" theory commands only a tiny portion of your credence (say, less than 1%). The implication that your actions might be guided exclusively by such a tiny portion of your credence distribution seems implausible. Finally, Gustafsson himself recants MFT (in Gustafsson 2022) because of its problems with sequential choice situations: It can recommend sequences of actions that are worse in expectation than an available alternative according to *every* moral theory in which the agent has positive credence.

An alternative approach, dubbed "My Favorite Option" (MFO) by Gustafsson and Torpman (2014), holds that a morally uncertain agent should choose the *option* that has the greatest probability of being morally right or permissible, taking the assessments of all moral theories in which she has positive credence into account. But this view faces difficulties closely analogous to MFT's. First, it will similarly recommend actions that are probably just slightly better than their alternatives, but have a substantial probability of being catastrophically morally worse. Second, just as MFT is sensitive to how we individuate *theories*, MFO is sensitive to how we individuate *options*. And third, MFO can generate cyclic pairwise comparisons of options (e.g., telling you to choose O_1 over O_2, O_2 over O_3, and O_3 over O_1), which can lead agents to make sequences of choices that are certainly worse than available alternatives—an even worse form of sequential failure than that resulting from MFT (Gustafsson & Torpman 2014: 165–6).

Both MFT and MFO are sensitive only to whether a given action is morally right or permissible according to a given moral theory. They take no account of the relative size of *differences* in moral value or choiceworthiness between alternative actions. But, as in the "species S" example above, it is intuitive that these cardinal facts make a difference to

how we ought to act under moral uncertainty: It is better, for instance, to do something that has a 51% probability of being just *slightly* wrong than to do something that has a 49% probability of being a moral atrocity. The idea that we should care about the magnitude as well as the probability of potential moral considerations under moral uncertainty is sometimes referred to as "moral hedging".

A natural suggestion, if one finds moral hedging plausible, is that we should use expectational decision principles to respond to moral as well as empirical uncertainty. That is, an agent making decisions under moral uncertainty should aim to maximize *expected moral value*, *expected rightness*, or *expected choiceworthiness*, with respect to her distribution of credence over rival moral theories as well as empirical hypotheses. Versions of this view have been defended by Lockhart (2000); Ross (2006); Sepielli (2009); MacAskill & Ord (2020); and MacAskill, Bykvist, & Ord (2020); among others. Though these authors use different terminologies, we will follow the most common terminology in the recent literature and refer to these views collectively as *maximize expected choiceworthiness* (MEC).

Any expectational approach to moral uncertainty faces a major challenge: To take an expectation over different moral theories, we must be able to *compare* the magnitude of differences in choiceworthiness according to rival theories. For instance, suppose you face a trolley problem where you can save five innocent people by killing one, and that you split your credence evenly between classical utilitarianism and a deontological theory that absolutely prohibits killing innocent people. Which option maximizes expected choiceworthiness in this situation depends on whether *the unchoiceworthiness of four (net) deaths, according to classical utilitarianism* is greater or less than *the unchoiceworthiness of violating the constraint against killing, according to deontology*. But does this sort of cross-theory comparison even make sense? What could make it the case

that one of these quantities is greater than the other, particularly given that at least one of the theories in question has to be false? And even if there are facts about intertheoretic choiceworthiness comparisons, how could we possibly know them? These difficulties comprise the *problem of intertheoretic comparisons*, one of the central problems in the literature on decision-making under moral uncertainty.

Some (e.g., Gracely 1996; Hedden 2016) have held that this problem is insuperable, and that we therefore must adopt some view like externalism or My Favorite Theory that does not require intertheoretic comparisons. But others have tried to rescue expectational approaches to moral uncertainty from the problem of intertheoretic comparisons. One strategy, sometimes called the "common ground" approach (MacAskill et al. 2020: 133ff), tries to find points of agreement between rival moral theories that can serve as a basis for intertheoretic comparisons. For instance, suppose you are uncertain between total utilitarianism, which tells you to maximize total welfare, and a pluralistic theory that tells you to maximize a weighted sum of welfare and beauty. Another way of describing this situation is that you are certain that you should maximize total value, and certain that welfare has value, but uncertain whether beauty also has value. At least on some ways of filling in the details of your belief state, it is natural to claim that you should treat the value of a unit of welfare as independent of the question of whether beauty has value—that is, a unit of welfare has the same value regardless of whether utilitarianism or pluralism is true. If so, then we have a way of making comparisons between these two theories: they value a unit of welfare equally, and agree about the magnitude of choiceworthiness differences in all choice situations where only welfare (and not beauty) is at stake. Approaches of broadly this sort are suggested by Sepielli (2009) and Tarsney (2018). The most obvious drawback of this approach is that its scope appears quite limited—it may let us make comparisons between pairs of total consequentialist theories with overlapping values, but does not offer any obvious suggestions for making

intertheoretic comparisons between, say, utilitarian and Kantian moral theories.

Another strategy, the "universal scale" approach, starts from the idea that there is a single cardinal scale of degrees of moral choiceworthiness. Any moral theory must, on this view, be in the business of mapping possible actions to degrees of choiceworthiness on this scale. And in a given choice situation, uncertainty between rival moral theories can be reconstrued as uncertainty about the value of each available option on this scale. Approaches in this spirit are suggested by MacAskill, Bykvist, & Ord (2020) and Carr (2020).

One objection to the universal scale approach is that it's not at all obvious whether rival theories do in fact make use of the same scale. For instance, in the spirit of orthodox expected utility theory, one might claim that cardinal choiceworthiness is just a convenient way of representing a moral theory's ranking of risky prospects, not a genuine quantitative property that theories attribute to options. In a similar spirit, does it really make sense to treat two moral theories as rivals if they differ in their choiceworthiness mappings while still agreeing about the ranking of all possible actions and outcomes? And even if it does, how could we ever get evidence for one theory over the other, or have any basis for a particular distribution of credence between them?

A third strategy, the *statistical normalization* approach, starts from the idea that under moral uncertainty, each moral theory should be given "equal say" in an agent's decisions—or rather, a say proportional to its probability. Different ways of measuring the "say" or influence of a moral theory lead to different precisifications of this principle. For instance, Lockhart argues that, in any given choice situation, the *range* of each theory's choiceworthiness assignment (the difference in choiceworthiness between the best and worst available options) should be treated as equal

(Lockhart 2000: 84). But this principle has serious drawbacks. For instance, because it considers each choice situation in isolation, it will evaluate tradeoffs between the same pair of theories differently in different choice situations, potentially leading to sequences of choices that are worse than available alternatives according to every moral theory in which the agent has positive credence (Sepielli 2013: 586–7). If we try to avoid this problem by equalizing the range of each theory's choiceworthiness assignment across all *possible* options, then the view can't accommodate theories like classical utilitarianism whose choiceworthiness scales are unbounded (Sepielli 2013: 588).

Cotton-Barratt, MacAskill, & Ord (2020) propose that we equalize the *variance* of each moral theory's choiceworthiness assignment over the set of all possible options, instead of the range. This approach can accommodate at least some unbounded theories, and has attractive formal properties that plausibly reflect the principle of giving each theory weight proportionate to its probability (Cotton-Barratt et al. 2020: 80–86). But it also has drawbacks. For instance, because there are infinitely many possible options, the variance of a theory's value assignment over all possible options can be well-defined only if we impose a *measure* on the set of possible options, and it's not obvious that there is any principled, non-arbitrary basis for choosing such a measure.

A further challenge for expectational approaches to moral uncertainty, in addition to the problem of intertheoretic comparisons, is the *structural diversity* of moral theories (Tarsney 2021). To even reach the point at which we confront the problem of intertheoretic comparisons, we must assume that each moral theory represents the choiceworthiness of options on a one-dimensional cardinal scale. But many moral theories don't seem to do this. For instance, Kantianism evaluates actions by classifying them as satisfying or violating the Categorical Imperative; it does not say things like

> murdering an innocent person is ten times as wrong as lying, which is three times as wrong as going a year without acting on your imperfect duty of beneficence.

Commonsense morality seems to recognize gradations of moral choiceworthiness (e.g., a supererogatory act can be better than a merely permissible act; one wrong act can be worse than another), but does not commonly issue judgements about the relative magnitudes of these differences. And many moral philosophers believe that some values are incomparable or only imperfectly comparable, which cannot be represented by a simple interval scale of choiceworthiness. The aspect of this problem that has received most attention to date is the problem of *merely ordinal* moral theories, which rank options without assigning cardinal degrees of choiceworthiness. In this context there is a natural analogy with voting theory: Just as voting methods (like first-past-the-post, approval voting, or instant runoff) aim to aggregate ordinal information about the individual preferences of voters into a social choice, analogues of these rules can be used to aggregate ordinal information about the preferences of moral theories into an overall verdict about what to do under moral uncertainty. For proposals along these lines, see Nissan-Rozen (2012), MacAskill (2016), and Tarsney (2019).

There is a final notable difficulty facing any attempt to incorporate moral uncertainty into our decision-making: Just as we are uncertain which moral theory is correct, we may also be uncertain *which approach to decision-making under moral uncertainty* is correct (e.g., we may be uncertain between MFT, MFO, and various combinations of MEC with particular methods of intertheoretic comparison). If we must take our uncertainty about "first-order" moral theories into account, shouldn't we also take account of this "second-order" uncertainty—and of our "third-order" uncertainty about how to respond to second-order uncertainty, and so on *ad infinitum*? Weatherson (2014, 2019) suggests that this regress

problem is fatal for the whole project of finding rational principles for responding to normative uncertainty. But others have tried to find solutions. Sepielli (2014) argues that, even if we can never take *all* our uncertainties into account when making decisions, it may still be *better* (e.g., more rational) to account for more of our uncertainties, including higher-order normative uncertainties. And Trammell (2021) shows that, under some moderately strong assumptions, the various metanormative theories in which an agent has credence will converge in their ranking of options as we ascend the hierarchy of higher-order norms, so that at a certain point the agent can be certain of which option is best even if she is still uncertain between competing norms. If the assumptions of these convergence results are often satisfied, or if agents are rationally required to satisfy them, then the regress may be less problematic than it first appears.

6. Small Probabilities and Extreme Stakes

A final question that looms large in the context of moral decision-making under uncertainty is how to handle small probabilities of extremely good or bad outcomes—or, more generally, small probabilities that an action has some extremely morally weighty feature. This is an important question for non-moral decision-making as well. To give two famous examples, Pascal's Wager and the St. Petersburg game are both cases of prudential decision-making under uncertainty where small probabilities of extreme outcomes can carry counterintuitive weight, and much ink has been spilled trying to avoid these counterintuitive implications (see the entries on Pascal's wager and the St. Petersburg paradox).

But the question of how to handle these extreme risks is even more pressing in the moral context. Each of the last three sections illustrates one reason why this is the case. Insofar as morality requires us to take account of the interests of remote strangers, it both raises the stakes relative to

prudential decision-making and makes prediction more difficult, so that small probabilities of extreme outcomes (like causing or preventing the birth of the next Hitler) are the norm rather than the exception. Insofar as morality imposes absolute or near-absolute constraints, small probabilities of violating those constraints can carry enormous weight. And insofar as we must take account of fundamental moral uncertainty, small credences in moral theories that assign extreme moral importance to a particular choice can hijack our deliberations (Ross 2006; Beckstead & Thomas forthcoming).

Should we allow small probabilities to carry such great weight in our decisions? In other words, should we accept the following thesis/theses?

> *Fanaticism*
> For any finite degree of choiceworthiness (or value, rightness, etc.) d and any probability p, there is a finite degree of choiceworthiness d^+ such that an action that has choiceworthiness d^+ with probability p and is morally neutral with probability $(1-p)$ is preferable to an action that has choiceworthiness d for sure; and there is a finite degree of choiceworthiness d^- such that an action that has choiceworthiness d for sure is preferable to an action that has choiceworthiness d^- with probability p and is morally neutral with probability $(1-p)$.

If moral value is unbounded, then expected value maximization is fanatical. For instance, an agent who maximizes expected total welfare will give potentially unlimited weight to arbitrarily low-probability outcomes, since there is no upper or lower limit to the number of welfare subjects in the world or, therefore, to total welfare. But expected value maximization is not the only way of being fanatical. For instance, a *risk-weighted* expected utility maximizer (Buchak 2013) will exhibit fanaticism as long as her utility function is unbounded (both above and below) and her risk function is strictly increasing.

Many people find fanaticism deeply counterintuitive, in the moral as well as the prudential domains. There are two standard strategies for avoiding it. One is to simply *ignore* sufficiently small probabilities. This idea has a long history (for a useful survey, see Monton 2019), and has recently been revived and defended by Nicholas J. J. Smith (2014, 2016) and Monton (2019). But this idea faces many compelling objections. To begin with, the idea of treating non-zero probabilities like zero probabilities seems *ad hoc*, and the choice of any particular threshold seems objectionably arbitrary. The most straightforward versions of small-probability discounting, which instruct an agent to simply ignore states or outcomes with probabilities below a certain threshold, are implausibly sensitive to how states and outcomes are individuated (Beckstead & Thomas forthcoming), and can violate dominance principles (Isaacs 2014). And even more sophisticated formulations like "tail-discounting" (Beckstead & Thomas forthcoming) are vulnerable to money pumps (Kosonen 2022: 196–235). Finally, in the long run, small probabilities can add up, and a policy of ignoring them can have the cumulative result of accepting very large probabilities of extreme losses or foregoing very large probabilities of extreme gains (Lundgren & Stefánsson 2020; Thoma 2022).

The other standard strategy for avoiding fanaticism is to require a *bounded* utility function (in the context of expected utility theory or something similar). This view is natural and popular in the context of individual prudential decision-making, where most goods have diminishing marginal utility. For instance, if you could have \$1 million for sure, or a 10^{-100} chance of receiving a blank check good for *any* finite amount of money, it would be quite reasonable to choose the former option. If you are an expected utility maximizer, then this choice implies that the utility you assign to money is bounded.

But in the moral context, the bounded utility approach looks less promising. Intuitively, while providing more of the same material benefit

to a single person has diminishing marginal value, providing a given benefit to more people does not. (Does saving *C*'s life count for less because you have already saved *A*'s and *B*'s?) So the use of a bounded utility function cannot easily be motivated by the thought that value or choiceworthiness is itself bounded. Also, boundedness in the moral context can have deeply counterintuitive consequences, requiring extreme risk aversion near the upper bound of your utility function and extreme risk-seeking near the lower bound. For instance, if the utility of saving lives is bounded, there will be some n such that saving n lives with probability p has greater expected utility than saving *any finite number of lives* with probability $0.99p$ (Beckstead & Thomas forthcoming).

The idea that the moral value of benefits to one person does not depend on how many others have been benefited in fact hints at a powerful class of arguments for fanaticism in the moral context. Any view that assigns decreasing (or increasing) marginal utility to benefiting additional people will violate compelling principles to the effect that if one prospect is more desirable than another from the standpoint of some individuals or subpopulations considered in isolation, and equivalent from the standpoint of all other individuals/subpopulations, then it is more desirable overall. Arguments of this sort, focusing on "separability" principles for non-overlapping subpopulations, are discussed by Wilkinson (2022), Beckstead & Thomas (forthcoming), and Russell (forthcoming). But a particularly simple version of the argument can be stated using individuals rather than subpopulations. (What follows is a variation on Kowalczyk (forthcoming).) Consider a choice between the two prospects described in Tables 1 & 2. Here you have the opportunity to benefit some very large number of individuals n. The choice is between certainty of benefiting one of them at random, and a $\frac{1}{n}$ chance of benefiting *all* of them. In the latter case, additionally, the benefit to each individual will be greater. From the perspective of each individual, therefore, the second option looks better, so there is a compelling case for choosing it. But in so doing, you are

foregoing a sure moral good (benefiting one individual) for a potentially-minuscule probability ($\frac{1}{n}$, for arbitrarily large n) of an astronomically large good (a greater benefit to n people).

	$S_1\,(\frac{1}{n})$	$S_2\,(\frac{1}{n})$	$\cdots$	$S_n\,(\frac{1}{n})$
p_1	1	0	$\cdots$	0
p_2	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
p_n	0	0	$\cdots$	1

Table 1. "Sure Thing"

	$S_1\,(\frac{1}{n})$	$S_2\,(\frac{1}{n})$	$\cdots$	$S_n\,(\frac{1}{n})$
p_1	$1+\epsilon$	0	$\cdots$	0
p_2	$1+\epsilon$	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
p_n	$1+\epsilon$	0	$\cdots$	0

Table 2. "Long Shot"

In Tables 1 & 2, rows correspond to people, columns correspond to (equiprobable) states of nature.

This case illustrates a simple argument for a form of fanaticism: for any *improvement* to the world (like increasing one person's welfare from 0 to 1) and any probability p, we can find a situation in which certainty of that improvement is less desirable than probability p of some even greater improvement. The argument needs only three premises: a weak form of *ex ante* Pareto (to establish that the option in Table 2 is better than the option in Table 1), the claim that randomizing the beneficiary of an improvement (as in Table 1) doesn't change its value, and transitivity. But the form of fanaticism thereby established is somewhat weak, and is compatible with moral views that don't seem intuitively "fanatical": for instance, a view that tells us to maximize expected average welfare in a population, with individual welfare being bounded above and below. (This is a form of bounded expected utility maximization, and so not fanatical in the strong sense defined above.) To turn the preceding argument into an argument for that stronger thesis, we need additional premises—for example, one could add the premise that for any outcomes $O_1 \succ O_2$, we can construct an outcome at least as good as O_1 by adding good lives to O_2, and an outcome at least as bad as O_2 by adding bad lives to O_1.

On the other hand, there are also compelling axiomatic arguments *against* fanaticism. In particular, if fanaticism is true then, given very minimal auxiliary assumptions, we can construct "improper prospects" that we must prefer to all of their possible outcomes (Russell & Isaacs 2021; Beckstead & Thomas forthcoming; Russell forthcoming).[20] Improper prospects have a number of counterintuitive consequences. In particular, an agent who regards some prospect as better (or worse) than all its possible outcomes thereby violates infinitary versions of the principles that characterize expected utility maximization (the independence axiom and the sure-thing principle), and thereby exhibits many of the foibles of non-expected utility maximizers. For instance, she will sometimes make sequences of choices that are certainly worse than available alternatives, and will pay to avoid information (Russell & Isaacs 2021; see also the closely related argument for boundedness in Hammond (1998)). Insofar as you think that agents should *not* behave in those ways (a standard justification for expected utility theory), this is a compelling argument against fanaticism.

Bibliography

Aboodi, Ron, Adi Borer, and David Enoch, 2008, "Deontology, Individualism, and Uncertainty: A Reply to Jackson and Smith", *Journal of Philosophy*, 105(5): 259–272. doi:10.5840/jphil2008105543

Anscombe, G. E. M., 1958, "Modern Moral Philosophy", *Philosophy*, 33(124): 1–19. doi:10.1017/S0031819100037943

Ashford, Elizabeth, 2003, "The Demandingness of Scanlon's Contractualism", *Ethics*, 113(2): 273–302. doi:10.1086/342853

Aumann, Robert J., 1962, "Utility Theory without the Completeness Axiom", *Econometrica*, 30(3): 445–462. doi:10.2307/1909888

Bader, Ralf M., 2018, "Stochastic Dominance and Opaque Sweetening", *Australasian Journal of Philosophy*, 96(3): 498–507. doi:10.1080/00048402.2017.1362566

Bales, Adam, Daniel Cohen, and Toby Handfield, 2014, "Decision Theory for Agents with Incomplete Preferences", *Australasian Journal of Philosophy*, 92(3): 453–470. doi:10.1080/00048402.2013.843576

Bales, R. Eugene, 1971, "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?", *American Philosophical Quarterly*, 8(3): 257–265.

Balfour, Dylan, 2021, "Pascal's Mugger Strikes Again", *Utilitas*, 33(1): 118–124. doi:10.1017/S0953820820000357

Beckstead, Nick and Teruji Thomas, forthcoming, "A Paradox for Tiny Probabilities and Enormous Values", *Noûs*, first online: 3 June 2023. doi:10.1111/nous.12462

Black, D., 2020, "Absolute Prohibitions Under Risk", *Philosopher's Imprint*, 20: article 20. [Black 2020 available online]

Blackorby, Charles, Walter Bossert, and David J. Donaldson, 2005, *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*, New York: Cambridge University Press. doi:10.1017/CCOL0521825512

Bovens, Luc, 2015, "Concerns for the Poorly Off in Ordering Risky Prospects", *Economics and Philosophy*, 31(3): 397–429. doi:10.1017/S0266267115000188

Bradley, Richard, 2017, *Decision Theory with a Human Face*, Cambridge: Cambridge University Press. doi:10.1017/9780511760105

——, 2022, "Impartial Evaluation under Ambiguity", *Ethics*, 132(3): 541–569. doi:10.1086/718081

Broome, John, 1991, *Weighing Goods: Equality, Uncertainty, and Time* (Economics and Philosophy), Cambridge, MA: Basil Blackwell.

doi:10.1002/9781119451266

——, 2004, *Weighing Lives*, Oxford/New York: Oxford University Press. doi:10.1093/019924376X.001.0001

Brown, Campbell, 2011, "Consequentialize This", *Ethics*, 121(4): 749–771. doi:10.1086/660696

Buchak, Lara Marie, 2013, *Risk and Rationality*, Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199672165.001.0001

——, 2017, "Taking Risks behind the Veil of Ignorance", *Ethics*, 127(3): 610–644. doi:10.1086/690070

——, 2023, "How Should Risk and Ambiguity Affect Our Charitable Giving?", *Utilitas*, 35(3): 175–197. doi:10.1017/S0953820823000055

Burch-Brown, Joanna M., 2014, "Clues for Consequentialists", *Utilitas*, 26(1): 105–119. doi:10.1017/S0953820813000289

Bykvist, Krister, 2020, "Consequentialism, Ignorance, and Uncertainty", in *The Oxford Handbook of Consequentialism*, Douglas W. Portmore (ed.), New York: Oxford University Press, 310–330. doi:10.1093/oxfordhb/9780190905323.013.8

Carlson, Erik, 1995, *Consequentialism Reconsidered* (Theory and Decision Library - Series A: Philosophy and Methodology of the Social Sciences, 20), Dordrecht/Boston: Kluwer Academic Publishers. doi:10.1007/978-94-015-8553-8

Carr, Jennifer Rose, 2020, "Normative Uncertainty without Theories", *Australasian Journal of Philosophy*, 98(4): 747–762. doi:10.1080/00048402.2019.1697710

Colyvan, Mark, Damian Cox, and Katie Steele, 2010, "Modelling the Moral Dimension of Decisions: Modelling the Moral Dimension of Decisions", *Noûs*, 44(3): 503–529. doi:10.1111/j.1468-0068.2010.00754.x

Cotton-Barratt, Owen, William MacAskill, and Toby Ord, 2020, "Statistical Normalization Methods in Interpersonal and

Intertheoretic Comparisons", *The Journal of Philosophy*, 117(2): 61–95. doi:10.5840/jphil202011725

Diamond, Peter A., 1967, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment", *Journal of Political Economy*, 75(5): 765–766. doi:10.1086/259353

Driver, Julia, 2012, *Consequentialism* (New Problems of Philosophy), London/New York: Routledge. doi:10.4324/9780203149256

Feldman, Fred, 2006, "Actual Utility, The Objection from Impracticality, and the Move to Expected Utility", *Philosophical Studies*, 129(1): 49–79. doi:10.1007/s11098-005-3021-y

Finnis, John, 2011, *Natural Law and Natural Rights*, second edition, (Clarendon Law Series), Oxford/New York: Oxford University Press.

Fishburn, Peter C., 1971, "A Study of Lexicographic Expected Utility", *Management Science*, 17(11): 672–678. doi:10.1287/mnsc.17.11.672

Foot, Philippa, 1967, "The Problem of Abortion and the Doctrine of the Double Effect", *Oxford Review*, 5: 5–15.

Frick, Johann, 2015, "Contractualism and Social Risk", *Philosophy & Public Affairs*, 43(3): 175–223. doi:10.1111/papa.12058

Fried, Barbara H., 2012, "Can Contractualism Save Us from Aggregation?", *The Journal of Ethics*, 16(1): 39–66. doi:10.1007/s10892-011-9113-3

Gewirth, Alan, 1981, "Are There Any Absolute Rights?", *The Philosophical Quarterly*, 31(122): 1–16. doi:10.2307/2218674

——, 1982, "There Are Absolute Rights", *The Philosophical Quarterly*, 32(129): 348–353. doi:10.2307/2218701

Gigerenzer, Gerd and Wolfgang Gaissmaier, 2011, "Heuristic Decision Making", *Annual Review of Psychology*, 62(1): 451–482. doi:10.1146/annurev-psych-120709-145346

Gracely, Edward J., 1996, "On the Noncomparability of Judgments Made by Different Ethical Theories", *Metaphilosophy*, 27(3): 327–332. doi:10.1111/j.1467-9973.1996.tb00212.x

Graham, Peter A., 2010, "In Defense of Objectivism about Moral Obligation", *Ethics*, 121(1): 88–115. doi:10.1086/656328

Greaves, Hilary, 2016, "Cluelessness", *Proceedings of the Aristotelian Society*, 116(3): 311–339. doi:10.1093/arisoc/aow018

——, 2017, "A Reconsideration of the Harsanyi–Sen–Weymark Debate on Utilitarianism", *Utilitas*, 29(2): 175–213. doi:10.1017/S0953820816000169

Greenwell, James R., 1977, "Abortion and Moral Safety", *Crítica: Revista Hispanoamericana de Filosofía*, 9(27): 35–48.

Gustafsson, Johan E., 2022, "Second Thoughts About My Favourite Theory", *Pacific Philosophical Quarterly*, 103(3): 448–470. doi:10.1111/papq.12408

Gustafsson, Johan E. and Olle Torpman, 2014, "In Defence of My Favourite Theory", *Pacific Philosophical Quarterly*, 95(2): 159–174. doi:10.1111/papq.12022

Hammond, Peter J., 1998, "Objective Expected Utility: A Consequentialist Perspective", in *Handbook of Utility Theory*, Salvador Barbera, Peter J. Hammond, and Christian Seidl (eds.), Dordrecht/Boston: Kluwer Academic Publishers, volume 1, 142–211 (ch. 5).

Hansson, Sven Ove, 2003, "Ethical Criteria of Risk Acceptance", *Erkenntnis*, 59(3): 291–309. doi:10.1023/A:1026005915919

Hare, Caspar, 2010, "Take the Sugar", *Analysis*, 70(2): 237–247. doi:10.1093/analys/anp174

Harman, Elizabeth, 2015, "The Irrelevance of Moral Uncertainty", in *Oxford Studies in Metaethics, Volume 10*, Russ Shafer-Landau (ed.), Oxford: Oxford University Press, 53–79 (ch. 3). doi:10.1093/acprof:oso/9780198738695.003.0003

Harsanyi, John C., 1955, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility", *Journal of Political Economy*, 63(4): 309–321. doi:10.1086/257678

——, 1977, "Morality and the Theory of Rational Behavior", *Social Research*, 44(4): 623–656.

Hausner, M., 1954, "Multidimensional Utilities", in *Decision Processes*, Robert McDowell Thrall, C. H. Coombs, and R. L. Davis (eds), New York: Wiley, 167–180.

Hedden, Brian, 2016, "Does MITE Make Right? On Decision-Making under Normative Uncertainty", in *Oxford Studies in Metaethics, Volume 11*, Russ Shafer-Landau (ed.), Oxford: Oxford University Press, 102–128 (ch. 5). doi:10.1093/acprof:oso/9780198784647.003.0005

Helgeson, Casey, 2020, "Structuring Decisions Under Deep Uncertainty", *Topoi*, 39(2): 257–269. doi:10.1007/s11245-018-9584-y

Howard-Snyder, Frances, 1997, "The Rejection of Objective Consequentialism", *Utilitas*, 9(2): 241–248. doi:10.1017/S0953820800005306

Hudson, James L., 1989, "Subjectivization in Ethics", *American Philosophical Quarterly*, 26(3): 221–229.

Huemer, Michael, 2010, "Lexical Priority and the Problem of Risk", *Pacific Philosophical Quarterly*, 91(3): 332–351. doi:10.1111/j.1468-0114.2010.01370.x

Isaacs, Yoaav, 2014, "Duty and Knowledge", *Philosophical Perspectives*, 28(1): 95–110. doi:10.1111/phpe.12042

Jackson, Frank, 1991, "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection", *Ethics*, 101(3): 461–482. doi:10.1086/293312

Jackson, Frank and Michael Smith, 2006, "Absolutist Moral Theories and Uncertainty":, *Journal of Philosophy*, 103(6): 267–283. doi:10.5840/jphil2006103614

——, 2016, "The Implementation Problem for Deontology", in *Weighing Reasons*, Errol Lord and Barry Maguire (eds.), New York: Oxford University Press, 279–292 (ch. 14). doi:10.1093/acprof:oso/9780199315192.003.0014

Kagan, Shelly, 1998, *Normative Ethics* (Dimensions of Philosophy Series), Boulder, CO: Westview Press.

Kant, Immanuel, 1797 [1996], "On a Supposed Right to Lie from Philanthropy (1797)", translated in *Practical Philosophy*, Mary J. Gregor (ed.), (The Cambridge Edition of the Works of Immanuel Kant), Cambridge: Cambridge University Press, 605–616.

Kolodny, Niko and John MacFarlane, 2010, "Ifs and Oughts":, *Journal of Philosophy*, 107(3): 115–143. doi:10.5840/jphil2010107310

Kosonen, Petra, 2022, "Tiny Probabilities of Vast Value", PhD Thesis, Worcester College, Oxford University. [Kosonen 2022 available online]

Kowalczyk, Kacper, forthcoming, "Saving Fanaticism", *Australasian Journal of Philosophy*.

Kumar, Rahul, 2015, "Risking and Wronging", *Philosophy & Public Affairs*, 43(1): 27–51. doi:10.1111/papa.12042

Lazar, Seth, 2017a, "Anton's Game: Deontological Decision Theory for an Iterated Decision Problem", *Utilitas*, 29(1): 88–109. doi:10.1017/S0953820816000236

——, 2017b, "Deontological Decision Theory and Agent-Centered Options", *Ethics*, 127(3): 579–609. doi:10.1086/690069

Lazar, Seth and Chad Lee Stronach, 2019, "Axiological Absolutism and Risk", *Noûs*, 53(1): 97–113. doi:10.1111/nous.12210

Lee-Stronach, Chad, 2018, "Moral Priorities under Risk", *Canadian Journal of Philosophy*, 48(6): 793–811. doi:10.1080/00455091.2017.1415104

Lenman, James, 2000, "Consequentialism and Cluelessness", *Philosophy & Public Affairs*, 29(4): 342–370. doi:10.1111/j.1088-4963.2000.00342.x

——, 2008, "Contractualism and Risk Imposition", *Politics, Philosophy & Economics*, 7(1): 99–122. doi:10.1177/1470594X07085153

Lockhart, Ted, 2000, *Moral Uncertainty and Its Consequences*, New York: Oxford University Press. doi:10.1093/oso/9780195126105.001.0001

Lundgren, Björn and H. Orri Stefánsson, 2020, "Against the De Minimis Principle", *Risk Analysis*, 40(5): 908–914. doi:10.1111/risa.13445

MacAskill, William, 2016, "Normative Uncertainty as a Voting Problem", *Mind*, 125(500): 967–1004. doi:10.1093/mind/fzv169

MacAskill, William, Krister Bykvist, and Toby Ord, 2020, *Moral Uncertainty*, Oxford: Oxford University Press. doi:10.1093/oso/9780198722274.001.0001

MacAskill, William and Toby Ord, 2020, "Why Maximize Expected Choice Worthiness?", *Noûs*, 54(2): 327–353. doi:10.1111/nous.12264

Marchau, Vincent A. W. J., Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper (eds), 2019, *Decision Making under Deep Uncertainty: From Theory to Practice*, Cham: Springer International Publishing. doi:10.1007/978-3-030-05252-2

Mason, Elinor, 2013, "Objectivism and Prospectivism About Rightness", *Journal of Ethics and Social Philosophy*, 7(2): 1–22. doi:10.26556/jesp.v7i2.72

McCarthy, David, 2015, "Distributive Equality", *Mind*, 124(496): 1045–1109. doi:10.1093/mind/fzv028

——, 2017, "The Priority View", *Economics and Philosophy*, 33(2): 215–257. doi:10.1017/S0266267116000225

McCarthy, David, Kalle Mikkola, and Teruji Thomas, 2020, "Utilitarianism with and without Expected Utility", *Journal of Mathematical Economics*, 87: 77–113. doi:10.1016/j.jmateco.2020.01.001

Mogensen, Andreas L, 2021, "Maximal Cluelessness", *The Philosophical Quarterly*, 71(1): 141–162. doi:10.1093/pq/pqaa021

Mogensen, Andreas L. and David Thorstad, 2022, "Tough Enough? Robust Satisficing as a Decision Norm for Long-Term Policy Analysis", *Synthese*, 200: article 36. doi:10.1007/s11229-022-03566-5

Monton, Bradley, 2019, "How to Avoid Maximizing Expected Utility", *Philosopher's Imprint*, 19: article 18. [Monton 2019 available online]

Myerson, Roger B., 1981, "Utilitarianism, Egalitarianism, and the Timing Effect in Social Choice Problems", *Econometrica*, 49(4): 883–897. doi:10.2307/1912508

Nissan-Rozen, Ittay, 2012, "Doing the Best One Can: A New Justification for the Use of Lotteries", *Erasmus Journal for Philosophy and Economics*, 5(1): 45–72. doi:10.23941/ejpe.v5i1.93

Nozick, Robert, 1974, *Anarchy, State, and Utopia*, New York: Basic Books.

Oddie, Graham and Peter Menzies, 1992, "An Objectivist's Guide to Subjective Value", *Ethics*, 102(3): 512–533. doi:10.1086/293422

Otsuka, Michael and Alex Voorhoeve, 2009, "Why It Matters That Some Are Worse Off Than Others: An Argument against the Priority View", *Philosophy & Public Affairs*, 37(2): 171–199. doi:10.1111/j.1088-4963.2009.01154.x

Parfit, Derek, 1984, *Reasons and Persons*, Oxford/New York: Clarendon Press. doi:10.1093/019824908X.001.0001

——, 2011, *On What Matters. Volume One*, Samuel Scheffler (ed.), (The Berkeley Tanner Lectures), Oxford ; New York: Oxford University Press. doi:10.1093/acprof:osobl/9780199572809.001.0001

Pfeiffer, Raymond S., 1985, "Abortion Policy and the Argument from Uncertainty":, *Social Theory and Practice*, 11(3): 371–386. doi:10.5840/soctheorpract198511315

Podgorski, Abelard, 2020, "Normative Uncertainty and the Dependence Problem", *Mind*, 129(513): 43–70. doi:10.1093/mind/fzz048

Portmore, Douglas W., 2017, "Uncertainty, Indeterminacy, and Agent-Centred Constraints", *Australasian Journal of Philosophy*, 95(2): 284–298. doi:10.1080/00048402.2016.1219376

Quiggin, John, 1982, "A Theory of Anticipated Utility", *Journal of Economic Behavior & Organization*, 3(4): 323–343. doi:10.1016/0167-2681(82)90008-7

Railton, Peter, 1984, "Alienation, Consequentialism, and the Demands of Morality", *Philosophy and Public Affairs*, 13(2): 134–171.

Regan, Donald, 1980, *Utilitarianism and Co-Operation*, Oxford: Clarendon Press. doi:10.1093/acprof:oso/9780198246091.001.0001

Rinard, Susanna, 2015, "A Decision Theory for Imprecise Probabilities", *Philosopher's Imprint*, 15: article 7. [Rinard 2015 available online]

Ross, Jacob, 2006, "Rejecting Ethical Deflationism", *Ethics*, 116(4): 742–768. doi:10.1086/505234

Russell, Jeffrey Sanford, forthcoming, "On Two Arguments for Fanaticism", *Noûs*, first online: 2 June 2023. doi:10.1111/nous.12461

Russell, Jeffrey Sanford and Yoaav Isaacs, 2021, "Infinite Prospects", *Philosophy and Phenomenological Research*, 103(1): 178–198. doi:10.1111/phpr.12704

Savage, Leonard J., 1954, *The Foundations of Statistics*, (Wiley Publications in Statistics), New York: Wiley.

Schoenfield, Miriam, 2014, "Decision Making in the Face of Parity", *Philosophical Perspectives*, 28(1): 263–277. doi:10.1111/phpe.12044

Sen, Amartya K., 1971, "Choice Functions and Revealed Preference", *The Review of Economic Studies*, 38(3): 307–317. doi:10.2307/2296384

Sepielli, Andrew Christopher, 2009, "What to Do When You Don't Know What to Do", in *Oxford Studies in Metaethics, Volume 4*, Russ Shafer-Landau (ed.), Oxford: Oxford University Press, 5–28 (ch. 1). doi:10.1093/oso/9780199566303.003.0002

——, 2010, *'Along an Imperfectly-Lighted Path': Practical Rationality and Normative Uncertainty*, PhD, New Brunswick, NJ: Rutgers, The State University of New Jersey. doi:10.7282/T3B56JWG

——, 2013, "Moral Uncertainty and the Principle of Equity among Moral Theories", *Philosophy and Phenomenological Research*, 86(3): 580–589. doi:10.1111/j.1933-1592.2011.00554.x

——, 2014, "Should You Look before You Leap?":, *The Philosophers' Magazine*, 66: 89–93. doi:10.5840/tpm20146690

——, 2016, "Moral Uncertainty and Fetishistic Motivation", *Philosophical Studies*, 173(11): 2951–2968. doi:10.1007/s11098-016-0645-z

——, 2018a, "How Moral Uncertaintism Can Be Both True and Interesting", in *Oxford Studies in Normative Ethics, Volume 7*, Mark C. Timmons (ed.), Oxford: Oxford University Press, 98–116 (ch. 5).

——, 2018b, "Subjective and Objective Reasons", in *The Oxford Handbook of Reasons and Normativity*, Daniel Star (ed.), Oxford: Oxford University Press, 784–799 (ch. 33).

Smart, J. J. C., 1973, "An Outline of a System of Utilitarian Ethics", in *Utilitarianism: For and Against*, by Bernard Williams and J. J. C. Smart, Cambridge: Cambridge University Press, 1–74.

Smith, Holly M., 2010, "Subjective Rightness", *Social Philosophy and Policy*, 27(2): 64–110. doi:10.1017/S0265052509990161

——, 2018, *Making Morality Work*, Oxford: Oxford University Press. doi:10.1093/oso/9780199560080.001.0001

Smith, Nicholas J. J., 2014, "Is Evaluative Compositionality a Requirement of Rationality?", *Mind*, 123(490): 457–502. doi:10.1093/mind/fzu072

——, 2016, "Infinite Decisions and Rationally Negligible Probabilities", *Mind*, 125(500): 1199–1212. doi:10.1093/mind/fzv209

Steele, Katie and H. Orri Stefánsson, 2021, *Beyond Uncertainty: Reasoning with Unknown Possibilities* (Cambridge Elements. Elements in Decision Theory and Philosophy), Cambridge/New York: Cambridge University Press. doi:10.1017/9781108582230

——, 2022, "Transformative Experience, Awareness Growth, and the Limits of Rational Planning", *Philosophy of Science*, 89(5): 939–948. doi:10.1017/psa.2022.55

Stefánsson, H. Orri and Richard Bradley, 2015, "How Valuable Are Chances?", *Philosophy of Science*, 82(4): 602–625. doi:10.1086/682915

Tarsney, Christian, 2018, "Intertheoretic Value Comparison: A Modest Proposal", *Journal of Moral Philosophy*, 15(3): 324–344. doi:10.1163/17455243-20170013

——, 2019, "Normative Uncertainty and Social Choice", *Mind*, 128(512): 1285–1308. doi:10.1093/mind/fzy051

——, 2021, "Vive La Différence? Structural Diversity as a Challenge for Metanormative Theories", *Ethics*, 131(2): 151–182. doi:10.1086/711204

Tenenbaum, Sergio, 2017, "Action, Deontology, and Risk: Against the Multiplicative Model", *Ethics*, 127(3): 674–707. doi:10.1086/690072

Thoma, Johanna, 2022, "Time for Caution", *Philosophy & Public Affairs*, 50(1): 50–89. doi:10.1111/papa.12204

——, 2023, "Taking Risks on Behalf of Another", *Philosophy Compass*, 18(3): e12898. doi:10.1111/phc3.12898

Thomas, Teruji, 2023, "The Asymmetry, Uncertainty, and the Long Term", *Philosophy and Phenomenological Research*, 107(2): 470–500. doi:10.1111/phpr.12927

Thomson, Judith Jarvis, 1976, "Killing, Letting Die, and the Trolley Problem":, *Monist*, 59(2): 204–217. doi:10.5840/monist197659224

Thorstad, David and Andreas Mogensen, 2020, "Heuristics for Clueless Agents: How to Get Away with Ignoring What Matters Most in Ordinary Decision-Making". *GPI Working Paper 2–2020*, University of Oxford: Global Priorities Institute. [Thorstad and Mogensen 2020 available online]

Trammell, Philip, 2021, "Fixed-Point Solutions to the Regress Problem in Normative Uncertainty", *Synthese*, 198(2): 1177–1199. doi:10.1007/s11229-019-02098-9

Weatherson, Brian, 2014, "Running Risks Morally", *Philosophical Studies*, 167(1): 141–163. doi:10.1007/s11098-013-0227-2

——, 2019, *Normative Externalism*, Oxford: Oxford University Press. doi:10.1093/oso/9780199696536.001.0001

Wilkinson, Hayden, 2022, "In Defense of Fanaticism", *Ethics*, 132(2): 445–477. doi:10.1086/716869

Williamson, Timothy, 2000, *Knowledge and Its Limits*, Oxford/New York: Oxford University Press. doi:10.1093/019925656X.001.0001

Yim, Lok Lam, 2019, "The Cluelessness Objection Revisited", *Proceedings of the Aristotelian Society*, 119(3): 321–324. doi:10.1093/arisoc/aoz016

Zimmerman, Michael J., 2008, *Living with Uncertainty: The Moral Significance of Ignorance* (Cambridge Studies in Philosophy), Cambridge/New York: Cambridge University Press. doi:10.1017/CBO9780511481505

Academic Tools

- How to cite this entry.
- Preview the PDF version of this entry at the Friends of the SEP Society.
- Look up topics and thinkers related to this entry at the Internet Philosophy Ontology Project (InPhO).
- Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

- Roussos, Joe, 2021, "Unawareness for Longtermists", 7th Oxford Workshop on Global Priorities Research, 24 June 2021. [Roussos 2021 slides available online (pdf)]

Related Entries

consequentialism | consequentializing | contractualism | decision theory | epistemology: Bayesian | ethics: deontological | Pascal's wager | probabilities, imprecise | rational choice, normative: expected utility | rational choice, normative: rivals to expected utility | risk | St. Petersburg paradox

Acknowledgments

For helpful feedback on a draft on this entry, the authors are grateful to Krister Bykvist, Chad Lee-Stronach, Andreas Mogensen, Katie Steele, David Thorstad, and Hayden Wilkinson.

Notes to Moral Decision-Making Under Uncertainty

1. Though often referred to as "Jackson cases", cases with the relevant structure are also given by Regan (1980: 265) and Parfit (2011: 159).

2. Though many philosophers have found this idea appealing, the concept of a moral principle being "usable" or "action-guiding" is notoriously difficult to explicate. For an important recent discussion, see Holly M. Smith (2018).

3. More precisely, some philosophers hold that the *normative* "ought", *practical* "ought", or *moral* "ought" is univocal. The word "ought" may

have entirely different meanings—for instance, the "ought" of natural expectation in a sentence like "The Moon ought to rise in the next ten minutes"—that are irrelevant for our purposes.

4. Although it is natural to frame these debates in terms of senses or meanings of English words like "ought", the important question for ethical purposes is not about the meanings of these words in ordinary language. Rather, the question is about the existence and importance of fact-relative, belief-relative, and evidence-relative normative standards or properties, for which a philosopher might *stipulatively* use terms like "ought" or "rightness" (without too much risk of misleading, insofar as the stipulated meanings depart from ordinary meanings).

5. For defenses of the "objectivist" view that privileges fact-relative moral concepts, see Carlson (1995), Graham (2010), and Driver (2012). For the "subjectivist" view that privileges belief-relative moral concepts, see Hudson (1989). For the "prospectivist" view that privileges the evidence-relative moral concepts, see Zimmerman (2008) and Mason (2013). Jackson (1991) and Howard-Snyder (1997) both reject objectivism while remaining non-committal between subjectivism and prospectivism. For "divider" views that distinguish objective and non-objective moral concepts without privileging one over the other, see Oddie & Menzies (1992; though their view is borderline "objectivist"), Holly M. Smith (2010), and Parfit (2011). Finally, see Kolodny & MacFarlane (2010) for an influential argument against the use of Jackson-style cases to motivate the "divider" position. For further citations to the very extensive literature on these questions, see fns. 1–3 of Mason (2013) and fn. 2 of Sepielli (2018b).

6. The use of the word "utility" here should be carefully distinguished from other uses in ethics; in particular, it has nothing directly to do with the "utility" in "utilitarianism". Further interpretive comments will follow.

7. For an introduction to these theorems, see section 2.2 of the entry on expected utility theory.

8. The forms of contraction and expansion consistency stated here are sometimes called "Property α" and "Property β", in reference to Sen (1971). As noted above, expected utility theory is usually presented as a theory about *preferences*, and thus deals with a binary relation $R(A, B)$ interpreted as "A is weakly preferred to B". In the context of permissibility, we can reinterpret $R(A, B)$ as "it is sometimes permissible to choose A when B is available". Given suitable background assumptions, the force of Properties α and β is that this R is a complete, transitive ordering, and that an option is permissible if and only if it is maximal with respect to R (i.e., A is permissible if and only if $R(A, B)$ for every other available option B) (Sen 1971: 8). Of course, other possible interpretations of $R(A, B)$ are salient in a moral context, such as "A is at least as good as B".

9. The sure thing principle originates in Savage (1954); our informal version matches his motivating discussion more closely than his formal statement.

10. See, e.g., the risk-weighted expected utility theory of Buchak (2013), inspired by Quiggin (1982).

11. See, e.g., the lexicographic expected utility theory of Hausner (1954) and Fishburn (1971).

12. A large literature starts from Aumann (1962); in the recent philosophical literature, Hare (2010) has been much discussed, particularly with respect to violations of stochasticism—see for instance Schoenfield (2014); Bales, Cohen, & Handfield (2014); Bader (2018).

13. In terms of our example, however, it is worth noting that the best-defended forms of utilitarianism, like those mentioned in the next paragraph, involve risk-neutrality.

14. Sometimes a distinction is drawn between "risk" and "uncertainty". The distinction is not always completely clear, but, roughly speaking, the former term covers cases where the decision-relevant probabilities are precise and accessible, and the latter term covers harder cases. In those terms, cluelessness about consequences presumably involves "uncertainty" rather than "risk".

15. To be a bit more careful: to tell which of two options has higher expected value, we do not necessarily need to know the expected value of each option separately. We just need to know whether the *difference* in expected value is positive or negative. (By analogy, if you see two people at a distance, you might know which one is taller while being very uncertain about their individual heights.) But the answer to this latter question can still depend sensitively on various probabilities and on the differences between them. To return to an earlier example, using my left hand rather than my right makes *roughly* zero difference to the probability of an extra typhoon; but, since the typhoon would be extremely destructive, even a small non-zero difference in the probability could make a decisive difference in expected value.

16. Constraints can also take a positive form; e.g., keep your promises even if that will result if fewer total instances of promise-keeping.

17. Relatedly, this view violates the continuity axiom of section section 2, which is used in the standard version of expected utility theory to get utilities that are numbers rather than vectors.

18. This reasoning may be too quick, since, as noted in section 2, the notion of an "outcome" in expected utility theory is quite flexible. For

example, Stefánsson & Bradley (2015) apply expected utility theory using outcomes that include facts about objective chances; they can therefore distinguish between an outcome in which *B* is not harmed, and there was no chance of *B*'s being harmed, from an outcome in which *B* is not harmed despite facing a significant objective risk. They argue that it could be rational to prefer the former option. Note, however, that the deontological rationales described in the main text also seem to cover subjective or evidential probabilities of harm.

19. This literature also tends to frame things less in terms of risks of violating rights or constraints, and more in terms of risks of harm. However, in rejecting the consequentialist position that a risky activity should be permitted as long as the benefits outweigh the harms, most participants in this literature arguably take there to be something like a (non-absolute) constraint against causing certain kinds of harm.

20. The best-known example of such a prospect is the St. Petersburg game, in which a fair coin is flipped repeatedly until it hands heads, with the player then receiving a payoff of $\$2^n$, where n is the total number of flips. The expected monetary reward from playing this game is

$$\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + \frac{1}{8} \times 8 \ldots = 1 + 1 + 1 \ldots = \infty.$$

So, it seems, a gambler who maximizes their expected monetary reward should be willing to pay *any* finite amount to play the St. Petersburg game, meaning that they strictly prefer the game to *all* of its possible outcomes (each of which is a finite amount of money). See the entry on the St. Petersburg paradox for more on this and related puzzles in infinite decision theory.

Copyright © 2024 by the authors
Christian Tarsney, Teruji Thomas, and William MacAskill