Intertheoretic Value Comparison: A Modest Proposal

Christian Tarsney

(Forthcoming in *Journal of Moral Philosophy*. Please cite the published version.)

Abstract

In the growing literature on decision-making under moral uncertainty, a number of skeptics have argued that there is an insuperable barrier to rational "hedging" for the risk of moral error, namely the apparent incomparability of moral reasons given by rival theories like Kantianism and utilitarianism. Various general theories of intertheoretic value comparison have been proposed to meet this objection, but each suffers from apparently fatal flaws. In this paper, I propose a more modest approach that aims to identify classes of moral theories that share common principles strong enough to establish bases for intertheoretic comparison. I show that, contra the claims of skeptics, there are often rationally perspicuous grounds for precise, quantitative value comparisons within such classes. In light of this fact, I argue, the existence of *some* apparent incomparabilities between widely divergent moral theories cannot serve as a general argument against hedging for one's moral uncertainties.

1 Introduction

Moral philosophers have recently begun paying systematic attention to the question of how agents rationally ought to act when they divide their beliefs among competing moral principles or theories that offer conflicting prescriptions with respect to some moral dilemma. A popular and intuitively appealing view is that agents in some circumstances should "hedge" for their moral uncertainties—that is, should be responsive both to the likelihood that a given moral theory is correct and to the strength of the reasons it posits for or against a given practical option. Such moral hedging can take many forms, but most straightforwardly, it might be expressed by a principle that directs agents to maximize "expected moral rightness," in a way formally analogous to the dictates of standard expected utility theory for agents who are uncertain of the consequences of their actions. This sort of hedging is endorsed by Lockhart (2000), Sepielli (2009, 2010), and MacAskill (2014), among others.

The idea of moral hedging is appealing both because of its potential strength—it promises to deliver a systematic, determinable resolution to all questions of rational choice under moral uncertainty—and because of the natural continuity it suggests between the requirements of rationality that govern choice under moral and empirical uncertainty. Hedging also strikes us as intuitively correct in many cases: for instance, if I judge that meat-eating is most likely permissible but may well be a grave moral wrong, I seem to have at least *some* subjective reason to avoid eating meat, that varies both with my credence that it is wrong to do so and with the degree of potential wrongness that my beliefs attribute. And it seems that such reasons can at least sometimes outweigh weak prudential reasons in favor of eating meat, as the hedging thesis implies.

But moral hedging also faces several significant obstacles and objections. Among these, the most central conceptually and the problem that has so far proven least tractable is the *problem of intertheoretic value comparisons* (PIVC): the problem of finding some non-arbitrary basis for comparing degrees of moral value and disvalue attributed to practical options by competing moral theories.[1] For instance, consider an agent who must decide whether to kill one to save five, and divides her beliefs evenly between classical utilitarianism, which directs her to kill the one, and an absolutist deontological theory, which directs her to let the five die. The possibility of resolving this dilemma by moral hedging seems to require that there be some (quantitative) answer to the question of how the moral value of saving a life, according to classical utilitarianism, compares with the wrongness of killing an innocent person, according to absolutist deontology. It is hard to imagine what could make any answer to this question—or at least, any intuitively plausible answer—correct, let alone how we could discover that answer. Suppose that an agent who divides her moral beliefs as just described ought to kill the one iff the number of lives she would thereby save is greater than or equal to seven. On what basis could we ever hope to establish that this was in fact the proper threshold, rather than six, or twelve, or anything else?

A number of philosophers offered solutions to PIVC, but these extant proposals are subject to compelling objections, and none has achieved widespread acceptance. The first such proposal, Ted Lockhart's "Principle of Equity among Moral Theories" (PEMT) holds that, in any choice situation, the maximum and minimum degrees of value assigned to any option by each moral theory in which the agent has positive credence should be treated as equal (Lockhart, 2000, p. 84). But this view has several fatal drawbacks, described in detail by Sepielli (2013): For instance, because PEMT generates inconsistent comparisons between the same pair of theories across different choice situations, it can require an agent to knowingly choose a course of action that is strictly dominated, i.e. worse than some available alternative according to every moral theory in which she has positive credence. Sepielli in turn proposes, without endorsing, a variant of Lockhart's principle that he calls the "Conceivability PEMT," on which the maximum and minimum *conceivable* degrees of value according to any pair of theories should be treated as equal. But, as he notes, this principle does not permit comparisons between theories, like most forms of consequentialism, according to which there are no maximum or minimum degrees of value.

In his (2009), Sepielli endorsed a very different approach according to which we may normalize a pair of theories with one another by finding some set of options such that the two theories agree on the *ratios of value differences* among those options, and infer from this that the theories regard the value differences themselves among these options as equal in magnitude, thereby establishing a "background ranking" of options that the two theories have in common. But as MacAskill (2014) points out, there may be several such sets of options for

[1] The identification of the problem under this name, so far as I can tell, is due to Ross (2006). But it is raised under different names by earlier sources, e.g. Hudson (1989), Gracely (1996), and Lockhart (2000).

a given pair of theories that generate inconsistent comparisons between the theories (and in subsequent work Sepielli abandons the proposal for this reason (Sepielli, 2010, pp. 180-181)).

Ross (2006) suggests that we can work backwards from our considered judgments that a particular response to a given dilemma would be most rational given a particular state of moral uncertainty to the intertheoretic value comparisons that support these judgments. But apart from worries about the reliability of these judgments of rationality, Ross's approach leaves us without any satisfying explanation for the comparisons we infer, without any understanding of *why* so many units of value according to T_1 are equivalent to the wrongness of such-and-such action according to T_2. And as critics of moral hedging (e.g. Hudson (1989), Nissan-Rozen (2015)) have doubted whether any such explanation is possible, this approach at best leaves a large part of the problem unresolved.

These difficulties led opponents of moral hedging to suggest that PIVC is simply insoluble, and that this constitutes a decisive objection to hedging.[2] The aim of the present paper is to answer this objection. My strategy is simple: show that there are at least some cases in which moral hedging is possible, and then argue that if such cases exist, the problem of intertheoretic value comparisons cannot constitute an argument against moral hedging in general—even if there turn out to be other cases in which intertheoretic value comparisons are genuinely impossible.

The approach to intertheoretic comparison that I will propose differs from those of Lockhart and Sepielli in that it normalizes the value scales of rival theories based on the shared content of those theories—for instance, their shared endorsement of certain kinds of goods as bearers of non-derivative moral value—rather than structural features of each theory (like the range of its value assignments, as in Lockhart's PEMT) or structural similarities between theories (like matching value difference ratios, as in Sepielli's background rankings approach). It differs from Ross's proposal in that it offers an explanation for why certain intertheoretic value comparisons hold, an explanation grounded in the common content of the theories being compared. But my approach also differs from these earlier proposals in its relative lack of ambition: I aim, in this paper, not to give a general account of intertheoretic value comparisons, but only to establish that such comparisons are sometimes possible (and in the process, to establish one sort of sufficient condition for intertheoretic comparability).

I begin in the next section by describing a simple case of moral uncertainty in which there seems to be a rationally perspicuous, non-arbitrary standard for intertheoretic value comparisons. This case suggests the more general idea that there are clusters of moral theories (what I will call "comparability classes") that share common principles strong enough to establish procedures for moral hedging, which sometimes will and sometimes will not take the form of expected value maximization. In §3, I argue that this approach generalizes beyond the simple sort of case described in §2, while conceding that its scope may still turn out to be relatively limited. In §4, I argue that so long as a compelling basis for intertheoretic value comparisons exists in some cases, PIVC cannot be taken as an objection to moral hedging as such—even if hedging is sometimes

[2] Among those who have argued against moral hedging are Hudson (1989), Gracely (1996), Gustafsson and Torpman (2014), Weatherson (2014), Harman (2015), Nissan-Rozen (2015), and Hedden (2016). Of these, Gracely, Gustafsson and Torpman, Nissan-Rozen, and Hedden lay particular stress on PIVC as an objection to hedging.

impossible, this is no reason to forgo it when it is. Instead, I will argue, if there is a line to be drawn between uncertainties for which a hedging procedure is appropriate and those for which it is not, that line should be drawn not between moral and empirical uncertainty but rather between those uncertainties that establish barriers to rational comparison and those that do not. §5 concludes by examining the implications of the comparability class approach for two other worries about moral hedging.

2 A case study for PIVC: Pluralistic consequentialism

We begin with a simple case. Consider an agent Clara whose moral beliefs are as follows: (i) She is certain that consequentialism is true. (ii) She is sure that pleasure and pain have basic (non-derivative) moral value and disvalue, respectively. But (iii) she is uncertain whether other things, like friendship, beauty/aesthetic appreciation, knowledge, and so forth have basic moral value as well. Thus, she divides her beliefs between monistic, hedonistic consequentialism and one or more forms of *value-pluralistic* consequentialism. Internal to any of the pluralistic theories she finds plausible, there is some basis for comparison (an "exchange rate") between hedons and the various other basic forms of value posited by that theory—none of the theories to which she assigns positive credence, that is, treat the various basic bearers of moral value as incomparable—but insofar as she is uncertain whether the monistic or which of the various pluralistic theories is correct, she is uncertain of the correct exchange rate between these various forms of value.[3]

It seems to me that, in Clara's particular state of moral uncertainty, PIVC has no great force, and that when her various moral theories come into conflict, she should experience no great difficulty in hedging for her uncertainties. To illustrate, start with a simplest version of the case: Suppose that Clara divides her beliefs between just two theories, the monistic hedonistic theory T_1 and a theory T_2 according to which there are two kinds of non-derivative goods, hedonic and aesthetic. We may stipulate an arbitrary unit of hedonic value, the hedon, and may then define a unit of aesthetic value, the aestheton, as the quantity of aesthetic value that is equal in value to a hedon according to T_2. Now suppose Clara assigns probability p to T_1 and probability $(1-p)$ to T_2. In this state, I propose, she can quite naturally calculate the expected value of any practical option O as follows.

$$\mathrm{EV}(O) = \mathit{hedons}(O) + (1-p)(\mathit{aesthetons}(O))$$

[3] By an "exchange rate" between two kinds of value-bearer V_1 and V_2 I mean a principle that specifies when a given quantity of V_1 has greater, less, or equal value to a given quantity of V_2 (and, where ratio scale comparisons are possible, specifies the ratio between the values of these two quantities). A first-order moral theory, like pluralistic consequentialism, that treats two value-bearers like pleasure and beauty as fully comparable will thus specify an exchange rate between them. But if an agent divides her beliefs between several first-order theories that posit different exchange rates between two kinds of value, she may appeal to a second-order theory of decision-making under moral uncertainty that specifies a new exchange rate, perhaps a weighted average of the exchange rates given by the various first-order theories in which she has positive credence, as I will shortly illustrate.

Prima facie, it seems quite plausible that an agent in Clara's doxastic state rationally ought to choose the option that maximizes this formula for expected value. The two theories in which Clara has positive credence agree on the value of hedonic goods—take it as a stipulation, to which we will return shortly, that T_1 and T_2 differ not at all in the grounds they offer for, or anything else they have to say about, the intrinsic value of a hedon, i.e. that Clara suffers from no uncertainty regarding the value of hedonic goods—and as a result T_2 contains within itself, in the exchange rate it posits between hedonic and aesthetic goods, a basis for comparing its own evaluation of practical options to the exclusively hedonistic evaluations of T_1.

The suggestion from this simplest version of the case of Clara can be easily generalized. Let Clara distribute her beliefs over any number of pluralistic theories, along with her monistic theory, so long as all those theories agree on the value of hedonic goods. Then we can simply create a unit of value for each theory equal to the quantity of value that it assigns to a single hedon (an empirically measurable unit of pleasure/pain, whose value, we have by stipulation every reason to believe, remains constant across theories), thereby normalizing the value scales of the various competing theories. Then we can simply compute the expected value of an option O across theories $\{T_1, T_2, ..., T_n\}$ by something very much like the standard formula for expected value:

$$\mathrm{EV}(O) = \sum_{i=1}^{n} (\mathrm{EV}(O)|T_i)(p(T_i))$$

This formula will be applicable even to an agent who distributes her belief over infinitely many competing theories, corresponding for instance to all possible real-valued exchange rates between hedonic and other forms of value over some bounded interval. In fact, we can generalize even further, dropping the assumption that there is a category of value-bearer common to every theory in which Clara has positive credence. Suppose, for instance, that Clara regards hedonic experience, beauty, and knowledge all as potential non-derivative value-bearers, and that her credences regarding the value of each of these goods are mutually independent, so that she has some credence in theories according to which all, none, or any other combination of the three are non-derivative goods. Comparisons between these various goods can nevertheless be made by appealing to the exchange rates given by the various pluralistic theories that value more than one of the goods in question, so long as these exchange rates are consistent across the relevant pluralistic theories. Thus, for instance, the monistic theory according to which only pleasure is a non-derivative good and the monistic theory according to which only knowledge is a non-derivative good are made comparable by the agent's credence in one or more theories according to which both are non-derivative goods.

What I have said so far involves an important presupposition, namely that when two theories agree that some feature of the world is a non-derivative bearer of moral value, those two theories attribute the *same kind* and *same degree* of value to that phenomenon—e.g., that the value of a hedon according to Clara's hedonistic theory is equal to the value of a hedon according to the pluralistic theory that also values aesthetic goods. Clearly, this need not always be the case. For instance, Clara might divide her metaethical beliefs between a robust moral realism and a fairly anemic anti-realism, and it might turn out that her credence in hedonistic consequentialism is mostly or entirely conditioned on her credence in robust realism while her credence in rival pluralistic theories is mostly or entirely conditioned on her credence in anti-realism. (Suppose she inclines toward a hedonistic view on which certain qualia have intrinsic value or disvalue entirely independent of our beliefs, attitudes, etc, which we are morally required to maximize. But if this view turns out to be wrong, she believes, then morality can only consist in the pursuit of whatever we contingently happen to value in some distinctively moral way, which includes pleasure but also knowledge, aesthetic goods, friendship, etc.) In this case, the sort of procedure that I have described no longer seems rationally compelling: although the two views to which she assigns positive credence both treat pleasure as a bearer of non-derivative moral value, there is no obvious reason to think that they assign the same *degree* of value to a given unit of pleasure.

But there is no reason to assume that such problems will always arise. An agent may be uncertain about first-order moral questions, like whether anything besides pleasure and pain has non-derivative value, without this reflecting any underlying metaethical uncertainties. An agent who divides her beliefs between various monistic and pluralistic theories might nevertheless be in no doubt as to the nature, basis, or degree of value possessed by some category of goods, like hedonic goods, that all the theories she entertains recognize as non-derivatively valuable.[4] The lack of any uncertainty concerning hedonic value makes it a constant feature of the various theories in which she has positive credence, and allows it to serve as a basis for normalization.

To sum up, then, the suggestion is this: An agent who is certain that some form of (maximizing, agent-neutral) consequentialism is correct, but uncertain—even quite radically uncertain—about what sorts of things have value or disvalue nevertheless can find in the shared (or overlapping) content of the various first-order theories to which she assigns positive credence sufficient basis for rationally perspicuous, non-arbitrary intertheoretic value comparisons. Of course, nothing I have said establishes incontrovertibly that an agent who divides her beliefs among such theories is rationally required to maximize expected value, computed in the sort of way I have described. I have only tried to establish that, contra the assertions of pessimists about PIVC, there is a reasonable, non-arbitrary way of computing expected value in the face of certain intertheoretic moral uncertainties.

Is the case we have been considering an isolated exception in this respect? There is no use denying that it is an especially favorable case—a case of intertheoretic moral uncertainty that "behaves" helpfully like empirical uncertainty and seems to present only one salient option for normalizing the various theories in which the agent has positive credence. Nevertheless, it seems to me, there is an underlying idea that has the potential to generalize quite widely: namely, that clusters of moral theories united by the right common content—and particularly, by common procedures for addressing empirical uncertainties about morally salient features of the world—can make use of those shared assumptions

[4] Alternatively, if she does have metaethical uncertainties about the nature of value, these may be probabilistically independent of her first-order moral beliefs. Just as metaethical uncertainties need not pose any problem for an agent maximizing expected utility in the face of empirical uncertainties, so long as the metaethical and empirical uncertainties are probabilistically independent, so those metaethical uncertainties need pose no obstacle to hedging for first-order moral uncertainties, so long as the two are independent.

as a basis for intertheoretic values comparisons. Let's call such clusters of moral theories whose shared features permit quantitative comparisons of moral value between their members *comparability classes*. In the next section, we will briefly examine two more potential comparability classes of theories, and consider how far the comparability class strategy can generalize. Then we will turn to the implications of the existence of such comparability classes for the broader problem of decision-making under moral uncertainty.

3 The scope of the strategy

To illustrate the general idea (and more especially, to show that we were not in the last section simply taking advantage of an idiosyncratic feature of consequentialist moral theories), let's consider another case far afield from the first: namely, absolutist deontological theories that categorically prohibit certain kinds of actions like telling a lie, breaking a promise, or killing the innocent. No less than consequentialists, defenders of such theories must find something to say about how agents ought to act under morally relevant empirical uncertainty, e.g. uncertainty whether some act would violate a deontological constraint. If I am absolutely prohibited from breaking my promises, what should I do when I am unsure whether I made some particular commitment as part of a long-ago act of promising? If I am absolutely prohibited from killing the innocent, how am I to assess courses of action that carry some (perhaps moderate to vanishingly small) *risk* of killing the innocent, or when I have good reason to kill (in defense of myself or others) someone who might be either an innocent or a malevolent threat?[5]

Let's imagine, then, a deontologist David who has the following simple view of how to deal with uncertainty: an absolute moral requirement is incumbent on an agent whenever his degree of belief that the conditions of that requirement are satisfied is greater than .5. So, for instance, if it is in his power to perform some action that he might or might not have promised to perform, he is absolutely required to perform it so long as he believes it more likely than not that he did so promise; he is absolutely prohibited from uttering a particular sentence if he believes it more likely than not that his intention in doing so would be to deceive; and so on.

This sort of view is open to serious objections, but it seems to me as plausible as anything else deontological absolutists can say about the problem of uncertainty, so let us suppose that this is how some plausible class of deontological

[5] This last possibility is suggested by Jackson and Smith (2006). They describe a case in which a skier is headed down a mountain in a direction that will trigger an avalanche, killing ten people, unless you shoot and kill him. You are uncertain whether the skier is ignorant of the danger he poses to the ten and therefore morally innocent, or intentionally trying to bring about their deaths. The deontological theory to which you subscribe permits (or requires) you to shoot the skier in the latter case but prohibits it absolutely in the former.

As Jackson and Smith point out, formulating the antecedents of deontological principles in terms of intentions—e.g. that we are prohibited not from breaking a promise or killing the innocent but from *forming an intention* to break a promise or kill the innocent—does not avoid the need to deal with uncertainty, both because (as even Kant admits) we can be quite radically uncertain about the contents of our own intentions and because it is far from obvious whether and in what cases forming an intention to perform some action that I believe I *may* have promised not to do, or to kill someone who *may* be innocent, should count as impermissibly forming an intention *to break my promise* or *to kill the innocent*.

theories deals with the problem.[6] Now, just as the injunction to maximize expected value held in common by the various consequentialist theories considered in the last section extends easily to encompass intertheoretic moral uncertainty (among theories of the relevant class) as well as empirical uncertainty, so too the probability threshold view of deontological obligations provides ready guidance to an agent like David who divides his beliefs between various deontological theories that have this threshold view in common.

Suppose, for instance, that David must decide whether to kill Thomas in order to prevent Thomas from killing ten innocent people. He is empirically uncertain whether Thomas is an innocent threat or a malevolent aggressor, but he is also *morally* uncertain, dividing his beliefs between one absolutist deontological theory that prohibits the killing of innocent threats in other-defense and another that permits it. David, then, must divide his beliefs among three salient possibilities: (i) Thomas is a non-innocent threat, and may therefore permissibly be killed to save a greater number. (ii) Thomas is an innocent threat, and innocent threats may permissibly be killed. (iii) Thomas is an innocent threat, and innocent threats may not permissibly be killed.

It seems quite natural to say, given that both moral theories to which David assigns positive credence agree on a probability threshold of .5 at which deontological prohibitions subjectively "kick in," that he is prohibited from killing Thomas iff the probability he assigns to possibility (iii)—that is, the conjunction of the empirical belief that Thomas is an innocent threat and the moral belief that it is absolutely impermissible to kill an innocent threat—is greater than .5. Absent some prior reason for David to treat moral and empirical uncertainty differently in his practical deliberations, the principle he accepts with certainty, that it is subjectively wrong or impermissible to choose some practical option O whenever O carries a risk greater than .5 of violating an absolute moral constraint, seems to address the former as much as the latter.

Again, there is no knockdown positive argument that this is the only rational way for an agent in David's position to deliberate, any more than there is a knockdown argument that the .5 probability threshold is the right way for deontologists to deal with empirical uncertainty. The point, as before, is simply that there is a rationally perspicuous, non-arbitrary way of responding to a particular kind of intertheoretic moral uncertainty that takes advantage of principles for dealing with empirical uncertainty shared by some class of moral theories over which an agent distributes his or her credence.

Let's consider one more case of intertheoretic uncertainty, drawn once again from the consequentialist side of the moral universe, that highlights one important difficulty for intertheoretic comparisons relevant to the discussion in the previous section. Consider an agent Simone who divides her credence between classical utilitarianism and a sufficientarian theory that prioritizes the interests of those below some threshold of wellbeing, granting their interests twice the weight of those above the threshold. It may seem non-obvious how Simone should compare these two theories when they come into conflict, since it is non-obvious whether the sufficientarian theory gives *more* weight to the interests of those below the sufficiency threshold, *less* weight to the interests of those above the sufficiency threshold, or some combination of both. This sort of case raises

[6] Two prominent responses to Jackson and Smith's (2006) uncertainty-based objection to absolutist deontology, namely Hawley (2008) and Aboodi et al. (2008), both defend versions of this "threshold" proposal.

the worry that even among maximizing consequentialist theories, comparability may be the exception rather than the rule.

But although this case is more difficult than the hedonism/pluralism case of the last section, it nevertheless seems to me that Simone may well be in a position to make practically useful comparisons between her utilitarian and sufficientarian theories. In general, the question of whether the sufficientarian theory values the interests of the worse off more, or the interests of the better off less, ought to have an answer, even if it is non-obvious. This answer is to be found by examining the content of the theory that underlies and explains its assignments of value to states of affairs or practical options. For instance, the version of the sufficientarian theory to which Simone assigns positive credence might hold that, in addition to the value of wellbeing, there is some other "distributive" value like fairness or equality that counts in favor of raising people up to a threshold of adequate wellbeing. Or perhaps it holds that *deprivation* of basic needs, opportunities, etc is a disvalue over and above the suffering and absence of happiness that accompany it. In either of these cases, it would be clear that the prioritarian theory does not assign *less* value to the interests of the better off, but rather assigns *more* value to the interests of the worse off. On the other hand, Simone's sufficientarian theory might be a "second-personal" version of consequentialism according to which, while the intrinsic value of human wellbeing does not diminish as absolute levels of wellbeing increase, once one has reached an adequate level of wellbeing one can no longer reasonably demand of other agents that they value your wellbeing equally with their own.[7] In this case, the sufficientarian theory attaches *less* weight to the interests of the better off than the utilitarian theory, while attaching *equal* weight to the interests of the worse off.

It's possible, of course, that the version of sufficientarianism to which our agent assigns positive credence simply posits as a brute fact that the interests of those below the threshold carry more weight. In this case, there would seem to be no basis internal to the theories for normalizing them in any particular way. Nonetheless, it might still be plausible to claim that the sufficientarian theory assigns *at least as much* weight to the interests of the less well off as classical utilitarianism, and assigns *no more* weight to the interests of the better off. In this case, we would have *rough comparability*: The range of allowable normalizations between the theories would be constrained by the requirement that the ratio r between the value of the interests of the least well off according to sufficientarianism and the value of their interests according to utilitarianism satisfy $1 \leq r \leq 2$. Such rough comparability would tell Simone determinately which course of action has highest expected moral value in many, though of course not all, cases where her two theories conflict.

Similar things might be said about other, structurally similar cases, for instance an agent who divides her beliefs between two or more versions of pluralistic consequentialism, all of which assign non-derivative value to both hedonic and aesthetic goods, but which differ concerning the relative weight of these goods.[8] There is at least a hope, though no guarantee, that the explanations

[7] Of course, this sort of theory would presumably differ from the versions given above in holding that every agent has, at least, the option of valuing her own interests equally with those of the less well off, even if she herself is above the sufficiency threshold for wellbeing.

[8] I thank an anonymous reviewer for this example, and for encouraging me to consider this category of cases.

these theories offer for their relative weightings of hedonic and aesthetic value will furnish some perspicuous basis for precise intertheoretic comparisons. And if this hope is frustrated, we may at least be able to make rough comparisons.

Assuming I am right that intertheoretic comparability exists in (at least some variants of) the cases we have so far discussed, how widespread is the underlying phenomenon? Is every moral theory comparable with at least *some* closely related alternative theories, or are there some theories such that *any* uncertainty, concerning even the smallest of moral minutiae, generates full-blown intertheoretic incomparability that would preclude moral hedging? And are comparability classes of moral theories typically small, permitting intertheoretic comparison and hedging only for a few limited forms of moral uncertainty among very similar theories (e.g. uncertainty about which features of the world are value-bearers or which categories of action are subject to absolute moral prohibition), or can we find plausible bases for comparability that unite larger classes of theories, e.g. all forms of maximizing act consequentialism?

These questions, it seems to me, can only be answered by examining many individual cases on their own terms. And I have only been able to discuss a few of the myriad sorts of moral uncertainty that might be thought to generate barriers to intertheoretic value comparisons. It must be conceded, then, that comparability classes of normative theories may turn out to be few, small, and far between. Nevertheless, I will argue in the next section, the existence of such classes is enough to answer the argument from PIVC to a general rejection of hedging, and even limited intertheoretic comparability may play an important role in a general procedure for rational decision-making under moral uncertainty.

4 Is incomparability anywhere a threat to comparability everywhere?

So far I have suggested that an agent who divides her beliefs among a class of sufficiently similar moral theories can possess a rational basis for intertheoretic value comparisons and hence for moral hedging. But I have not addressed the (doubtless more realistic) case of an agent who assigns positive degrees of belief to a much more diverse class of theories, a class that shares no obvious structural features that would naturally ground such comparisons—e.g. an agent who assigns positive credence both to a maximizing consequentialist theory and to an absolutist deontological theory.

Toward the end of this section I will mention some possible approaches to these "hard cases" of intertheoretic value comparison, but I have no general solution to offer. Absent such a solution, the pessimist about intertheoretic comparisons may argue as follows: "The principle of hedging under moral uncertainty requires that agents be able to compare degrees of value across theories to which they assign positive degrees of belief, in general. Even if there are some easy cases in which it looks like such comparisons can be made, as long as the hard cases remain unresolved, the problem of intertheoretic value comparisons as a whole is unresolved and so still poses a decisive obstacle to the principle of intertheoretic moral hedging."

My reply is simply this: Someone who opposes any form of hedging for moral uncertainty, but shares the ordinary view that agents should be responsive to empirical risks and uncertainties, must justify drawing a line between the moral and the empirical. What the above cases of intertheoretic comparability show is that PIVC cannot justify drawing such a line. Rather, absent some *other* argument that agents should not be responsive to their moral uncertainties,[9] problems of incomparability suggest the need for special principles of rationality to handle *incomparability*, rather than special principles (or a special absence of principles) for responding to moral as opposed to empirical uncertainty.

The existence of plausible procedures for value comparison in certain cases of intertheoretic moral uncertainty, taken together with the various forms in *in*comparability that can arise in the absence of any moral (or other normative) uncertainty, suggests that the problem of incomparability is simply orthogonal to the distinction between moral and empirical uncertainty. Just as intertheoretic value comparisons do not *always* generate (even apparent) incomparability, so (apparent) incomparability of normative considerations does not arise only in the context of intertheoretic comparisons. Many philosophers, for instance, have held that certain pairs of moral values are genuinely incomparable—that is to say, they have advocated first-order moral theories that generate *intra*theoretic incomparability. An agent might believe, say, that patriotic and familial obligations are genuinely incomparable, such that when these values come into conflict (as when she feels called to volunteer for a war of national self-defense, but cannot do so without abandoning an ailing relative), there is no uniquely rational resolution. Incomparability may arise in more mundane, non-moral contexts as well—for instance, if I am shopping for home decor and must choose between satisfying my own aesthetic sensibilities and pleasing my spouse, whose tastes I regard as gauche.[10] Finally, even on a simple utilitarian theory of value that does not admit this sort of incomparability, purely empirical uncertainty can give rise to a different species of incomparability, as it does in the Pasadena Game (Nover and Hájek, 2004), for which the expected values of playing and not playing appear to be incomparable.

But it would be absurd, on the basis of such cases, to deny the rationality of quantitative value comparisons in general, to claim that the existence of one or more of these forms of incomparability shows that the procedure of expected value maximization is unreasonable in any context and hence to claim that there is no rational basis, say, for investors in the stock market to choose investments that maximize their expected financial returns. Likewise, it is equally absurd to suggest that the (apparent) impossibility of quantitative comparison between utilitarian and deontological moral considerations should also preclude comparisons, e.g., among possible values in a consequentialist scheme under conditions of more limited moral uncertainty.[11]

[9] Many such arguments have been attempted; see for instance Weatherson (2014), Harman (2015).

[10] I thank an anonymous reviewer for this example.

[11] MacAskill (2013) points out that incomparability can be "infectious" in that an agent who has any positive credence, however slight, that incomparable values are at stake in a given choice situation and tries to calculate the expected value of her various options will in general find them to be undefined. But again, this worry is not unique to the problem of moral uncertainty, and is not solved by ignoring moral uncertainty. The investor who has some vanishingly small credence that her investment decisions implicate values like patriotism and family that she judges, as a matter of first-order moral theory, to be incomparable, will find herself in a similar predicament. The problem of infectious incomparability, in this respect, is structurally analogous to more familiar decision-theoretic problems like Pascal's Wager (a problem, one might say, of "infectious infinities"), and must likewise be resolved in a general

If there is genuine incomparability between the considerations put forward by certain pairs of moral theories, then a division should be drawn between the principles of rationality to be employed where comparison is possible and those to be employed where it is not. What might such a division look like? I will close this section by describing, very speculatively, a picture that seems to emerge from the idea that it is sometimes but not always possible to make non-arbitrary intertheoretic comparisons of value.

Very generally, the picture is that the sorts of two-level procedures for practical deliberation that opponents of moral hedging have suggested may be recast with the boundary between levels drawn not in terms of empirical vs. moral uncertainty but rather in terms of uncertainty within vs. between comparability classes of moral theories whose shared assumptions can ground intertheoretic value comparisons.[12]

Such a two-level procedure, as advocated by the opponents of hedging, directs an agent to respond to her empirical uncertainties in ordinary, intuitive ways (e.g. by maximizing expected value), but with respect to her moral beliefs, they direct her either to "take her best guess" and base her deliberations, judgments of the value of consequences, etc on the moral theory to which she assigns the largest portion of her credence (the so-called My Favorite Theory view, advocated *inter alia* by Gracely (1996) and Gustafsson and Torpman (2014)) or to simply act on the *true* moral theory, regardless of her moral beliefs or the evidence available to her (the view advocated *inter alia* by Weatherson (2014), Harman (2015), and Hedden (2016)).[13]

My suggestion instead is that if such a two-level procedure is appropriate, the role in such procedures that opponents of hedging give to moral theories (as either subject to a "best guess" principle or completely belief-insensitive in their deliberative role) is better given to comparability classes of moral theories. Thus it might be, for instance, that the agent who finds herself faced with a trolley dilemma and divides her beliefs between utilitarian and deontological theories has no alternative but to take her best guess as to which sort of theory is correct—that is, there may turn out to be no rational procedure by which she can weigh these very different sorts of moral considerations against each other. But an agent whose relevant moral uncertainties are *within* a class of sufficiently similar deontological or consequentialist theories can do better than guessing, and should. And an agent who distributes her beliefs over various consequentialist and deontological theories may at least weigh some of her uncertainties, e.g. between versions of consequentialism that accept different theories of value, before incomparability forces her to adopt another decision procedure (say, act-

way by decision theory.

[12] Somewhat more precisely, a comparability class can be defined as a class of comprehensive moral theories (maximal consistent sets of moral propositions) that bear the relation of comparability to some particular theory. I will assume, though I have not argued, that intertheoretic comparability is an equivalence relation, i.e. transitive, reflexive, and symmetric. Reflexivity and symmetry seem uncontroversial. It is less obvious, but still quite plausible, that intertheoretic comparability must be transitive, i.e. that if comparison is possible between the moral considerations offered by theories T_1 and T_2 and likewise between T_2 and T_3, then comparison must be possible between T_1 and T_3.

[13] These views may be described, more succinctly but in a way that somewhat disguises the two-level structure, by saying that an agent subjectively ought to choose the option that is implied to be most subjectively choiceworthy by the combination of the agent's empirical belief state and either (a) the moral theory to which she assigns the greatest portion of her credence or (b) the true moral theory.

ing on the aggregated verdict of the consequentialist theories in which she has positive credence, if she regards consequentialism in general as more plausible than deontology).[14]

Nonetheless, I have not denied that there may be valid bases for intertheoretic comparison besides the shared-content approach I have described in this paper. So two possibilities remain open: Either there is some non-obvious basis for intertheoretic value comparisons between very diverse moral theories, in which case (setting aside other objections) agents can engage in "global" moral hedging, or else rational deliberation for an agent who is widely uncertain about both empirical and moral matters will involve a two-level procedure of comparing and weighing considerations within comparability classes of theories followed by some other (e.g. best-guess) procedure between comparability classes.

One version of the former approach is advocated by MacAskill (2014), who argues that where no other basis for comparability exists, moral theories should be normalized at the *variance* of their value assignments (just as Sepielli's Conceivability PEMT would normalize theories at the *range* of their value assignments). Such proposals are intuitively attractive to the extent they can vindicate the common judgment that certain comparisons are possible even among vastly disparate theories. For instance, consider an agent who divides her beliefs equally between classical utilitarianism and Kantianism, and must decide whether to push one innocent person in front of a trolley, causing her death, in order to both save the life of a second innocent person and prevent $1000 worth of damage to a third innocent person's car. It seems intuitive that the wrongness of killing an innocent person in this way according to Kantianism is greater than the rightness of preventing $1000 worth of damage to an automobile according to utilitarianism. (If the intuition is not immediate, suppose that the car's owner is prosperous, well-insured, and has two more cars at home.)

If a strategy like MacAskill's can succeed, even with less then perfect generality (e.g., as MacAskill suggests, allowing comparisons between cardinal but not ordinal theories), then co-membership in a comparability class of theories is not a necessary condition for intertheoretic comparability. My purpose has only been to argue that it is a *sufficient* condition, and that this defeats the argument from the problem of intertheoretic value comparisons to the conclusion that one ought never hedge for one's moral uncertainties.[15]

[14] This approach to choice under moral uncertainty would have the advantage of at least mitigating the most important objection to the My Favorite Theory view, namely the problem of *theory individuation*. Gustafsson and Torpman, who give the most detailed defense of MFT in the literature to date, stipulate that an agent should treat two moral theories as distinct unless she is certain that they will never yield different practical prescriptions (Gustafsson and Torpman, 2014, p. 13). But as MacAskill points out, this implies that an agent who distributes her beliefs over many fine-grained moral theories may well be required to act on a theory in which she has minuscule credence, even when this theory prescribes a course of action that she believes with near-certainty is much worse than some available alternative (MacAskill, 2014, pp. 24-25). Since comparability classes of moral theories are, plausibly, much less fined-grained than individual theories, a version of MFT that allowed hedging within comparability classes would at least weaken the intuitive force of this objection.

[15] I thank an anonymous reviewer for encouraging me to clarify this point.

5 Conclusion

I have argued that that PIVC is not an objection to moral hedging as such, since within certain classes of moral theories, intertheoretic value comparisons appear to be straightforwardly possible and well-motivated. And I have suggested, more tentatively, that we should look for a distinction between the principles of rationality that govern not empirical vs. moral uncertainty, but rather uncertainty within vs. between comparability classes of moral theories.

Let me close by drawing out the implications of this idea for two other objections to moral hedging. The first, alluded to in passing in the first section, is a worry about "grounding": If the true comprehensive moral theory is the sum of all moral truths, and this theory holds that the considerations distinctively forwarded by other, false moral theories have no (non-derivative) moral significance, then what moral facts could possibly serve to "ground" or make true any purported exchange rate between the values of rival moral theories?[16]

A second, closely related objection holds that moral hedging initiates a vicious regress: If, before acting on any normative principle, we must always weigh the possibility that the principle is mistaken, this requirement will apply as well to those "second-order" principles that we use to do the weighing, e.g. to the principles that tell us how to hedge for our first-order moral uncertainties. But then if we are at all uncertain about these second-order principles, we will need "third-order" principles that tell us how to weigh the demands of competing second-order principles against one another, and so on *ad infinitum*.

The more modest idea of hedging within comparability classes of moral theories suggests a response (or at least a partial response) to both these objections. Our original agent Clara who divides her beliefs among a variety of monistic and pluralistic consequentialist theories can say quite easily what grounds the comparisons she makes between the prescriptions of these various theories: namely, the comparisons between categories of value-bearer internal to the theories themselves. Even if the pluralistic theories that establish exchange rates between, say, hedons and aesthetons, turn out to be false, the rationality of the hedging procedure can be adequately grounded by her credence in these theories plus true principles of rationality that tell her, where possible, to maximize the expected value of her choices, or more generally to seek the good, avoid

[16] This line of objection to hedging has been suggested by several philosophers, though to the best of my knowledge it has not been identified as a distinct line of argument in the recent literature. I take this worry about grounding to be, for example, at least one of the lines of argument suggested in the following passage from Hudson (1989): "Even mere axiological uncertainty within an unquestioned subjective consequentialist framework is unhedgeable. Suppose the agent assigns probability 0.6 to the view that pleasure-minus-pain is the only intrinsic good, and 0.4 to the view that the good is self-realization. And suppose she must choose between an act that produces ten hedons and two reals and one that produces nine hedons and thirty reals. ('Reals' are the units in which self-realization is measured.) Which act should she do? The two axiological theories lead to different answers. Since the hedonic theory is more probable, perhaps she should accept its answer. But the self-realization theory seems to find more of a difference between the two actions, and perhaps this should outweigh its slightly lesser probability. But wait—is a difference of twenty-eight reals really greater than a difference of one hedon? What is the common measure between hedons and reals? *Note that the agent, for all her uncertainty, believes with complete confidence that there is no common measure: she is sure that one or the other—pleasure or self-realization—is intrinsically worthless.* Under the circumstances, the two units must be incomparable by the agent, and so there can be no way for her uncertainty to be taken into account in a reasonable decision procedure." (Hudson, 1989, p. 224; italics mine)

risks of serious wrongdoing, etc. Likewise, insofar as the principles that ground intertheoretic comparisons are shared by all theories in a comparability class, the regress problem does not afflict hedging procedures so long as they are kept internal to such classes. Conditioning one's credences on the disjunction of all theories composing such a class, the principles on which the appropriate hedging procedure is based (e.g. that one should maximize expected value, or should not take actions that carry a probability greater than .5 of violating a deontological prohibition) are assigned probability 1, as logical implications of every theory in the class.

Looking for comparability classes of moral theories within which non-arbitrary procedures for intertheoretic value comparison can be established is, then, at least a fruitful starting point for proponents of moral hedging, serving to weaken many of the standard objections to more "global" hedging procedures. How easy it will be to find bases for comparison among other kinds of moral theories (contractualist, rule consequentialist, virtue ethical...) and how wide or narrow these classes will ultimately turn out to be remain open questions. It seems possible that broader classes of moral theories (e.g. consequentialist or deontological theories in general) might offer more vague or tenuous bases for intertheoretic comparisons, while within particular subclasses more definite and compelling comparisons will be possible. But in any case, so long as there are clear and compelling comparisons to be made in even a few cases, the problem of intertheoretic comparisons should not lead us to abandon the hope of rationally accounting for our moral uncertainties in decision-making.

References

Aboodi, R., A. Borer, and D. Enoch (2008). Deontology, Individualism, and Uncertainty: A Reply to Jackson and Smith. *Journal of Philosophy 105*(5), 259–272.

Gracely, E. J. (1996). On the Noncomparability of Judgments Made by Different Ethical Theories. *Metaphilosophy 27*(3), 327–332.

Gustafsson, J. E. and O. Torpman (2014). In Defence of My Favourite Theory. *Pacific Philosophical Quarterly 95*(2), 159–174.

Harman, E. (2015). The Irrelevance of Moral Uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Volume 10. Oxford University Press.

Hawley, P. (2008). Moral Absolutism Defended. *Journal of Philosophy 105*(5), 273–275.

Hedden, B. (2016). Does MITE Make Right? On Decision-Making under Normative Uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Volume 11. Oxford University Press.

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

Jackson, F. and M. Smith (2006). Absolutist Moral Theories and Uncertainty. *Journal of Philosophy 103*(6), 267–283.

Lockhart, T. (2000). *Moral Uncertainty and Its Consequences.* Oxford University Press.

MacAskill, W. (2013). The Infectiousness of Nihilism. *Ethics 123*(3), 508–520.

MacAskill, W. (2014). *Normative Uncertainty.* Ph. D. thesis, University of Oxford.

Nissan-Rozen, I. (2015). Against Moral Hedging. *Economics and Philosophy 31*(3), 1–21.

Nover, H. and A. Hájek (2004). Vexing Expectations. *Mind 113*(450), 237–249.

Ross, J. (2006). Rejecting Ethical Deflationism. *Ethics 116*(4), 742–768.

Sepielli, A. (2009). What to Do When You Don't Know What to Do. *Oxford Studies in Metaethics 4*, 5–28.

Sepielli, A. (2010). *'Along an Imperfectly-Lighted Path': Practical Rationality and Normative Uncertainty.* Ph. D. thesis, Rutgers University Graduate School - New Brunswick.

Sepielli, A. (2013). Moral Uncertainty and the Principle of Equity among Moral Theories. *Philosophy and Phenomenological Research 86*(3), 580–589.

Weatherson, B. (2014). Running Risks Morally. *Philosophical Studies 167*(1), 141–163.