# Does MITE Make Right?
# On Decision-Making Under Normative Uncertainty

Brian Hedden

## 1 Introduction

We are not omniscient agents. Therefore, it is our lot in life to have to make decisions without being apprised of all of the relevant facts. We have to act under conditions of uncertainty. This uncertainty comes in at least two kinds. First, there is ignorance of *descriptive* facts; you might be ignorant of the potential causal impacts of the various actions available to you. For instance, you might be unsure whether to give the pills to your headache-suffering friend because you are uncertain whether they are painkillers or rat poison. Second, there is ignorance of *normative* facts, or facts about whether a certain action or outcome is good or bad, permissible or impermissible, blameworthy or praiseworthy, etc.[1] For instance, you might know exactly what would happen (descriptively speaking) if you (or your partner) had an abortion and what would happen if you didn't, and yet still be uncertain about whether having an abortion is a morally permissible thing to do.[2] Due to the ubiquity of normative, and not just descriptive, uncertainty, we might want a theory that provides some guidance about how to take this normative uncertainty into account in deciding what to do. While I will be concerned with specifically *moral* uncertainty, much of what I say will carry over to other cases of normative uncertainty, such as uncertainty about what would be instrumentally rational to do or what it would be epistemically rational to believe.

At this stage-setting phase, some terminology will be helpful. Let us say that what you *objectively* ought to do depends only on how the world in fact is, and not on how you believe the world to be. For Utilitarians, what you objectively ought to do is whatever will in fact maximize happiness, irrespective of your beliefs about what will maximize happiness. For non-consequentialists, what you objectively ought to do might depend, for instance, on facts about whether

[1] Moral non-cognitivists might resist talk of moral facts. But I do not take this paper to be committed one way or the other regarding moral cognitivism vs. non-cognitivism. I speak of moral facts merely for the sake of convenience. See Sepielli (2012) for discussion of how the problem of what to do under conditions of normative uncertainty arises even for non-cognitivists.

[2] This example is from Sepielli (2009).

some act would in fact cause an innocent person to die (thereby violating that person's rights), irrespective of your beliefs about whether it would cause an innocent person to die.

And let us say that what you *subjectively* ought to do depends in some way on your descriptive beliefs about how the world is. For consequentialists, what you subjectively ought to do might be whatever will maximize *expected* world value (expected total happiness, for utilitarians), relative to your beliefs. And for non-consequentialists, what you subjectively ought to do might depend on whether you *believe* that some act would cause an innocent person to die. (Note that my usage of the term 'subjective ought' differs from that of some authors, who define what you subjectively ought to do as whatever you believe that you objectively ought to do. There are at least two ways in which my usage of the term differs from theirs. First, their usage is incompatible with an expectational account of what you subjectively ought to do, such as the consequentialist one just mentioned. Second, their usage makes the subjective ought simultaneously sensitive to both descriptive and normative uncertainty.)

Neither the objective *ought* nor the subjective *ought*, on my usage, is sensitive to your moral uncertainty. We might then introduce a super-subjective *ought* and say that what you super-subjectively ought to do depends on both your descriptive uncertainty and your moral uncertainty. This gives us a tripartite distinction:

> Objective *ought*: Insensitive to your descriptive and moral uncertainty.
>
> Subjective *ought*: Sensitive to your descriptive uncertainty but insensitive to your moral uncertainty.
>
> Super-subjective *ought*: Sensitive to your descriptive and moral uncertainty.

Before moving on, let me flag that making these distinctions does not commit one to the claim that the English word 'ought' is ambiguous or that it admits all and only these three readings. It may be, for instance, that modals are highly context-sensitive, so that different contexts can give rise to all sorts of readings of 'ought' claims. My aim here is simply to highlight some possible senses of 'ought' that may be of particular interest to normative theorists.

The remainder of this paper is dedicated to evaluating the prospects for theorizing about decision-making under moral uncertainty. My evaluation, which is largely negative, has two parts. I begin by examining what has already emerged as the preeminent proposal for what you super-subjectively ought to do. This proposal takes the dominant theory of what you subjectively ought to do, namely expected value theory, and attempts to extend it to take into account your moral uncertainty as well. I argue that this proposal is unworkable.

In the second part of my evaluation, I question whether we should want an account of decision-making under moral uncertainty in the first place. I tentatively suggest that a super-subjective *ought* has no important role to play in our normative theorizing and should thus be abandoned. There is no normatively interesting sense of *ought* in which what you ought to do depends on your uncertainty about (fundamental) moral facts.[3] In this respect, moral uncertainty is importantly different from descriptive uncertainty.

## 2 Does MITE Make Right?

While theorizing about what you objectively and subjectively ought to do has a long and distinguished history, theorizing about the super-subjective *ought*, about decision-making under normative uncertainty, is still in its infancy. But already, a preeminent theory has emerged. This theory incorporates insights from the preeminent theory of the subjective *ought*, namely expected value theory. For this reason, it will be helpful to start with a brief overview of this theory.

Suppose that we can represent your doxastic, or belief-like, state with a probability function $P$ and that the value function $V$ represents how good or bad different outcomes $O_i$ are. If we are interested in what you *morally* ought to do, then $V$ can be thought of as representing moral goodness in some way, while if we are interested in what you *prudentially* ought to do, then it can represent your own preferences or levels of happiness. Since we are concerned with morality, let us understand $V$ in the former way. Then, we say that what you subjectively ought to do is to make-true the act-proposition with the highest expected moral value, defined thus:

*Expected Moral Value*:
$\text{EMV}(A) = \sum_i P(O_i \mid A)V(O_i)$[4]

Given the attractiveness of the expected value maximization framework for theorizing about the subjective *ought*, it is tempting to try to extend it to the super-subjective *ought*. If it is possible to represent all moral theories in expected value terms (this assumption will be questioned shortly[5]), then there is

[3] This is compatible with the idea that there may be natural language uses of 'ought' where the context is such as to give rise to a reading on which it is sensitive to your moral uncertainty. I am just denying that such a sense of 'ought' is important for purposes of normative theorizing.

[4] This is the formula for *evidential* expected value, rather than *causal* expected value. The debate over evidential decision theory and causal decision theory is an important one, but it is not our topic, and so I set it aside.

[5] See Sen (1982), Oddie and Milne (1991), Dreier (1993, Smith (2009), Colyvan, Cox, and Steele (2010), and Portmore (2011) among others, for discussion of this question outside the context of moral uncertainty. In my view, the considerations raised in section 2.2, among others, show that not all moral theories can be represented in expected value terms (or 'consequentialized,' insofar as this means being represented using a value function). More exactly, I take these considerations to show that not all moral theories can be represented in the same expected value maximization framework. Perhaps there are different modifications of the expected value framework that can helpfully represent different moral theories, but they cannot all be squeezed into the same framework. But that seems to be what is necessary in order to do the relevant trade-offs and aggregations needed to yield a theory about what on ought to do in light of one's moral uncertainty.

an apparently straightforward way in which to extend the expected value framework to deal with moral uncertainty as well. Expected moral value ($EMV$) is an *intra*theoretical notion. When we take the expected moral value of an action on each moral theory and sum them up, weighted by the probability of each theory, we get an *inter*theoretical notion, which we can call the 'intertheoretic expectation.'

*Intertheoretic expectation*:
$$\mathrm{IE}(A) = \sum_i P(T_i)EMV_i(A) = \sum_i P(T_i) \sum_j P(O_j|A)V_i(O_j)$$

Now, the proposal is that what you super-subjectively ought to do is to make-true the act-proposition with the highest intertheoretic expectation. Let us call this theory 'MITE,' for: **M**aximize **I**nter**T**heoretic **E**xpectation. MITE is a natural extension to the super-subjective *ought* of expected value theory as a theory of the subjective *ought.* Expected value theory evaluates an action by looking at how objectively good (or bad) an action would be in different states of the world and discounting that goodness by your degree of belief that that state of the world is actual. MITE evaluates an action by looking at how (subjectively) good (or bad) an action would be according to each moral theory you take seriously and discounting that goodness (or badness) by your degree of belief that that moral theory is correct.

Versions of MITE have been defended by Lockhart (2000), Ross (2006), and Sepielli (2009, dissertation), and it has swiftly established itself as the dominant theory of the super-subjective *ought.* This is no accident. MITE has the attractive feature of taking into account both how confident you are in each moral theory *and* how good or bad the given act would be, according to each of those moral theories. (Later I will be questioning whether it makes sense to speak of how good or bad an act is, according to different moral theories, but for now I grant the intuition that such talk does make sense.)

By contrast, a decision rule which just recommended acting in accordance with the moral theory to which you assign highest credence would ignore facts about the relative goodness or badness of acts according to the different moral views.[6] You might be 51% confident that having an abortion would be slightly morally better than not having one, and 49% confident that having an abortion would be absolutely monstrous, but this decision rule would say that you should just go with the view you're 51% confident in. Similarly, a maximin-style decision rule which recommends ranking acts according to their worst possible moral badness and then performing the highest act in that ranking would ignore your differing levels of confidence in each moral theory. (In the next section, however, I will be questioning whether there are any grounds for making these sorts of comparisons between how good or bad a given act is, according to different moral theories.)

[6] This view is sometimes called the 'My Favorite Theory' view, and is defended by Gracely (1996) and Gustafsson and Torpman (2014).

For this reason, I would venture so far as to say that when it comes to trying to devise a formal theory of what you super-subjectively ought to do, MITE (or some slight variant thereof) is the only game in town. This is important, since if MITE ultimately fails, as I will argue it does, then this casts serious doubt on the prospects for coming up with *any* formal theory of what you super-subjectively ought to do.

In 2.1 and 2.2, I consider two serious problems for MITE that show that its ambitions must be considerably scaled back. The first is the problem of intertheoretic value comparisons, first noted by Hudson (1989), Gracely (1996), and Lockhart (2000). To employ MITE, we must make precise comparisons of 'degrees of wrongness' across moral theories. I argue that there is no principled way to make these comparisons, unless we start off with a considerable number of judgments about what agents in various circumstances super-subjectively ought to do. Thus, MITE can at best aspire to take us from a smaller set of judgments about the super-subjective *ought* to a larger set of such judgments. The second problem is the impossibility of adequately representing certain sorts of moral theories, such as theories which distinguish between supererogatory and merely permissible acts, in expected value maximization terms, as MITE requires. If there are moral theories that cannot be squeezed into the expected value maximization framework that MITE presupposes, then MITE cannot say anything about what an agent who assigns any credence to such theories super-subjectively ought to do. Thus MITE cannot provide a general framework for decision-making under moral uncertainty.

## 2.1 Axiological Uncertainty and the Problem of Intertheoretic Value Comparisons

Let us begin with a type of moral uncertainty which would seem to be naturally and fruitfully dealt with by MITE. Consider an agent who is certain that (maximizing) consequentialism is correct; that is, she is certain that one ought to maximize value. However, she is uncertain about what is of value. She doesn't know what the right axiology is. It would seem that we should be able to straightforwardly give her advice about what to do by calculating the expected moral values of the available actions, relative to the value function corresponding to each possible axiology, and summing up those expected moral values, weighted by her degree of belief that the corresponding axiology is correct, thus arriving at an intertheoretic expectation for each action.

But even in this highly artificial case, we already run into problems. In particular, we run into the problem of calibrating value functions. As we know from decision theory, a preference ordering (satisfying certain axioms) over worlds and prospects (gambles) does not uniquely determine a value function. Instead, such a preference ordering only determines a value function which is unique at most up to addition of a constant and multiplication by a positive scalar.[7] As such, if

[7] For the systems of von Neumann and Morgenstern (1944) and Savage (1954), if your preferences satisfy their axioms, you are representable as an expected utility maximizer with

the value function $V$ represents a given set of preferences, so does the function $aV + b$, for real numbers $a$ ($> 0$) and $b$. Axiologies generally only give us a preference ordering, but in order to apply the expected value framework to cases of axiological uncertainty, we need to fix on one value function corresponding to each axiology. And it is doubtful whether there is any principled reason for privileging any one function from axiologies to value functions over the other possible such functions. This is the *problem of intertheoretic value comparisons*.

The thrust of this problem can be seen through an example which is well-known from Parfit (1984). Even if one is certain that happiness is what matters, one can be uncertain about whether worlds are ranked by *total* happiness or by *average* happiness. This uncertainty will be important in situations where one has the option of implementing a policy which will increase the world's population, but at the cost of decreasing average happiness. In order to give guidance to the agent making this choice using MITE, we have to choose value functions to correspond to Totalism and to Averagism. However, it appears that any such choice will be arbitrary and have unintuitive consequences.[8]

Suppose we start with a simple proposal – for Totalism we let the value of a world be the total happiness in that world, while for Averagism we let the value of a world be the average level of happiness. Unfortunately, this will have the result that for most real-life cases where one can substantially increase population at the cost of decreasing average happiness, our framework will recommend doing what Totalism recommends unless the agent is overwhelmingly confident that Averagism is correct.

Suppose that the agent has the choice of increasing the world's population from 6 billion to 24 billion people at the cost of halving the average happiness level. Let the present average happiness level be x ($x > 0$). Then, for Totalism, the difference between the expected moral value of increasing the world's population and the expected moral value of the status quo will be $24,000,000,000 \times (x/2) - 6,000,000,000x = 6,000,000,000x$. For Averagism, the difference between the expected moral value of increasing the population and the expected moral value of the status quo is $-(x/2)$.

Crunching the numbers, maximizing intertheoretic expectation will recommend that the agent implement the population-increasing policy (i.e. doing what Totalism recommends) unless she is over 99.9999999916% confident that Averagism is right. But this seems crazy.

We could perhaps improve things by representing Averagism not by the value function that assigns each world its average happiness as its value, but rather by a value function that assigns each world some large multiple of its average happiness as its value. But this proposal is not without its own problems.

No matter what value functions we use to represent Averagism and Totalism, once we fix on proposed decrease in average happiness, Averagism will swamp

a utility (or value) function that is unique up to positive linear transformation. In Jeffrey's (1983) system, the uniqueness condition for utility functions is more complicated, but nonetheless it is true that if $V$ represents your preferences, so does $aV + b$ ($a > 0$).

[8] William MacAskill recently informed me that he also uses Totalism and Averagism to illustrate this point, though he attributes it to Toby Ord. See MacAskill (2014, 93-4).

Totalism for smaller population increases while Totalism will swamp Averagism for larger population increases. This is perhaps natural enough. After all, in situations where one can increase population by decreasing average happiness, Totalism will say that the moral significance of the situation increases with the size of the possible increase in population, while Averagism will say that the moral significance of the situation does not depend on the size of the possible population increase. So we would expect Averagism to outweigh Totalism for small possible population increases, and we would likewise expect Totalism to outweigh Averagism for very large possible population increases. The problem is that representing Totalism and Averagism by particular value functions requires us to choose a point along the continuum of possible population increases where Totalism starts to outweigh Averagism (for a given reduction in average happiness). And any such choice will seem arbitrary and unmotivated. There is nothing in the moral theories themselves that tells us how to make intertheoretic value comparisons.[9]

Can we make any plausible non-question-begging stipulations about interthe-

[9] The astute reader may notice a structural similarity between the problem of intertheoretic value comparisons for MITE, and the familiar problem of interpersonal comparisons of utility for theories of social choice. One difference, however, is that we may have some grip on how to make interpersonal comparisons of utility that doesn't depend just on the functions that we'd get if we used Ramsey's (1931) method to construct utility functions for the individuals involved. For one, our shared biology may provide some grounds for calibration–it seems plausible that two people undergoing the same painful medical procedure, with each protesting as loudly as the other and displaying similar patterns of neuronal activity, perspiration, and other common indicators of discomfort, should be treated as suffering a similar level of disutility, at least for the purposes of social choice. While such considerations may help us ground interpersonal comparisons of utility, it's not obvious whether there's anything that could play a similar role in grounding intertheoretic comparisons of value.

While I'll discuss a different method for attempting to solve this problem from Sepielli (2009) later in this section, in more recent work (Sepielli (2010, ch 4)), he offers a strategy that's somewhat analogous to the one I've just suggested might work in the case of interpersonal utility comparisons. He suggests that we might be able to appeal to conceptual connections between various normative concepts in order to ground intertheoretic value comparisons. Just as we might ground interpersonal utility comparisons by assuming that people in similar behavioral and neurological states are undergoing similar levels of disutility, we might ground intertheoretic value comparisons by assuming, for instance, that if two theories recommend similar degrees of blame for an act, that they each regard the reasons against that act as equally weighty. While Sepielli acknowledges that he hasn't provided a detailed, psychologically realistic account of the various conceptual connections between normative concepts of the sort he thinks would solve the problem of intertheoretic value comparisons, there are reasons for skepticism about the prospects for any such strategy. For example, two moral theories might disagree about how much we should blame somebody for acting in a certain way for reasons that have nothing to do with what they say about the reasons in favor of acting in that way (Gustafsson and Torpman (2014) and MacAskill (2014) also make this point). One theory might imply that we ought never blame anybody because it implies that justified blame would require contra-causal free will, while the other theory might be compatibilist about blame. Similar issues will also arise with consequentialist theories on which whether blame is recommended in a given circumstance depends not on the wrongness of the act in question, but rather on the consequences that would result from blaming. I raise this example to motivate skepticism that there is anything like a silver bullet that will allow us to determine that two theories must be interpreted as assigning some act equal value, so long as they agree on some other normative claim.

oretic value comparisons? In the remainder of this section, I look at three prominent proposals for doing so and find them wanting. Start with Lockhart (2000), who proposes a *Principle of Equity among Moral Theories* (PEMT), according to which all moral theories should be deemed to have the same amount of moral rightness at stake in any given situation. In each situation, the worst available actions according to each moral theory should be assigned the same (low) expected moral value, and similarly for the best available actions according to each moral theory. This is a version of the 'zero-one' rule, a proposal for solving the problem of interpersonal comparisons of utility by scaling each person's utility function to the zero-one interval. (Note that the PEMT will likely require us to use different value functions to represent a given moral theory in different choice situations.)

Unfortunately, the PEMT is implausible (see Ross (2006) and Sepielli (2013)). It arbitrarily rules out the possibility of situations in which moral theories would seem to differ dramatically in how morally significant they consider the choice at hand. Consider again the case of Averagism and Totalism. We can imagine a scenario in which one has the option of creating on another planet a population of ten billion people who are all just slightly less happy than the average here on earth - the difference between our average happiness and theirs is equivalent, say, to the difference between not having a hangnail and having one. The PEMT rules out by fiat the possibility of saying that this is a situation that carries far more weight for Totalism than for Averagism. Now, I am not claiming that this in fact is a situation that carries more weight for Totalism than for Averagism. After all, I am denying the possibility of making such intertheoretic value comparisons. My claim is simply that there is no intuitive support for the PEMT's claim that this is a situation that is equally weighty for Averagists and Totalists, and that more generally, moral theories cannot differ in how morally significant they consider a given choice to be.[10]

[10] Of course, we could modify the PEMT and instead stipulate that all moral theories should be treated as having the same maximum and minimum *possible* moral value at stake. That is, we consider the worst possible actions (not holding fixed a given choice situation) according to the various theories and make sure that they are all assigned the same (very low) expected moral value, and we also consider the best possible actions according to the competing theories and assign them all the same (very high) expected moral value. But this too is implausible. First, there is little reason to think that there will be worst and best possible actions for given moral theories, or even that expected moral value should be bounded for every moral theory (Sepielli (2013)). Certainly, utilitarians will likely think that possible acts grow better and better without bound as more and more happiness is created, and also that acts grow worse and worse without bound as more and more suffering is created. Second, some moral theories may just think that no possible situation can be terribly significant from a moral standpoint. Various moral nihilistic views hold that no acts are morally better than any others. Note, however, that such nihilistic theories are independently problematic for MITE, since some versions of decision theory prohibit all acts and outcomes being equally preferred. For instance, Savage's (1954) postulate P5 says that it is not the case that for all pairs of acts, one is at least as good as the other. One can also imagine slight deviations from moral nihilism which hold that no acts are are substantially morally better than any others (MacAskill (2014, 135)). It would be a distortion of what such a view says to represent it as being such that its best and worst possible acts have the same expected moral values as the utilitarian's best and worst possible acts, respectively. This is especially relevant for Ross (2006), who

Next consider an interesting proposal made by Sepielli (2009) (though Sepielli (2010) disavows it). Sepielli's approach relies on the existence of some background agreement among moral theories that will serve as a fixed point that we can use to make the requisite intertheoretic value comparisons.[11] The idea is to find at least three actions or outcomes A, B, and C such that all of the moral theories the agent takes seriously agree that A is better than B, which is better than C and also agree about the ratio of the value difference between A and B and the value difference between B and C. We then stipulate that the value functions chosen to represent each moral theory must agree in the numbers they assign to A, to B, and to C.

Consider Averagism and Totalism again. They agree about the one-person case. They agree that a world A where there is one person with happiness level 10 is better than a world B where the one person has happiness 4, which in turn is better than a world C where the one person has happiness 2. Moreover, they agree on the ratio of value differences between A and B, and B and C; they agree that the value difference between A and B is three times the value difference between B and C. So, on Sepielli's proposal, we just pick three numbers x, y, and z to serve as the values of A, B, and C for both Averagism and Totalism, with the constraints that $x > y > z$ and $x - y = 3 \times (y - z)$. So, for instance, we can assign world A value 10, world B value 4, and world C value 2. And, having set down these values, we fill in the rest of Averagism's value function and the rest of Totalism's value function in the usual way.

This proposal has some intuitive appeal, but it will not provide a general solution to the problem of intertheoretic value comparisons. First, there is no guarantee that there will always be even this minimal sort of background agreement among all of the moral theories to which the agent assigns some credence (Gustafsson and Torpman (2014)). Sepielli's approach to the problem of intertheoretic value comparisons will not work in these cases, and so MITE will not provide a fully general framework for decision-making under conditions of moral uncertainty.

Worse, there are cases in which Sepielli's proposal will lead to contradiction.[12] This problem can arise when theories agree on more than one ratio of value differences. Indeed, this will happen in the case of Averagism and Totalism. As noted, Averagism and Totalism agree about the ratio of value differences between A and B, and B and C. But they also agree about a lot of other ratios of value differences. Consider, for examples, worlds D, E, and F. World D contains two people, each with happiness level 10; world E contains two people, each with happiness level 4; and world F contains two people, each with happiness level 2. Averagism and Totalism agree that the degree to which

employs MITE for the purpose of arguing that moral theories that hold that there is little moral difference between the acts available to us should be treated as false for the purposes of deliberation, since having some credence in such theories will not affect which act has highest intertheoretic expectation. This result is impossible if the PEMT or modifications thereof are adopted.

[11] Ross (2006) briefly considers a proposal like this.

[12] I recently learned that Gustafsson and Torpman (2014) independently sketched this sort of problem.

D is better than E is three times the degree to which E is better than F.

Now, we cannot apply Sepielli's proposal both to A, B, and C and to D, E, and F without contradiction. Suppose that we start with A, B, and C. We'll set the values of A, B, and C as, say, 10, 4, and 2 (respectively) for both Averagism and Totalism. But then, Averagism and Totalism must differ in the values they assign to D, E, and F. Averagism must assign worlds D, E, and F values 10, 4, and 2 (respectively), while Totalism must assign D, E, and F values 20, 8, and 4 (respectively). Similarly, if we start by applying Sepielli's proposal to D, E, and F, Averagism and Totalism will agree on the values of D, E, and F but differ in the values they assign to A, B, and C. So, Sepielli's proposal leads to contradiction if we try to apply it both to A, B, and C and also to D, E, and F. More generally, contradiction threatens whenever moral theories agree about more than one ratio of value differences, for the constraints that result from applying Sepielli's proposal to one ratio of value differences may be incompatible with the constraints that result from applying it to a different one.

Finally, consider a proposal which explicates intertheoretic value comparisons in terms of their practical implications. This strategy is explicit in Ross (2006) and Riedener (2015), and also hinted at in Sepielli (unpublished). Ross (2006, 763) outlines the strategy thus:

> [W]e can explicate intertheoretic value comparisons in terms of claims about what choices would be rational assuming that the ethical theories in question had certain subjective probabilities. Thus, to say that the difference in value between ordering the veal cutlet and ordering the veggie wrap is one hundred times as great according to Singer's theory as it is according to the traditional moral theory is to say, among other things, that if one's credence were divided between these two theories, then it would be more rational to order the veggie wrap than the veal cutlet if and only if one's credence in Singer's theory exceeded .01.

But this proposal is circular, if MITE's ambition is to provide a framework which takes as input an agent's credences in moral theories (and credences about descriptive matters of fact) and outputs what the agent super-subjectively ought to do, without presupposing any facts about what agents super-subjectively ought to do in various situations (see also Gustafsson and Torpman (2014)). After all, Ross's proposal is to start with facts about what agents super-subjectively ought to do in certain cases and use those facts to reverse-engineer the desired intertheoretic value comparisons. But we could scale back MITE's ambitions. Instead of trying to use MITE to yield what agents super-subjectively ought to do given only their credences in moral theories, we could instead content ourselves with starting out with some facts about what agents super-subjectively ought to do in some circumstances (arrived at by some independent means, such as brute intuition) and then just using MITE to arrive at further facts about what agents super-subjectively ought to do in other circumstances. MITE could be thought of simply as a framework for imposing consistency on our judgments about what agents in different states of uncertainty super-subjectively ought to do. This is how many decision theorists think of expected utility theory, as simply requiring a certain coherence among your preferences and decisions.

If we scale back MITE's ambitions in this way, then Ross's observation does solve our problem. Riedener (MS) proves that if our judgments about what agents in various states of uncertainty super-subjectively ought to do obey certain decision-theoretic axioms, and if each moral theory's 'preferences' obey the same decision-theoretic axioms, then there is a choice of value functions to represent each moral theory such that an act $A$ is super-subjectively better than $B$ just in case the Intertheoretic Expectation (IE) of $A$ is higher than that of $B$, relative to the aforementioned choice of value functions.

I am unsatisfied. There is an analogy between Riedener's proof and Harsanyi's (1955) proposed solution to the problem of intertheoretic comparisons of utility. Harsanyi proves (with some supplemental assumptions, which I set aside) that if there are 'social preferences' that satisfy standard decision-theoretic axioms, and if each individual's preferences also satisfy those axioms, then there is a choice of individual utility functions such that the social preferences can be represented by a social utility function which is the weighed sum of those individual utility functions. Importantly, however, Harsanyi's theorem doesn't tell us how to pick an individual utility function to represent a given individual's preferences unless we already have the social utility function in hand. For this reason, Harsanyi's proposal leaves much to be desired. As one with Utilitarian sympathies (with utility understood as a representation of preferences), I would have liked to be told how to start off with individual's preferences and construct a social preference ordering therefrom, but I am instead told that if I start off with individual's preferences and a social preference ordering, then there is a way of fixing the zero point and scale of each individual's utility function such that social utility can be thought of as a weighted sum of individual utility. But I have no independent way of arriving at judgments about the social preference ordering. Insofar as I am a Utilitarian, I think that any facts about social betterness must be rooted in prior facts about individuals' preferences. I don't come up with judgments about social betterness through brute intuition, for instance.

Similarly, I might want to be told how to start off with my credences in moral theories and use them to derive a verdict on what I super-subjectively ought to do, but instead the Ross/Sepielli/Riedener approach tells me that if I start off with credences in moral theories *and* facts about the 'preferences' of the super-subjective *ought*, then there is a way of fixing the zero point and scale of each moral theory's value function such that the super-subjective *ought* can be thought of as mandating IE-maximization relative to those choices of zero points and scales. But I have little or no independent grip on (alleged) facts about super-subjective betterness. I, for one, have few if any brute intuitions about what agents super-subjectively ought to do in a various cases (with the possible exception of extreme cases, such as where one assigns all but a vanishingly small probability to one theory's being true). And while this is simply an autobiographical report, I suspect that most readers will likewise find themselves with few if any firm intuitions about what agents super-subjectively ought to do in various cases. Note that the case of ordinary decision theory is importantly different. Expected utility theory may just be a framework for imposing consistency on preferences, but it is still of some use since I come to the table with many preferences arrived at independently of thinking about expected utility theory.

In sum, if MITE is understood modestly, as a framework for imposing consistency on our judgments about the super-subjective *ought*, it is of little value unless we start off with at least some such judgments which are arrived at by independent means. But I am skeptical of whether we can or do arrive at such independent judgments about the super-subjective *ought*.

## 2.2 Options and Non-EVM-Representable Theories

Many moral theories cannot be represented in expected value maximization terms. For example, many moral theories hold that morality shouldn't be overly demanding. Morality gives us *options*.[13] According to these views, some actions are supererogatory, while others are merely permissible. For instance, giving a large proportion of one's time and money to charity is a wonderful thing to do, but it isn't required. After all, these theorists say, morality leaves us space to pursue our own goals and projects.[14]

Options are a challenge for MITE because on the face of it, they seem to say that one needn't always maximize value, whereas MITE requires all theories to be put in an expected-value maximization framework.[15] At first blush, these theories seem to differ from consequentialist theories not in their value theories, but in their decision rules. Some moral theories involving options, such as Slote's Satisficing Consequentialism (1984) are explicitly presented as differing from maximizing consequentialist views in employing a different decision rule.[16]

Now consider an agent who gives some credence to a moral theory which accepts a utilitarian axiology but gives the agent options. For instance, suppose that it says that while it's best to give as much money as possible to charity, one is only required to give away $1,000. How should defenders of MITE deal with this agent's state of uncertainty?

If we represent this options theory using a utilitarian value function and then plug it into MITE, then we effectively ignore the fact that the theory says that

[13] The term 'options' comes from Kagan (1989).

[14] See Williams (116-117, from Smart and Williams (1973)) for a famous defense of this claim, presented as an argument against Utilitarianism.

[15] See Sepielli (2010) for further interesting discussion of various issues regarding moral uncertainty and supererogation, although his focus is considerably different from ours.

[16] The problem of supererogation is discussed by Lockhart (2000, ch. 5), but he takes the strategy of arguing against moral theories involving supererogation. I am inclined to agree with him that the true moral theory, whatever it is, will not involve supererogation. But in the context of defending a framework for decision-making under moral uncertainty, this move is beside the point. As long as an agent could reasonably have some credence in a options or supererogation theory (even if such a theory is in fact false), then a theory of the super-subjective *ought* must be able to say something about that case.

there are options! We would be ignoring the distinction between this options-based utilitarian view and standard utilitarianism. In the extreme case in which the agent is certain of that options theory, MITE would say that she is super-subjectively obligated to maximize total happiness and give as much money as possible to charity. This is the wrong result. So clearly some added complexity is required if MITE is to deal with options theories.

We might try to somehow 'average' the relevant decision rules in aggregating her moral uncertainty. That is, in the case where an agent divides her credence between an options theory and a standard maximizing consequentialist theory, we might not only try to weight the different value functions by her credence in the corresponding theories, but also try to weight the different decision rules (maximization versus satisficing, for instance) by her corresponding credences. But it is doubtful whether any sense can be made of the notion of 'averaging' decision rules. What would it be, for instance, to average maximization with satisficing?

A more promising approach for the defender of MITE would be to draw a distinction between different senses in which a theory might be associated with a value function. Suppose our options theory $T$ makes various claims about which outcomes are better than others and by how much, and that these claims can be unified by representing $T$ as endorsing a value function (as is the case for our options theory which accepts the utilitarian axiology). Call this value function $T$'s *explicit value function.* We have already seen that options theories cannot be interpreted as requiring that agents maximize value according to their *explicit* value functions (else there would be no supererogatory acts, according to the theory). However, perhaps it will be possible to represent theories like $T$ as recommending that agents maximize expected value, so long as the value function whose expectation they're asked to maximize is not $T$'s explicit value function, but rather one reverse-engineered by looking at which actions $T$ recommends in which choice-situations. Call this sort of reverse-engineered value function an *implicit value function.*

Will such implicit value functions always exist? Sepielli (2009) and Ross (2006) both suggest that arguments ultimately inspired by Ramsey (1931) show that they will. Roughly, the idea is that Ramsey showed that if an agent's preferences satisfy certain axioms, then they can be represented with a value function. So for any moral theory, we can just imagine an agent who always prefers to act as the theory recommends, and then use Ramsey's method to construct the implicit value function of the theory, which can then be used together with the other theories the agent takes seriously to generate intertheoretic expectations for actions. (What Sepielli and Ross are appealing to here is known as a *Representation Theorem*, which says that if an agent has preferences which satisfy such-and-such axioms, then she can be represented as a agent who maximizes expected value, relative to some probability-utility function pair $< P, U >$ which is unique up to certain sorts of transformations, which differ depending on the axiom system in question. See von Neumann and Morgenstern (1944), Savage (1954), and Jeffrey (1983) for examples of Representation Theorems.)

But there are reasons to doubt whether we really can represent options theories using implicit value functions in this way. One main reason is that the options theory's preferences are likely to violate standard decision-theoretic axioms. In particular, the preferences of the options theory are likely to be *negatively intransitive.* That is, there will likely be acts A, B, and C such that neither of A and C is preferred to the other (in the sense that, given a choice between the two, neither is required), neither of B and C is preferred to the other, and yet A is preferred to B. For instance, let A be giving $1,000 to charity online (so that it arrives immediately), let B be giving $1,000 to charity by snail mail (so that it arrives after some delay), and let C be saving the money. Neither of A and C is preferred to the other, since both are permissible, and similarly for B and C. But A is preferred to B; if you're going to give to charity, you ought to choose the option that gets the money there more quickly if that requires no extra cost or effort. If the options theory has such negatively intransitive preferences, then it cannot be represented in EMV-maximization terms.[17]

Even setting this aside, it seems unlikely that any implicit value function assigned to an options theory would yield plausible results when plugged into MITE. For the implicit value function cannot assign the supererogatory act a higher expected moral value than the merely permissible one, for this would mean that in the limiting case where the agent is certain of that options theory, she would be *required* to perform the supererogatory act. And the implicit value function cannot assign the supererogatory and the merely permissible acts equal expected moral values, for then options theories can be easily swamped by other theories when we apply MITE to a morally uncertain agent. Consider an agent who gives some credence to an options theory which says that donating to charity is supererogatory while saving is merely permissible. The other theory

[17] This points merits some clarification. In an important and underappreciated paper, Oddie and Milne (1991) prove that in a certain sense of 'representation,' any moral theory whatsoever (subject to two constraints mentioned below) can be represented in EMV-maximization terms, relative to some agent-neutral value function. But their interpretation of what it is for a moral theory to be represented by another differs importantly from the interpretation that is relevant in the context of evaluating MITE. Oddie and Milne assume that each moral theory (i) has finitely many deontic categories (where deontic categories are things like supererogatoriness, obligatoriness, permissibility, wickedness, etc.), and (ii) that the moral theory gives a partial ordering of these deontic categories (supererogatoriness will be ranked higher than impermissibility, for instance). Then, they prove that for each such moral theory $M$, there is an agent-neutral value function $V$ such that, if act $A$'s deontic category is ranked at least as highly as act $B$'s according to M, then the expected value of $A$ is at least as great as the expected moral value of $B$, relative to value function $V$. But importantly, as Carlson (1995) notes, Oddie and Milne allow one moral theory to count as representing another even if the former does not even contain the same deontic categories as another. This is relevant because expected value theory as standardly interpreted employs just two deontic categories - permissibility (corresponding to having maximal expected value) and impermissibility (corresponding to having sub-maximal expected value). So on Oddie and Milne's criterion of representation, a theory on which $A$ is supererogatory and $B$ is merely permissible is adequately represented by a value function which assigns greater value to $A$ than to $B$ and hence deems $A$ to be obligatory and $B$ to be impermissible. This may be fine for some purposes. But in the context of MITE, it is unacceptable, for it does not enable us to respect the original moral theory's distinction between the supererogatory and the merely permissible. In effect, squeezing the supererogation theory into the EMV-maximization framework needed for MITE obliterates distinctions that the theory deems to be of fundamental importance.

to which the agent assigns some credence is a mild egoist theory that says that saving is slightly better than donating. For the options theory, on the proposal under consideration, donating and saving have the same expected moral value. For the mild egoist theory, saving has a slightly higher expected moral value than donating. Applying MITE, we will get the result that the agent ought to save her money no matter what (non-zero, real-valued) credence she assigns to each theory. This is an implausible result. Even if the agent is overwhelmingly confident that donating is supererogatory and saving merely permissible, a tiny degree of confidence that saving is required will tip the balance in favor of saving, so long as we represent options theories as assigning supererogatory and merely permissible acts the same expected moral value.

I am skeptical that there is any satisfactory way to squeeze options theories into MITE's expected value maximization framework.[18] But even if I am wrong about the case of options, it is overwhelmingly likely that very many moral theories that are worth taking seriously will be unable to be squeezed into this framework. They will have 'preferences' that fail to satisfy the axioms of the relevant Representation Theorem (see MacAskill (2014)). Just to take one possible example, an absolutist moral theory, on which some acts (murder, say) are absolutely prohibited, might have 'preferences' which fail to satisfy the Continuity axiom of Von Neumann and Morgenstern's decision theory. Suppose that our absolutist moral theory says that murdering one person (M) is worse than the status quo (S), which is worse than rescuing one person (R). That is, $M < S < R$. Moreover, murdering is absolutely prohibited, which on this theory means that if you're uncertain whether some act would result in murdering someone or saving someone, it's wrong to do it. In particular, for any probability $p$, an act with probability $p$ of resulting in M and probability $1-p$ of resulting in R is worse than the status quo $S$. This violates the Continuity axiom[19], which says:

> *Continuity*: If $A \leq B \leq C$, then there exists some positive probability $p$ such that:
>
> $(p)A + (1-p)C \sim B$ (where $\sim$ is the relation of indifference)

So, an absolutist moral theory on which it is impermissible to run any risk at all of murdering someone, even for the sake of having a chance of rescuing someone will have 'preferences' which violate one of the standard axioms of decision theory. As a result, that absolutist moral theory's verdicts will not be representable by a value function.[20]

[18] Recently, Ben West suggested to me that it may be possible to represent options theories in EU-maximization terms using vector-valued value functions. I will not pursue this strategy here.

[19] An absolutist moral theory would also likely violate the Archimedean axiom adopted by many decision theories, which in effect says that no options are infinitely good or infinitely bad. See Sepielli (2009) and Smith and Jackson (2006) for further discussion of absolutist moral theories in the context of decision-making under moral (Sepielli) and descriptive (Smith and Jackson) uncertainty.

[20] One might simply reject the Continuity axiom (and the Archimedean axiom) and assign

Even if options-based moral theories, absolutist moral theories, and others whose preferences cannot be represented by a value function are ultimately false, it seems that insofar as any moral uncertainty at all is rationally permissible, it should be rationally permissible to assign some positive credence to one of these problematic types of moral theory. If so, then there are moral theories which it can be rational to take seriously and which are such that if you do take them seriously, MITE cannot say anything about what you super-subjectively ought to do.

There is a general lesson here. MITE, and probably any plausible theory of the super-subjective *ought*, requires that the different moral theories in which an agent has some credence be translated into a common currency so as to allow them to be weighed up against each other.[21] But moral theories differ radically, and often in deep, structural ways. There is no reason to think that all respectable moral theories, from consequentialism, to Kantianism, to absolutist theories, to Ross-style pluralist theories (perhaps involving incommensurability, or Chang's (1997) 'parity'), to virtue ethical theories, will all be amenable to being squeezed into a common framework, whether that common framework is an expectational decision-theoretic one, or something else entirely. This doesn't necessarily mean that not all moral theories can be put into some decision-theoretic framework or other, but it is important to be careful about quantifier scope. It may be that, for each moral theory, there is some formal decision-theoretic framework that can (in some sense) represent it,[22]) but I am deeply skeptical that there will be some formal decision-theoretic framework that can be used to represent each moral theory. Instead, different departures from orthodox expected value theory (the system of Savage (1954), say) will be needed for different moral theories; some may require infinite values, others may require sets of value functions, still others may require a non-maximizing rule, and so on. For some purposes, like coming up with a way to think about how that

absolutely prohibited actions a negative infinite value. But then absolutist moral theories will swamp non-absolutist theories. It may be possible to attempt to avoid this swamping by representing Absolutist theories using a variety of technical devices, such as context-dependent value functions which, in any context, always assign values in such a way as to prohibit the absolutely prohibited action (Sepielli (2010)). Or perhaps surreal numbers will be of help (see Hàjek (2003) for discussion of surreal numbers in the context of decision theory). This technical moves may help the defender of MITE avoid uncomfortable conclusions when faces with absolutist theories - e.g., that if you given any credence to an absolutist moral theory, it will swamp all other theories to which you give some credence in virtue of its involving infinite values and disvalues. But it is difficult to see how one would motivate a particular choice among the various technical devices that might be wheeled in to help deal with absolutist theories, and yet different choices will yield different recommendations from MITE in various situations. At any rate, the present point is simply that many moral theories would seem, on the face of it, to violate standard decision-theoretic axioms needed to get representation theorems off the ground.

[21] An exception is the view that one super-subjectively ought to take the theory in which one has highest credence, and then simply act on its basis. See Gracely (1996) and Gustafsson and Torpman (2014) for a defense of this approach. Unfortunately I do not have the space to argue against it here.

[22] See footnote 5 above for references to discussions of attempts to find decision-theoretic representations of various moral theories.

theory should say you ought to act under descriptive uncertainty, it may only be important that each moral theory be representable in some formal decision-theoretic framework or other. But for the purpose of coming up with a formal framework for decision-making under moral uncertainty, it is crucial that each moral theory be representable in the same formal decision-theoretic framework (or common currency, as I put it earlier). And this, I am arguing, is not the case.

# 3 Whither the Super-Subjective Ought?

In the previous two sections, I have argued that MITE is unlikely to succeed as a theory of what a morally uncertain agent super-subjectively ought to do. If my arguments are sound, what does that mean for the super-subjective *ought*? I see three possibilities. First, perhaps we just need to pull up our socks and continue the hard work of trying to devise an adequate decision theory for the super-subjective *ought*. This strikes me as unattractive. The problems I have raised seem like in-principle problems, not likely to be solved through technical subleties.

Second, we might hold that there are facts about what one super-subjectively ought to do in most, or perhaps all, possible situations, but that these facts cannot be encapsulated in any formal or otherwise finitely statable theory. Perhaps there is little to be said by way of exceptionless principles, save for extreme cases (e.g., that if you are certain that $A$ is not morally worse than $B$, and not certain that $B$ is not worse than $A$, then you super-subjectively ought to do $A$).[23] This would amount to a sort of particularism about the super-subjective *ought*.

I have no compelling argument against this second option, but I want to explore a third, perhaps more radical, response. I want to suggest that perhaps there is no need to come up with a theory of the super-subjective *ought*, for the super-subjective *ought* has no clear role to play in our normative theoriz-

[23] Note that one who adopts the third option I consider (below), which denies the existence of a super-subjective *ought*, can still hold that there is something wrong with someone who is certain that $A$ is better than $B$ but then goes on to do $B$. But the explanation of what is wrong with that person will be different. If fundamental moral facts are *a priori*, then there is a sense in which one always ought to believe the true moral theory (though this this need not entail that one is blameworthy for having false moral beliefs, as Harman (2011) holds; the sense of *ought* may be purely epistemic, for instance). Then, if the true moral theory is one on which $A$ is better than $B$, our imagined agent is criticizable for simply for acting wrongly, while if the true moral theory is one on which the $A$ is not better than $B$, then our imagined agent is criticizable for having a moral belief that she ought not have. So in essence, we can account for what's wrong with an akratic agent by appealing (perhaps among other things) to a wide-scope norm stating that one ought to be such that if one believes one ought to do $A$, then one does $A$. But there are multiple ways ot satisfy such a wide-scope norm. One can make the antecedent of the embedded conditional false, or one can make the consequent true. In my view, if the moral belief refered to in the antecedent is false, then one ought to make the antecedent false (i.e. not have the false moral belief), while if that moral belief is true, then one ought to make the consequent true (i.e. perform the action that is in fact morally required).

ing. This discussion will be regrettably brief and speculative. I cannot show conclusively that the super-subjective *ought* has not role to be play in our theorizing. Instead, I proceed by looking at three main motivations for introducing the subjective *ought* to supplement the objective one, and then showing how we might resist the thought that these motivations carry over to motivate the introduction of a super-subjective *ought*. This discussion will clarify what kinds of commitments will likely have to be take on board by someone who wishes to adopt this third, more deflationist, response.

Start by recapping three interrelated motivations for bringing in the subjective *ought*. First, what you objectively ought to do often depends on factors inaccessible to you. You might be in no position to know that the pills in your bottle are rat poison, and you justifiably take them to be painkillers. In this case, even though you objectively ought not given them to your friend, you are not in a position to know that you ought not do so. Second, and relatedly, the objective *ought* is insufficiently action-guiding. It does not give advice to the deliberating agent that she can effectively use to determine what to do. Third, non-culpable ignorance of the facts which determine what you objectively ought to do is typically an excusing factor. Suppose you give your friend the pills, and after taking them he writhes around on the floor foaming at the mouth, and then dies. While you helped cause his death, you are not blameworthy for it, since you were justifiably ignorant of the fact that the pills were rat poison. On the basis of these considerations, we then introduce the subjective *ought*, which is intended to (i) be such that what you subjectively ought to do doesn't depend on things inaccessible to you, (ii) is action-guiding, and (iii) links up more closely with blame- and praiseworthiness than does the objective *ought*. (It is not clear that (i) should be regarded as a separate motivation, since it may be that the only grounds for wanting an *ought* which is always accessible to you is that accessibility is required for action-guidingness and blameworthiness.) The subjective *ought* is supposed to satisfy these demands by making what you ought to do depend on your credences in the relevant descriptive propositions, rather than on which of the relevant descriptive propositions are in fact true.

Now, there are serious questions about whether the subjective *ought* really can satisfy these demands, especially in light of Williamson's (2000) Anti-Luminosity Argument. If Williamson is right, then there are *no* conditions that are such that whenever they obtain, you are in a position to know that they obtain. Even the facts about your own doxastic state that determine what you subjectively ought to do may be inaccessible to you. And in a case where you are not in a position to know what your own beliefs or credences are, the subjective *ought* may not be fully action-guiding, and your self-ignorance might excuse you from any blame stemming from your failure to do what you subjectively ought to do. But set these issues aside. After all, my aim is not to defend the subjective *ought* but to oppose the super-subjective *ought*. What I now want to do is suggest that these considerations - accessibility, action-guidingness, and links with blame- and praiseworthiness - might not carry over to motivate the introduction of a super-subjective *ought* to supplement the objective and subjective ones.

First, even if descriptive facts may often be inaccessible to you, it is not clear that normative facts are likewise inaccessible. If fundamental moral truths are *a priori*, then there is a sense in which any agent is in a position to know the moral truth. There is no in-principle obstacle to her coming to know the moral facts. Moreover, your evidence (whatever it is), will entail each of the fundamental moral truths By contrast, your evidence will often not entail, or even support, the true descriptive propositions that are relevant in a given decision situation. Of course, even if the fundamental moral truths are *a priori*, this does not mean that they are obvious. But it is not clear that we should demand a sense of *ought* on which what you ought to do depends only on factors that are obvious as opposed to merely knowable in some weaker sense.

Admittedly, I have not argued that in fact fundamental moral truths are *a priori*. While I find this claim plausible (after all, the sorts of considerations typically given for or against particular moral theories tend to be of the *a priori* variety), it is certainly open to dispute. Some theorists might doubt that fundamental moral truths are even necessary (and it's unlikely that they would be contingent *a priori*), while others might hold that they are necessary *a posteriori*, in which case fundamental moral truths might be no more accessible than descriptive necessary *a posteriori* truths like the proposition that Hesperus is Phosphorus. Nevertheless, those theorists sympathetic to an *a priori* conception of ethics should hold that fundamental moral truths are unlike even very unobvious descriptive truths in being in-principle accessible.

Second, consider the morally uncertain agent's felt need for some sort of guidance. Certainly, such an agent will wish she knew what morality demands of her, and she will often have reason to deliberate further (though if she must act now, she may need to simply make a decision and defer deliberation until later). But reasons to deliberate further may be ordinary, garden variety epistemic and moral reasons. We have epistemic reasons to deliberate about matters of great importance in our lives. And the true moral theory $T$, whatever it is, will often want the agent to deliberate further about morality, since deliberating (insofar as it is reliable) will lead her to beliefs which better approximate $T$, and (insofar as her motivational state is sensitive to her moral beliefs) this will lead her to act in accordance with $T$ more often.

So a theory of the super-subjective *ought* is not needed to account for why uncertain agents often ought to continue deliberating about morality (and indeed, it gives no special role to deliberation anyway). The super-subjective *ought* really aims to earn its keep by giving agents guidance about how to hedge their bets, morally speaking. That is, it tells them how to act so as to minimize their expected degree of wrongness. But there is a case to be made that a desire for guidance about how to engage in moral hedging involves an objectionable sort of moral fetishism, so that a morally good agent would not look to a theory of the super-subjective *ought* like MITE to guide her actions in the first place.

Michael Smith (1994, 75) distinguishes between caring about morality *de dicto* and caring about morality *de re*:

> Good people care non-derivatively about honesty, the weal and woe

> of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like, not just one thing: doing what they believe to be right, where this is read *de dicto* and not *de re*. Indeed, commonsense tells us that being so motivated is a fetish or moral vice, not the one and only moral virtue.

Now, Smith uses the allegedly fetishistic character of *de dicto* concern for morality to argue against judgment externalism, the view that it is possible to judge that an action is morally required without being in any way motivated to perform that action. The details of Smith's anti-externalist argument needn't occupy us here, since the internalist/externalist debate is not our topic, and in any event I am persuaded by criticisms of Smith's argument by Shafer-Landau (1998), and Svavarsdóttir (1999), and especially Dreier (2000).[24] But Harman (2011) and Weatherson (2013) have recently raised this moral fetishism objection against theories on which an agent's moral beliefs affect how she ought to act. It is easy to overstate the case, however (and I suspect that Harman, at least, has). An agent who feels the need to deliberate further about some moral matter needn't always be fetishistic. This is especially clear where the agent's deliberation concerns thick moral concepts like *fairness* or *respect*, rather than thin ones like *wrongness* or *permissibility*.[25] And we do want agent's motivational states to somehow be sensitive to their beliefs about morality; else what is the point in debating moral matters? (Indeed, see Dreier (op cit) for discussion of how to explain why good, well-motivated agent's motivations are sensitive to their moral beliefs without attributing to them *de dicto* concern for morality.)

So my narrow, and hopefully more cautious, claim is just that the kind of motivation involved in moral hedging is objectionably fetishistic, even if a felt need to deliberate further, and a general sensitivity of one's motivational state to one's beliefs about morality, are not. But reasons to deliberate further, or to have a motivational state that is responsive to beliefs about morality, can be accounted for without positing a super-subjective *ought*.

Third, and finally, while it is quite clear that (non-culpable) ignorance of relevant descriptive facts often excuses you from blame, it is rather controversial whether (non-culpable) ignorance of fundamental moral facts likewise exculpates. Harman (2011) has recently argued that it does not.[26] She argues that it is possible to come to have deeply false moral beliefs without having been epistemically irresponsible in any way (unless failure to know *a priori* facts itself constitutes epistemic irresponsibility), but that in such cases an agent who acts on those false moral beliefs still strikes us as blameworthy. As just one example, she considers people who protest at abortion clinics and yell at the women and

[24] See also Lillehammer (1997) for an argument that *de dicto* concern for morality needn't be fetishistic in the first place. By contrast, Dreier agrees that *de dicto* concern for morality is objectionably fetishistic but argues that we can explain the fact that good, strong-willed agents are motivated to act in accordance with their moral beliefs without attributing to them such *de dicto* concern for morality.

[25] This point is emphasized by Sepielli (unpublished).

[26] See Zimmerman (1997) and Rosen (2004) for defenses of the opposing view.

doctors going inside. Assume that abortion is morally permissible and that it 'is wrong to yell at women outside abortion clinics: these women are already having a hard time and making their difficult decision more psychologically painful is wrong' (458). While perhaps many of these particularly strident protesters have been epistemically irresponsible in coming to their beliefs, it is not plausible to think that this is the case for all of them. But nevertheless, these protesters are blameworthy for the distress they cause. Arpaly (2003), using Smith's *de dicto/de re* distinction, argues for the same conclusion, writing that:

> An action is blameworthy just in case the action resulted from the agent's caring inadequately about what is morally significant - where this is not a matter of *de dicto* caring about morality but *de re* caring about what is in fact morally significant.

Now, it is clear that we are not inclined to excuse Hitler, say, from blame simply on account of his erroneous moral beliefs (if indeed he believed he was acting rightly). But matters are less clear, and so the stance of Harman and Arpaly is less compelling, in cases where the stakes are smaller or where the moral ignorance or error is less egregious. Vegetarians typically do not have strong negative reactive attitudes when their friends or colleagues eat meat. But this may not be because the carnivores do not merit blame, but rather because we are generally disinclined to hold others to a higher standard than we hold ourselves. Often when we judge that someone acting wrongly, we do not blame them if we could easily see ourselves acting in that manner. And it is coherent to judge that someone is blameworthy despite not actually having a strong negative reactive attitude toward that person.[27] So the fact that an agent's false moral beliefs may sometimes make us disinclined to blame her, not because moral ignorance is itself exculpatory, but rather because we could easily see ourselves being in her situation.

Now, even if you are not convinced, and believe that (non-culpable) moral ignorance *is* exculpatory (whether always, often, or just sometimes), we can still resist the thought that this means there must be an *ought* that is sensitive to moral uncertainty. For it is possible for some factor to be exculpatory without there being an *ought* that is specially sensitive to that factor. For instance, if an agent commits a violent act, the fact that he had a brutal, abusive upbringing can excuse him from blame, or at least mitigate his blameworthiness. But that does not mean that there is a special *ought* which is sensitive to the degree to which one's upbringing was normal. There is no sense in which he ought to have done as he did. Similarly, it may be that moral ignorance is exculpatory without there being a super-subjective *ought*.

[27] Consider a case not involving false moral beliefs, but rather akrasia. I believe that I and other reasonably well-off people are morally required to give very large portions of our wealth to the distant needy, but do not have strong negative reactive attitudes towards people who don't do so, even when they also believe they are so obligated. That's because I myself don't give tons of money away! The people who don't give generously to charity (myself included) are blameworthy, even though few actually experience an attitude of blame toward them (even those who are convinced about our obligations toward the distant needy).

This concludes my tentative argument that the introduction of the super-subjective *ought* is unmotivated. Admittedly, it is not a water-tight case. Perhaps one of the three motivations for introducing the subjective *ought* does carry over to the case of moral uncertainty. Perhaps there are other possible motivations for introducing a super-subjective *ought* besides the three that I have considered here.

But I hope at least to have done some softening-up work to suggest that before we become invested in solving the technical problems that face particular theories of decision-making under moral uncertainty, such as MITE, or accept a form of particularism about the super-subjective *ought*, we should get clearer about whether and why we wanted such a *ought* in the first place. Until a strong case is made that we need the super-subjective *ought* to play certain well-defined roles in our normative theorizing, we should be neither surprised nor worried when attempts to theorize about a super-subjective *ought* run into trouble. The default position should be that there are no rules for how to act in light of moral uncertainty; beliefs about descriptive matters make a difference to how you ought to act, while beliefs about moral matters do not. What you ought to do, in any moral sense of *ought*, depends on which moral theory is in fact true, not on your (possibly mistaken) beliefs about what morality requires.[28]

**References**

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency.* New York: Oxford University Press.

Carlson, E. 1995. *Consequentialism Reconsidered.* Dordrecht: Springer.

Chang, R. 1997. 'Introduction.' In R. Chang (ed), *Incommensurability, Incomparability, and Practical Reasoning.* Cambridge: Harvard University Press.

Colyvan, M., Cox, D., and Steele, K. 2010. 'Modelling the Moral Dimensions of Decisions.' *Noûs* 44, 503-29.

Dreier, J. 1993. 'Structures of Normative Theories.' *The Monish* 76, 22-40.

Gracely, E. 1996. 'On the Noncomparability of Judgments Made by Different Ethical Theories.' *Metaphilosophy* 27, 327-332.

Gustafsson, J. and Torpman, O. 2014. 'In Defence of My Favourite Theory.' *Pacific Philosophical Quarterly* 95, 159-74.

Hàjek, A. 2003. 'Waging War on Pascal's Wager.' *Philosophical Review* 112, 27-56.

[28] Thanks to Caspar Hare, Stefan Riedener, and especially Daniel Greco, who collaborated on an ancestor of this paper. Thanks also to audiences at the 2010 Australasian Association of Philosophy Conference, Oxford University, and the 2014 Wisconsin Metaethics Workshop. This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

Harman, E. 2011. 'Does Moral Ignorance Exculpate?' *Ratio* 24, 443-68.

Harsanyi, J. 1955. 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility.' *Journal of Political Economy* 63, 309-21.

Hudson, J. 1989. 'Subjectivization in Ethics.' *Americal Philosophical Quarterly* 26, 221-9.

Jeffrey, R. 1983. *The Logic of Decision.* Chicago: University of Chicago Press.

Kagan, S. 1989. *The Limits of Morality.* New York: Oxford University Press.

—2012. *The Geometry of Desert.* New York: Oxford University Press.

Lockhart, T. 2000. *Moral Uncertainty and its Consequences.* New York: Oxford University Press.

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

Oddie, G. and Milne, P. 1991. 'Act and Value: Expectation and the Representability of Moral Theories.' *Theoria* 57, 42-76.

Parfit, D. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

Portmore, D. 2011. *Commonsense Consequentialism.* New York: Oxford University Press.

Ramsey, F. 1931. 'Truth and Probability.' In his *The Foundations of Mathematics and other Logical Essays.* New York: Routledge.

Riedener, S. 2015. *Maximizing Expected Value under Axiological Uncertainty.* Ph.D. thesis, University of Oxford.

Rosen. G. 2004. 'Skepticism about Moral Responsibility.' *Philosophical Perspectives* 18, 295-313.

Ross, J. 2006. 'Rejecting Ethical Deflationism.' *Ethics* 116, 742-68.

Savage, L. 1954. *The Foundations of Statistics.* New York: John Wiley and Sons.

Sen, A. 1982. 'Rights and Agency.' *Philosophy and Public Affairs* 11: 3-39.

Sepielli, A. Unpublished. 'Normative Uncertainty, Intertheoretic Comparisons, and Conceptual Role.' University of Toronto.

—2009. 'What to Do When You Don't Know What to Do.' *Oxford Studies in Metaethics* 4, 5-28.

—2010. *Along an Imperfectly-Lighted Path: Practical Rationality and Normative Uncertainty.* Ph.D. thesis, Rutgers University.

—2012. 'Normative Uncertainty for Non-Cognitivists.' *Philosophical Studies* 160, 191-207.

—2013. 'Moral Uncertainty and the Principle of Equity Among Moral Theories.' *Philosophy and Phenomenological Research* 86, 580-9.

Slote, M. 1984. 'Satisficing Consequentialism.' *Proceedings of the Aristotelian Society, Supplementary Volumes* 58, 139-63.

Smart, J. and Williams, B. 1973. *Utilitarianism: For and Against.* Cambridge: Cambridge University Press.

Smith, M. 1994. *The Moral Problem.* Oxford: Blackwell.

—2009. 'Kinds of Consequentialism.' In E. Sosa and E. Villanueva (eds), *Philosophical Issues: Metaethics.* New York: Wiley-Blackwell.

Smith, M. and Jackson, F. 2006. 'Absolutist Moral Theories and Uncertainty.' *Journal of Philosophy* 103, 267-83.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior.* Princeton: Princeton University Press.

Weatherson, B. 2013. 'Running Risks Morally.' *Philosophical Studies* 167, 1-23.

Williamson, T. 2000. *Knowledge and its Limits.* Oxford: Oxford University Press.

Zimmerman, M. 1997. 'Moral Responsibility and Ignorance.' *Ethics* 107, 410-26.