Normative Uncertainty without Theories

Jennifer Rose Carr

How should an agent act under normative uncertainty? We might extend the orthodox theory of rational choice to the case of uncertainty between competing normative theories. But this requires that the values assigned by different normative theories be comparable. This paper defends a strategy for avoiding the need for intertheoretic value comparisons: instead of comparing competing moral theories, I argue that values can be represented in terms of a *de dicto* specification of value. I provide a decision theory for *de dicto* values that generalises expected utility theory and compare the proposal with alternative strategies for avoiding the problem of intertheoretic comparisons.

Keywords: moral uncertainty, incomparability, moral hedging

A rational person can be uncertain about whether it's permissible to eat lobsters, to vote for a lesser evil political candidate, to push a racist president in front of a train to save the lives of five dogs, and so on. This uncertainty can be normative uncertainty: uncertainty not about descriptive facts (e.g., whether lobsters feel pain) but about normative facts.[1] What should an agent do when faced with normative uncertainty?

A natural thought: we should generalise our best theory for choices under descriptive uncertainty: expected utility theory. Many philosophers have pursued this thought, defending the

[1] I use 'normative facts' in a minimalist, realist- and antirealist-friendly sense. It's an open question whether normative antirealism is compatible with rational normative uncertainty. See Smith [2002]; Staffel [2019].

decision rule Maximise Expected Intertheoretic Utility [Lockhart 2000; Ross 2006; Sepielli 2009; MacAskill 2014]. But this decision rule faces a serious challenge: it requires the utilities assigned by different moral theories to be comparable. There must be a shared unit of value between theories that fundamentally disagree about the nature of value. Many have argued that no such shared unit of value exists and concluded that normative uncertainty is irrelevant to morally appropriate action [Hudson 1989; Gracely 1996; Hedden 2012].

This paper offers an account of how normative uncertainty can affect appropriate action without reliance on intertheoretic utility comparisons. Section 1 introduces the orthodox decision rule for action under descriptive uncertainty: Maximise Expected Utility. I then characterise normative uncertainty and its most widely defended 'metanormative' decision theory, Maximise Expected Intertheoretic Utility, which requires that each moral theory be representable by a utility function. Section 2 explains the problem of intertheoretic utility comparisons with attention to the analogy with interpersonal utility comparisons.

Section 3 proposes a strategy for avoiding the problem of intertheoretic utility comparisons, using what I call '*de dicto* utilities'. Instead of assigning utility functions to different moral theories and then trying to determine how they compare, we set aside moral theories and focus on hypotheses about a utility function specified *de dicto*: *the actual moral utility function* (notated 'u'), whichever it is. I provide a metanormative decision rule for *de dicto* utilities (Maximise Expected Moral Utility) and explain the substantive differences between intertheoretic utilities and *de dicto* utilities. Section 4 addresses some worries about the proposal. Section 5 compares it with alternative proposals for avoiding the problem of intertheoretic utility comparisons.

1 Background

1.1 Normative uncertainty

The orthodox decision theory for the subjective *ought* is expected utility theory. A decision problem is represented as an ordered quadruple $\langle \mathcal{S}, \mathcal{A}, c, u \rangle$. $\mathcal{S}$ is a set of (mutually exclusive, exhaustive) possible states of the world, which determine the outcomes of an agent's acts. $\mathcal{A}$ is a set of propositions characterising mutually exclusive acts available to the agent. c is the agent's probability function, defined over $\mathcal{S}$, and u is the agent's utility function, defined over a space of outcomes $\mathcal{O}$. $\mathcal{O}$ is the set of possible conjunctions of states and acts, of the form "$s \wedge a$", where $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Let $c(s \mid\mid a)$ be the probability of s given that the agent performs a. Each $a \in \mathcal{A}$ is assigned an expected utility:

$$Eu(a) = \sum_{s \in \mathcal{S}} c\,(s \mid\mid a) u(s \wedge a)$$

Expected utility theory requires agents to conform to the decision rule Maximise Expected Utility:

Maximise Expected Utility: Choose an act that maximises expected utility (***Eu***).

Alongside descriptive uncertainty, rational agents face normative uncertainty.

Go Vegan? Sally is uncertain about whether non-human animals have moral standing. She is certain that if animals don't have moral standing, then it's a little better for her to eat meat, eggs, and dairy occasionally for gustatory and social reasons. She's also certain that if animals do have moral standing, then it's *badly* morally wrong for her to eat non-vegan foods.

Sally's decision problem could be represented as follows (where columns represent competing moral theories, rows represent acts, cells represent outcomes, and contents of cells represent the goodness or badness of outcomes):

	animals have moral standing	*animals don't*
stay non-vegan	very, very bad	fine and tasty
go vegan	fine	fine

Should Sally go vegan? Arguably, the answer to this question is sensitive to Sally's confidence that animals have moral standing: unless she's highly confident that they don't, she ought to hedge and avoid eating them.

In what sense of 'ought'? An agent **supersubjectively ought**[2] to ϕ iff she subjectively ought to ϕ, given her state of descriptive and normative uncertainty. This *ought* is *super*subjective in that it's sensitive to more informational limitations than the subjective *ought*, which is sensitive only to descriptive uncertainty. The **metanormativist**[3] holds that rational doxastic attitudes toward normative propositions can impact how an agent morally ought to act. In other words, there is a normatively significant supersubjective *ought* that sometimes yields different prescriptions from the correct objective or subjective *ought.*

[2] I borrow this term from Hedden [2016].

[3] I borrow this term from MacAskill [2014]. Harman [2011] calls this position 'uncertaintism'.

Metanormativism is controversial, for a variety of reasons.[4] An important class of objections contend that an adequate metanormative decision theory—a decision theory for supersubjectively permissible action under normative uncertainty—is impossible. The aim of this paper is not to provide a positive argument for metanormativism, but to defend it against this class of objections.

1.2 Expected utility theory for normative uncertainty

I assume that each moral theory is representable by a utility function.[5] Let O be a set of alternative possible outcomes. Utility functions are used to represent assessments of utility

[4] See Harman [2011]; Weatherson [2014]; Hedden [2016].

[5] This is controversial. One might hold, e.g., that the prohibitions of deontological theories can only be *absolute* within decision theory if we represent prohibited acts as having negative infinite utility. Infinite utilities fit poorly in standard decision theory. (They violate the von Neumann and Morgenstern [1944] 'Continuity' axiom.) They also lead to counterintuitive recommendations: for example, that an act with any positive probability (no matter how low) of leading to a violation of a prohibition, and zero probability of leading to the fulfilment of an obligation, will have negative infinite expected utility, and can therefore not be preferable to acts that are known to be prohibited. Colyvan, Cox, and Steele [2010] argue that many deontological intuitions can be captured with utility functions that assign prohibited acts finite but very low utility.

Other worries stem from the possibility of theories that treat some pairs of acts as incomparable in value (e.g., pluralist theories that rank one act as more valuable than another along one dimension of value, but less valuable along another). Such theories also fit poorly within standard decision theory (violating the von Neumann and Morganstern 'Completeness' axiom). I discuss such theories separately in Carr [manuscript]. See Portmore [2007]; Brown [2011] for discussion of further worries about whether moral theories can be represented with utility functions. Many controversies can be sidestepped if we

(e.g., desirability, moral value, goodness, worthiness, etc.), where these assessments may be normative or non-normative, agent-relative or agent-neutral. The term 'utility' here is meant to cast a broad net, compatible with deontological and other non-utilitarian conceptions of moral goodness, value, rightness, and so on: whatever it is in virtue of which some actions are more choiceworthy than others. (Others use 'value' or 'choiceworthiness' for this purpose; I prefer 'utility' because I focus on the relation with interpersonal utility comparisons.) Insofar as a theory makes moral distinctions at all, I assume, there are utility functions that can represent it, even if these utility functions are not particularly fine-grained.

How ought an agent act under normative uncertainty? The default option in the literature is a generalisation of the decision rule:

Maximise Expected Intertheoretic Utility (MEIU): Choose an act that maximises expected intertheoretic utility ($\boldsymbol{Eu_I}$, defined below).

Versions of MEIU are defended by Lockhart [2000], Ross [2006], Sepielli [2009], and MacAskill [2014]. I'll characterise a neutral version below.

Let $\mathcal{T}$ be the set of total moral theories. Different moral theories may have different commitments about how an agent *subjectively* ought to act. Some will require maximising expected utility: perhaps causal, perhaps evidential. Some will treat descriptive uncertainty as morally irrelevant, so that the subjective *ought* according to t collapses into the objective *ought* according to t. And so on.

---

allow that a utility function need not represent every morally relevant feature of a theory; that some theories may be representable with multiple utility functions, and not all must be relevant to metanormative decision problems; and that which utility function is relevant in a metanormative decision context may be context-sensitive.

I assume that for all $t \in \mathcal{T}$, all available acts can be assigned real-valued representations of their *subjective moral utility*. I do not assume that the subjective value according to t of an act is always identical to the act's expected t-utility. Each total theory—comprising a first-order moral theory and a theory of how to act under rational descriptive uncertainty—is then representable with a **subjective utility function**. The utility functions at issue in this paper are all subjective.

For each $t \in \mathcal{T}$, call this function 'u_t'. For theories that treat rational descriptive uncertainty as morally irrelevant, the subjective utility function may be identical to the objective utility function. For theories according to which an agent t-subjectively ought to maximise causal expected utility, the t-subjective utility of an act may equal the expected causal t-objective utility of the act. Mutatis mutandis for evidential decision theory. And so on. Using subjective utility functions allows us to freely alternate between talk of the utility of outcomes and the utility of acts.

The expected intertheoretic utility of an act a can then be defined as:

$$Eu_I(a) \quad = \sum_{t \in \mathcal{T}} c\,(t \,||\, a) u_t(a)$$

MEIU requires maximising Eu_I. Return to the **Go Vegan?** example. We can fill in utilities for t_1, according to which animals have moral standing, and t_2, according to which they don't:

	t_1	t_2
stay non-vegan	-1000	3
go vegan	0	0

For each act, MEIU takes the probability-weighted average of the utility that each theory assigns to the act, and requires choosing the act that maximises this quantity. Given our

assignment of utilities, even if Sally is 99% confident that animals don't have moral standing, MEIU requires her to hedge her moral bets and go vegan.

2 The problem of intertheoretic utility comparisons

Let u and u' be distinct utility functions. Call a claim of either of the following forms a *utility comparison*:

$$u(x) \geq u'(y) \qquad \text{comparison of } \textit{levels}$$

$$u(x_0) - u(x_1) \geq u'(y_0) - u'(y_1) \qquad \text{comparison of } \textit{units}$$

MEIU requires maximising *intertheoretic* utility, and so presupposes that the utilities assigned by different theories are, at minimum, amenable to comparison of units.[6] Full comparability between theories requires that they fit together into an intertheoretic utility function, a universal utility scale that represents the correct commensuration of different theories. But if such a universal scale does not exist, there are no grounds for comparisons of levels or units of moral utility. *The problem of intertheoretic utility comparisons* is the worry that there are simply no facts in the empirical or normative world that could determine how different theories compare in how they value what they value.[7]

To explain this problem—arguably MEIU's greatest challenge—I begin by rehearsing analogous objections to inter*personal* utility comparisons.

[6] Comparisons of levels are unnecessary for MEIU. If we increase one theory's utilities by k, in inequalities of expected utilities for acts, k will cancel out.

[7] This problem was introduced by Hudson [1989].

2.1 The indeterminacy of interpersonal utility comparisons

In economics and philosophy, interpersonal utility comparisons are traditionally treated as indeterminate. For example, Robbins [1932] argues: 'Introspection does not enable *A* to measure what is going on in *B*'s mind, nor *B* to measure what is going on in *A*'s. There is no way of comparing the satisfactions of two different people' (140). Similarly, Arrow [1951] accepts the premise that 'interpersonal comparison of utilities has no meaning and, in fact, that there is no meaning relevant to welfare comparisons in the measurability of individual utility' (9).[8]

This orthodoxy stems from the view that interpersonal utility comparisons cannot have empirical significance. Why not? An agent's utilities must be manifestable in her preferences and choice behaviour. Agents, it's traditionally assumed, make choices that approximate maximising expected utility. But if Bob maximises expected utility according to utility function u, then he also maximises expected utility according to $u'(\cdot) = 10^{23}u(\cdot) + 198.39$. It makes no empirical difference to Alex's preferences or behaviour, or Bob's preferences or behaviour, if the numbers used to represent Bob's utilities are all millions of times greater than the numbers used to represent Alex's utilities (or all trillions of times lesser, or all multiplied by 5, or…). The same acts will maximise expected utility for Alex however we represent Bob, and vice versa.

Traditionally, any utility function u is treated as *informationally equivalent*—equivalent for the purposes of representing the relevant desires or values—to any other utility function u' where

[8] The orthodoxy isn't universal: Harsanyi [1977] defends interpersonal comparisons using *extended preferences* (preferences between being in *A*'s position with *A*'s preferences vs. being in *B*'s position with *B*'s preferences), together with the (questionable) assumption that these extended preferences are universal. Sen [1970, 1979] argues that interpersonal comparisons are necessary for social choice theory and surveys options for formal representations of comparability.

there is a positive affine transformation θ such that $u'(x) = \theta\big(u(x)\big)$ for any x in the domain of u. A *positive affine transformation* is a function $\theta\colon \mathbb{R} \to \mathbb{R}$ such that there exists an $a \in \mathbb{R}_{>0}$ and a $b \in \mathbb{R}$ where, for any $r \in \mathbb{R}$, $\theta(r) = ar + b$.

So on the traditional picture, an agent's utilities are represented with an interval scale. Its zero point doesn't represent anything about the agent's desires (e.g., the boundary between what the agent values and disvalues). Nor do ratios of the utilities it assigns to different outcomes. To assign one outcome utility 32 and another utility 64 is not to represent the first as half as good as the second. (Compare: $32°\text{F}$ is not half as warm as $64°\text{F}$, nor is $0°\text{C}$ $(= 32°\text{F})$ zero eighteenths as warm as $18°\text{C}$ $(\approx 64°\text{F})$.)

An agent's desires can therefore be represented with any of an uncountable set of informationally equivalent utility functions. For each utility function u', call the set of utility functions to which u' is informationally equivalent "$[u']$."[9] Each $[u']$ forms an equivalence class. The information that u' carries about its representational target is the information about which the elements of $[u']$ are unanimous. For example, given our assumptions about informational equivalence, we can expect all u in $[u']$ to be unanimous about whether $u(x_1) > u(x_2)$, or that $\frac{u(x_1)-u(x_2)}{u(x_3)-u(x_4)} = r$. So from any claim about an agent's utility function u', the only information to be extracted about the agent's desires is information that also holds for every other element of $[u']$. That's all the information that's needed to rationalise the agent's behaviour.

Similarly for any claim about two agents' utility functions. We can learn something about Alex and Bob if we're told that both prefer x_1 to x_2 (namely, that for all $u \in [u^{alex}] \cap [u^{bob}]$, $u(x_1) > u(x_2)$). But we do not learn anything about Alex or Bob from the claim that Alex's

[9] For readability I use double quotes in place of Quine quotes.

utility function assigns a larger number to x_1 than does Bob's. The assumption of informational equivalence of utility functions up to positive affine transformation ensures that it's not the case that for all $u_i^{alex} \in [u^{alex}]$ and $u_j^{bob} \in [u^{bob}]$, $u_i^{alex}(x_1) \geq u_j^{bob}(x_1)$ or that $u_i^{alex}(x_1) - u_i^{alex}(x_2) \geq u_j^{bob}(x_1) - u_j^{bob}(x_2)$. Interpersonal utility comparisons are therefore meaningless.

2.2 The indeterminacy of intertheoretic utility comparisons

In order to apply certain decision rules (MEIU, along with related possible decision rules, e.g., a Buchakian risk-weighted alternative to MEIU [Buchak 2013]), the utility functions representing different moral theories must be comparable.

But as with interpersonal utility comparisons, there's arguably no way to ground intertheoretic utility comparisons. In the **Go Vegan?** case, if t_1 says that staying nonvegan is 1000 t_1-utiles worse than going vegan, and t_2 says that it's 3 t_2-utiles better, that does not entail that the moral stakes are higher for t_1, or even that eating animals is worse according to t_1 than it is eating animals according to t_2. There are strong reasons to reject the idea that moral utility functions carry more cardinal information, or enough cardinal information to underwrite the possibility of intertheoretic utility comparisons.

Most flat-footedly, there are no empirical or normative facts that determine the right scale for particular theories: why would some theory assign the outcome of, say, hugging one's mother utility 4 rather than 1.204×10^{93}? Why would the difference in utility assignments between hugging one's mother and eating poisoned tamales be 9323 rather than 52? As many philosophers have noticed, there are fundamental problems for the possibility of locating such comparisons. Hudson [1989] uses the example of comparing a theory that values only pleasure (measured in 'hedons') with a theory that values only self-realisation (measured in 'reals'):

> 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-realisation—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. Clearly this second-order hedging is impossible. (225)

If intertheoretic utility comparisons are impossible, this poses an existential challenge to metanormative decision theories that rely on utility comparisons across different moral theories (including MEIU and related theories). This in turn undermines the plausibility of a substantive supersubjective *ought*. MEIU is generally taken to be the most promising metanormative decision theory. Familiar decision theories that aren't impacted by the problem of intertheoretic utility comparisons are less attractive either for their limited scopes (e.g. theory-wise moral dominance principles) or for their counterintuitive results (e.g. the so-called 'My Favourite Theory').[10] If the most plausible, systematic decision theories for decisions under moral uncertainty are metaphysically guaranteed to fail, it would be reasonable to conclude that moral uncertainty is normatively inert. Many have argued for precisely this conclusion on the basis of the problem of intertheoretic value comparisons (e.g. Hudson [1989]; Gracely [1996]; Hedden [2016]).

Caveat #1: This paper doesn't take a stand on whether intertheoretic utility comparisons are in fact possible. The challenge may well be answerable. There is a case to be made that interpersonal utility comparisons are possible. Sen [1970], for example, notes that ordinary

[10] *My Favourite Theory* requires agents to conform to the moral theory in which their credence is highest. See [Lockhart 2000; MacAskill 2014] for objections; see [Gustafsson and Torpman 2014] for a defence of a sophisticated variant.

intuitions favour the idea that the stakes of a decision are sometimes higher for one person than another, which requires the possibility of interpersonal utility comparison. Sen [1970] and List [2003] explain possible conditions under which such comparisons may be determinate. Similarly, Ross [2006] and MacAskill [2014] note that ordinary intuitions favour the idea that in some cases, the stakes of a decision are higher according to one moral theory than another. So the problem of intertheoretic utility comparisons may simply show flaws in our assumptions about the representation of value in utility functions.

This paper doesn't hinge on whether objections to intertheoretic utility comparisons are conclusive. The claim here is that a substantive supersubjective *ought*, and the use of metanormative decision theories, do not depend on the possibility of intertheoretic utility comparisons. If such comparisons are impossible, I'll argue, a close analogue of MEIU and its relatives are still usable.

Caveat #2: Another problem sometimes goes under the name 'the problem of intertheoretic utility comparisons'. This is the problem of comparing 'cardinal theories' (theories that draw cardinal distinctions in the utilities they assign, such that these theories are not adequately representable by an ordinal ranking of outcomes or options) with 'merely ordinal theories' (theories which are adequately representable by an ordinal ranking).[11] I defend a solution to this problem separately in Carr [manuscript].[12]

[11] Like cardinal theories, merely ordinal theories may be representable with utility functions, but with different constraints about which utility functions are informationally equivalent. Suppose a merely ordinal theory that imposes a total preorder on outcomes is representable with utility function u. u will then be informationally equivalent to any utility function u' where there is a positive monotonic

3 Metanormative decision theory without theories

3.1 Utilities de dicto

In traditional decision theories for descriptive uncertainty, we typically imagine that the agent's uncertainty about the utility of their acts results from uncertainty about the outcomes their acts will bring about. We assume that agents are uncertain about what descriptive properties their acts might have: their monetary value, the number of lives lost, the quantity of happiness generated, etc. Given a maximal specification of these descriptive features, we assume that the agent will have a determinate utility assignment to the outcome.

But this way of grounding utilities in descriptive features of outcomes is inessential to decision theories. Decision theories don't care about *why* you assign the utilities you do to the worlds you do. They only require that your individuation of states is fine-grained enough to specify a utility for each possible outcome of your acts. (That is, the partition $\mathcal{O}$ of outcomes must be sufficiently fine-grained that for every $o \in \mathcal{O}$, for every $w, w' \in o$, $u(w) = u(w')$.) We can represent an act a with sequence of dummy outcomes $\langle o_1, \ldots, o_n \rangle$, that don't pick out any specific propositions. As long as we have a corresponding sequence of credences $\langle r_1, \ldots, r_n \rangle$ that sum to 1, such that $c(o_i \mid\mid a) = r_i$, and a sequence of utilities $\langle r_1', \ldots, r_n' \rangle$, such that $u(o_i) = r_i'$, then we can assign a an expected utility. It's not necessary to have any information about what in the world these dummy outcomes represent.

---

transformation θ such that $u'(x) = \theta(u(x))$. A *positive monotonic transformation* is a function $\theta: \mathbb{R} \to \mathbb{R}$ such that, for all $r_i, r_j \in \mathbb{R}$, if $r_i > r_j$, then $\theta(r_i) > \theta(r_j)$.

[12] That strategy is compatible with use of *de dicto* utilities, introduced below, but does not depend on it.

Indeed, it's plausible that agents must sometimes make decisions where their only conception of possible outcomes of their acts is in terms of the outcomes' utilities. Such cases can arise when an agent lacks the conceptual resources to entertain possible outcomes of her acts, or lacks the phenomenological information necessary to assign these outcomes utilities. For example: Paul [2014, 2015] argues that in deciding whether to become a parent, one cannot so much as entertain what it's like to have a child, any more than Jackson's [1982] Mary can entertain what it's like to see the colour red when she's spent her entire life in a black and white room. This forms the basis for Paul's argument that choosing whether to have a child cannot be a rational decision.

Pettigrew [2015] and Dougherty, Horowitz, and Sliwa [2015] have argued that in such cases, adequate dummy outcomes can be individuated solely by their utilities, rather than the empirical properties that help to determine those utilities. We can think of the act of having a child as having possible outcomes of the form: *I'll be in a state with utility 1; I'll be in a state with utility 24;* and so on. The agent's credences in these possible outcomes determine the expected utility of having a child.

It's debatable whether this strategy adequately addresses Paul's challenge, for reasons specific to the context of transformative experience.[13] But I'll show that an analogous strategy—using dummy outcomes specified purely in terms of utility assignments—works well in the context of normative uncertainty.

[13] Paul [2014: 128] argues that reasoning with dummy outcomes lacks *authenticity*, which requires 'choosing after assessing our preferences from our first-personal point of view and then living with the results'. Isaacs [2019] offers a variety of objections to Pettigrew's strategy, the most compelling of which don't apply in the context of moral uncertainty.

For this, we need a different representation of both the rational agent's state of uncertainty and the utility function at stake. Instead of using an intertheoretic utility function specified *de re*,[14] we use a utility function *de dicto* (under a description): the utility function determined by whichever moral theory is in fact correct, as determined by the moral truths of the actual world; the actual moral utility function, delivered by pure normative reality.

Let $\mathtt{u}$, in typewriter font, be shorthand for the definite description *the actual moral utility function* (relative to a selected scale). Read "$\mathtt{u}(a) = n$" as *the actual moral utility of a is n.* Let $\Lambda \subset \mathbb{R}$ be a set of real numbers that contains all of the moral utility assignments that an agent considers possible for a. (Assume for convenience that Λ is finite.) The expected moral utility of a is represented as follows:

$$E\mathtt{u}(a) = \sum_{\lambda \in \Lambda} \lambda c(\mathtt{u}(a) = \lambda \mid\mid a)$$

We can then construct decision theories that make use of $\mathtt{u}$: for example,

Maximise Expected Moral Utility (MEMU): Choose an act that maximises $E\mathtt{u}$.

The moral utility function $\mathtt{u}$ is not intertheoretic. It isn't generated by attributing utility functions to individual moral theories and then determining a conversion rate between them. It simply represents hypotheses about how objectively morally good or bad different available acts might be. These are hypotheses about possible features of the actual moral utility function. Because there are no moral theories being compared, no intertheoretic utility comparisons are needed.

[14] I use the terminology of *de re* and *de dicto* somewhat nonstandardly: a *de re* specification of a utility function is one that epistemically entails, for each o, the precise utility of o. So if a name for a utility function is specified descriptively (for example, if we take the description 'Sally's utility function' and assign the entity actually occupying that role the name 'Julius'), the name does not count as specifying a utility function *de re*.

The difference between MEIU and MEMU isn't that they make conflicting predictions. It's that on the assumption that intertheoretic utility comparisons are impossible, MEMU can still make predictions, whereas MEIU cannot. If intertheoretic utility comparisons *are* possible, then in cases where MEIU can make predictions, MEMU's predictions may entirely coincide with MEIU's, at least given reasonable constraints on the relation between credences in moral theories assigned and credences about the actual moral utilities.[15]

3.2 How de dicto utility hypotheses differ from intertheoretic comparisons

Objection. How is this not merely a redescription of intertheoretic utility maximisation? The state space of MEIU is the set of epistemically possible moral theories. The elements of the state space for MEMU must also specify, somehow or other, competing hypotheses about the moral utilities of all acts under consideration. So why not call these hypotheses 'moral theories'? And if that characterisation is accurate, then aren't they still susceptible to the problem of intertheoretic utility comparisons?

Reply. We've assumed that there are no meaningful comparisons of units or levels between different moral theories. But this doesn't entail that it's meaningless to compare hypothetical features of a utility function specified *de dicto*.

To see this, it's helpful to return to the problem of inter*personal* utility comparisons. It's been argued that there is no empirical significance to be assigned to interpersonal utility comparisons. By contrast, we can make empirically significant comparisons between scenarios in which an individual agent's utility function is other than how it actually, presently is: for example:

[15] Indeed, whether the two decision rules coincide in such cases might provide a test for whether an agent's credences in moral theories and credences in actual utilities are epistemically rational.

- *Cross-temporal comparisons:* It's an empirical fact that I used to assign lower utility to drinking coffee than I presently do.
- *Counterfactual comparisons:* If I assigned higher utility to drinking coffee than I in fact do, then (*ceteris paribus*) I would drink coffee more often.
- *Hypothetical indicative comparisons:* I can wonder about another agent's utility function (what's his utility for drinking coffee?), entertain different possible utility functions that the agent might have (does he assign drinking coffee higher utility than drinking horchata?), and receive empirical evidence that confirms some hypothesis over another (he ordered coffee rather than horchata).

Objection. Why aren't such comparisons ruled out by the same considerations that rule out interpersonal utility comparisons?

Reply. They had better not be, because it's clear that these comparisons are empirically significant! If some model of an agent's desires or values rules out the possibility of making such comparisons, that's evidence that the model is inadequate.[16]

Happily, that isn't the case. In these comparisons, we retain fixed points: we can compare my actual preferences to what my preferences would be in a scenario in which some of my utility function is held fixed, but the utility I assign to drinking coffee goes up. There are meaningful comparisons to be made between an agent's different possible utility functions even in cases where the vast majority of the agent's utility assignments change. All that's required for such comparison is the assumption of two fixed points: two outcomes with different utilities.

[16] See Briggs [2015] for discussion of both the importance of, and the challenges to, cross-temporal intrapersonal utility comparisons.

Fixed points can establish a basis for constructing utility functions that represent hypotheses about features of the actual moral utility function, u. Suppose we arbitrarily select two outcomes, x and y, such that the agent knows $\mathrm{u}(x) \neq \mathrm{u}(y)$, that are not at issue for the decision at hand. The agent may be uncertain about the moral utilities of x and y; she may even be uncertain which has higher utility. Then we can arbitrarily select $n, m \in \mathbb{R}$, $n > m$, and stipulate that whichever of x and y has higher utility will have utility n, and the other utility m. From these stipulations, we can construct different hypotheses about other outcomes' utilities: that $\mathrm{u}(z) = 2(n-m) = 2(\max(\mathrm{u}(x), \mathrm{u}(y)) - \min(\mathrm{u}(x), \mathrm{u}(y)))$, or that $\mathrm{u}(z) = \frac{1}{2}(n-m)$, or…

Notice: on this account, alternative hypotheses about u aren't selected from the space of utility functions and then scoured for points of comparability (as many characterisations of intertheoretic utility comparison assume: for example Ross [2006], discussed below). Instead, we choose features of u to fix, and then construct alternative hypotheses around these features.

4 Objections and replies

4.1 Arbitrary scale

Objection. What sense does it make to talk about 'the' moral utility function? Surely u has informational equivalents. No facts about the world, even objective normative facts, determine particular real number assignments to acts or states of affairs: nothing about Kartik's adopting a rescue dog ties the act to the number 12 rather than 713. But using *de dicto* utilities seems to require the presupposition that acts have a determinate quantitative level of moral utility. After all, it's a decision theory for choice under uncertainty about this very quantity.

Reply. Let's maintain the assumption that the actual moral utility function is unique only up to positive affine transformation. Choosing any specific utility function from $[\mathrm{u}]$ is arbitrary. Still, there are non-arbitrary moral facts. These are encoded in ratios of utility differences: for any x_1, x_2, x_3, x_4, there is a unique real number r such that for all $\mathrm{u}' \in [\mathrm{u}]$, $\frac{\mathrm{u}'(x_1)-\mathrm{u}'(x_2)}{\mathrm{u}'(x_3)-\mathrm{u}'(x_4)} = r$. Given our assumptions about informational equivalence, these quantities are genuinely determined by the normative facts, and they can serve as the objects of uncertainty for MEMU. I use the alternative representation only for ease of exposition. For any given decision problem, a conventional scale can be selected arbitrarily. We can then flesh out the description determining *de dicto* utilities relative to specific stipulations. We might use descriptions like "$\mathrm{u}(\cdot \mid \mathrm{u}(x) = n, \mathrm{u}(y) = m)$" as shorthand for the description: 'the actual moral utility of ·, on a scale where the utility of max(u(x), u(y)) is n and of min(u(x), u(y)) is m'.

4.2 What about genuine intertheoretic uncertainty?

Objection. Decision theories that make use of u, like MEMU, treat normative uncertainty as uncertainty about u. But there are other forms of normative uncertainty—in particular, uncertainty about which moral theory is correct. (We can, e.g., be uncertain of whether Singer's [1975] utilitarianism is correct.)

Reply. MEMU and its relatives don't rule out intertheoretic normative uncertainty. Rather, they say that intertheoretic normative uncertainty needn't be the form of uncertainty that's relevant to metanormative decision theory.

Objection. Still, these decision theories ignore intertheoretic normative uncertainty. But this form of uncertainty shouldn't be normatively inert.

Reply. Intertheoretic uncertainty can affect metanormative decision theory indirectly, through its effect on rational uncertainty about features of $\mathfrak{u}$. There are presumably epistemic coherence requirements connecting these two forms of uncertainty. Given that these requirements are more substantive than probabilistic coherence, this requirement goes beyond subjective bayesianism.[17] This leads to a more general worry:

4.3 Are these credences unconstrained?

Objection. On this proposal, what an agent supersubjectively ought to do is determined entirely by her credences in propositions about $\mathfrak{u}$. So with the right credences, any action at all could be warranted by MEMU. This makes MEMU too permissive.

Reply. This is a problem for effectively any normative theory that makes permissible action sensitive to doxastic states. I accept that what an agent subjectively and supersubjectively ought to do depends not on the agent's actual credences, but on the credences that are rational given her evidence.[18] So MEMU will be permissive only to the extent that our moral epistemology is permissive.

A popular permissive epistemic theory is *subjective bayesianism*, the thesis that a credence function is rational if it satisfies the probability axioms and is updated by conditionalisation on the agent's total evidence. Pairing MEMU with subjective bayesianism risks leaving metanormative decision theory too permissive. But the same is true about expected utility theory for purely descriptive uncertainty. Subjective bayesianism permits updating on evidence in intuitively irrational ways, which warrant intuitively irrational choices. Consider an update policy that, conditional on experiential evidence as of a busy highway in front of me, has me

[17] Thanks to Elizabeth Harman for this helpful objection.

[18] For discussion of this question, see Harman [2011].

assign credence .999 to the proposition that there's a field of wildflowers in front of me. This doesn't violate the laws of probability, and is therefore permissible according to subjective bayesianism. But this update warrants assigning maximal expected utility to the act of running forward with my kite (unwittingly into traffic).

In short: this objection poses a problem for subjective bayesianism rather than for MEMU.

Objection. If there are substantive epistemic constraints beyond probabilism and conditionalisation on credences about $\mathtt{u}$, it would be helpful to know what they are. What determines which credences in propositions about $\mathtt{u}$ are epistemically rational?

Reply. This is a question for moral epistemology, not moral decision theory. This division of labour is well-precedented. Epistemologists who aren't subjective bayesians haven't managed to specify general constraints that determine, for example, how many spotted treefrogs one would have to see in order to rationally have credence greater than .8 that all treefrogs are spotted. This doesn't cast doubt on the viability of expected utility theory.

5 Comparison to alternative accounts

Sections 3 and 4 showed that metanormative decision theory can proceed without intertheoretic utility comparisons. A variety of other strategies have been proposed that aim to show that intertheoretic utility comparisons are in fact possible. I'll discuss two clusters of such 'commensuration strategies': those that appeal to *intuitive* intertheoretic agreement and those that appeal to *structural* intertheoretic agreement.

5.1 Intuitive intertheoretic agreement

One form of commensuration strategy, defended in [Ross 2006], involves finding intertheoretic fixed points: pairs of cases where theories *intuitively* agree. The proposal runs as follows: take

two moral theories, t_1 and t_2. Look for two outcomes, x and y, such that it's known that t_1 and t_2 fully agree in their moral evaluations of x and y, and such that t_1 and t_2 assign x a greater utility than y. (Two theories fully agree in their moral evaluations of x and y just in case x and y don't differ in relation to any determinants of moral value about which the theories disagree.) We then define the difference in utility between x and y as one unit of utility for both theories. We select a $u'_{t_1} \in [u_{t_1}]$ and a $u'_{t_2} \in [u_{t_2}]$ such that $u'_{t_1}(x) - u'_{t_1}(y) = 1 = u'_{t_2}(x) - u'_{t_2}(y)$. These utility functions, Ross argues, are comparable.[19] u′

This strategy presupposes the existence of correct intertheoretic utility comparisons. After all, it requires there to be an independent fact in moral reality that determines that, whatever scale we choose, t_1 and t_2 agree about the difference in utility between x and y. (Indeed, it requires the existence of competing moral theories that are *fully* comparable with respect to utility units.) This is a substantive and controversial metaphysical commitment. One might simply reject all intertheoretic utility comparisons: trying to find a conversion rate across theories would be like trying to find a conversion rate between centimetres and degrees Fahrenheit. Decision theories that make use of *de dicto* utilities don't require the possibility of such comparisons, and are therefore less metaphysically committal.

Ross's commensuration strategy also requires *knowledge* of places where t_1 and t_2 agree, which places a stronger epistemic condition on the usability of metanormative decision theory than do *de dicto* utilities. The latter only require knowledge that two outcomes are morally unequal. This is not an entirely trivial epistemic requirement, but it's extremely weak: it's sufficient to know that the utility of eating a poisoned tamale is not equal to the utility of hugging your mom. (You don't even need to know which is better!)

[19] With respect to units, which is all that's necessary for MEIU.

Another form of commensuration strategy also appeals to intuitive intertheoretic agreement, but with attention to specific shared commitments about *contributory* values in competing theories. This strategy, defended in [Tarsney 2018], groups theories into 'comparability classes' according to local agreements about dimensions or loci of value:

> 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 recognise as nonderivatively valuable. 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 normalisation. (331)

Here, the epistemically possible monistic and pluralistic theories fall into a shared comparability class because of their agreement about the contributory value of hedonic goods.

Within comparability classes, utility comparisons are determinate and can figure into decision theories like MEIU. There may be no guarantee of utility comparisons across distinct comparability classes.[20]

There are cases where this commensuration strategy risks commitment to inconsistencies. Suppose there are three theories, t_1, t_2, and t_3, that each share some commitments with both of the other two theories with respect to three dimensions of contributory value: units of pleasure (*hedons*), units of aesthetic beauty (*aesthetons*), and units of self-realisation (*reals*). Each theory

[20] Tarsney's aim is to show that the problem of intertheoretic utility comparison does not block the possibility of a substantive supersubjective *ought* governed by metanormative decision theory. So he defends only the existential claim that there are cases where intertheoretic utility comparisons are possible.

is only committed to the value of two of these and denies the value of the third. Where two theories agree about some dimension of value, their agreement is total: they have the same claims about the nature, basis, and degrees of value of each.

	values	**total intuitive agreement**	**contributory value comparison**
t_1	hedons	with t_3 about hedons	utility of 1 hedon =
	aesthetons	with t_2 about aesthetons	utility of 1 aestheton
t_2	aesthetons	with t_1 about aesthetons	utility of 1 aestheton =
	reals	with t_3 about reals	utility of 1 real
t_3	reals	with t_2 about reals	utility of 1 real =
	hedons	with t_1 about hedons	utility of 2 hedons

The agreement between t_1 and t_2 about the value of aesthetons generates comparability between their assessments of hedons and reals: 1 hedon has the same value as 1 real. But since t_1 also agrees with t_3 about the value of hedons, we are equally committed to the claim that 2 hedons are worth 1 real. These are incompatible commitments.

Tarsney acknowledges the possibility of inconsistencies and suggests that where they arise, the three theories cannot be grouped into a shared comparability class. But in the above example, each pair exhibits all the qualities that would otherwise be sufficient, on this commensuration strategy, for comparability. If these qualities aren't sufficient for comparability in cases where inconsistencies arise, then it's hard to see why they are ever sufficient for comparability. Worse, novel theories could be constructed ad hoc to generate inconsistencies within comparability classes. So Tarsney's tactic for avoiding inconsistencies risks undermining the explanation of how intertheoretic comparisons are possible.

5.2 Structural intertheoretic agreement

Lockhart [2000] proposes a strategy for determining intertheoretic utility comparisons in terms of structural features of decision problems. In each decision problem, we treat the maximum utility attainable among the agent's options according to each theory as tied; similarly for the minimum attainable utility. The utility functions are therefore locally normalised at each decision problem.

One challenge for this strategy is that it's unable to represent theories as disagreeing about the stakes associated with a decision. (See [Ross 2006; Sepielli 2013] for discussion.) It requires that in the **Go Vegan?** example, the utility of missing out on tasty animal products must be as low, according to the pro-omnivorism theory, as the utility of eating meat, according to the pro-veganism theory. This is counterintuitive: however these theories are precisified, the stakes of the decision are intuitively much greater on the latter theory than on the former.

Another challenge: as Sepielli [2013] shows, this commensuration strategy sometimes rationalises cyclical preferences, because the intertheoretic utility function changes between decision problems. It also makes agents' preference rankings between two options depend on irrelevant alternatives.[21] Both of these properties are widely regarded as sufficient for irrationality.

A variant on this strategy would globally normalise instead of locally normalising. This would require equalising the maximum utilities, and equalising the minimum utilities, assigned to any *possible* outcome (even outcomes that can't result from any acts available to an agent in a particular decision problem). The problem for this strategy is that it rules out what should be

[21] In the interest of space, I won't summarise Sepielli's arguments.

easy cases for metanormative decision theory—totalist consequentialisms—that don't have upper or lower bounds for possible utility assignments.

Sepielli [2009] also proposes a strategy for determining comparisons in terms of structural features: if a certain kind of ratio of utility differences is equal between two theories, then they have enough cardinal structure in common to determine a unit comparison between theories. Specifically, Sepielli argues, for any x, y, z in the domain of the utility functions for theories t_1 and t_2, if

$$\frac{u_{t_1}(x) - u_{t_1}(y)}{u_{t_1}(y) - u_{t_1}(z)} = \frac{u_{t_2}(x) - u_{t_2}(y)}{u_{t_2}(y) - u_{t_2}(z)}$$

then $u_{t_1}(x) - u_{t_1}(y) = u_{t_2}(x) - u_{t_2}(y)$. Note that unlike Ross's proposal, Sepielli's doesn't require *intuitions* about comparability. It looks solely to structural features of utility functions to determine a comparison of their units. But as MacAskill [2014] notes (attributing the observation to Toby Ord), this proposal turns out to generate inconsistent commensurations.[22] Sepielli [2010] rejects this commensuration strategy on these grounds.

[22] These inconsistencies arise when the relevant equality of ratios of utility differences holds for more than one triad of outcomes. A simple example:

[figure 1]

On this proposal, the shared ratio of utility differences for x_1, x_2, and x_3 determines the conversion rate: 1 t_1-utile = 1 t_2-utile. The shared ratio of utility differences for x_3, x_4, and x_5 determines the conversion rate: 1 t_1-utile = 2 t_2-utiles. This entails that 1 t_2-utile = 2 t_2-utiles, which is a contradiction.

6 Conclusion

It's often argued that the problem of intertheoretic utility comparisons provides a fatal objection to the idea that normative uncertainty plays a substantive role in appropriate moral deliberation or moral evaluation. This paper shows that this is not so. It develops a systematic decision rule for decision-making under moral uncertainty that minimally generalises widely accepted decision rules and does not require use of intertheoretic comparisons. So it provides a principled answer to the question of how normative uncertainty influences appropriate action.[23]

University of California, San Diego

7 References

Arrow, Kenneth 1951. *Social Choice and Individual Values*, New York: Wiley.

Briggs, R.A. 2015. Transformative Experience and Interpersonal Utility Comparisons, *Res Philosophica* 92/2: 189–216.

Brown, Campbell 2011. Consequentialize This, *Ethics* 121/4: 749–71.

Buchak, Lara 2013. *Risk and Rationality*, Oxford: Oxford University Press.

Carr, Jennifer. Manuscript. The Hard Problem of Intertheoretic Comparison.

Colyvan, Mark, Damian Cox, and Katie Steele 2010. Modelling the Moral Dimension of Decisions, *Noûs* 44/3: 503–29.

Dougherty, Tom, Sophie Horowitz, and Paulina Sliwa 2015. Expecting the Unexpected, *Res Philosophica* 92/2: 301–21.

[23] Thanks to Ryan Doody, Melissa Fusco, Elizabeth Harman, Seth Lazar, Chip Sebens, Brian Talbott, and audiences at Hebrew University of Jerusalem and at the California Institute of Technology.

Gracely, Edward 1996. On the Noncomparability of Judgments Made by Different Ethical Theories, *Metaphilosophy* 27/3: 327–32.

Gustafsson, Johan, and Olle Torpman 2014. In Defence of My Favourite Theory, *Pacific Philosophical Quarterly* 95/2: 159–74.

Harman, Elizabeth 2011. Does Moral Ignorance Exculpate? *Ratio* 24/4: 443–68.

Harsanyi, John C. 1977. Rational Behavior and Bargaining Equilibrium in Games and Social Situations, Cambridge: Cambridge University Press.

Hedden, Brian 2012. Options and the Subjective Ought, *Philosophical Studies* 158/2: 343–60.

Hedden, Brian 2016. Does Mite Make Right? Decision-Making Under Normative Uncertainty. In *Oxford Studies in Metaethics 11*, ed. Russ Schafer-Landau, Oxford: Oxford University Press.

Hudson, James 1989. Subjectivization in Ethics. *American Philosophical Quarterly* 26/3: 221–29.

Isaacs, Yoaav 2019. The Problems of Transformative Experience. *Philosophical Studies*, https://doi.org/10.1007/s11098-018-01235-3.

Jackson, Frank 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32/127: 127–36.

List, Christian 2003. Are Interpersonal Comparisons of Utility Indeterminate? *Erkenntnis* 58/2: 229–60.

Lockhart, Ted 2000. *Moral Uncertainty and Its Consequences*, New York: Oxford University Press.

MacAskill, William 2014. *Normative Uncertainty*. PhD thesis, University of Oxford. https://www.academia.edu/8473546/Normative_Uncertainty.

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

Paul, L. A. 2014. *Transformative Experience*. New York: Oxford University Press.

Paul, L. A. 2015. What You Can't Expect When You're Expecting, *Res Philosophica* 92/2: 1–23.

Pettigrew, Richard 2015. Transformative Experience and Decision Theory, *Philosophy and Phenomenological Research* 91/3: 766–74.

Portmore, Douglas 2007. Consequentializing Moral Theories. *Pacific Philosophical Quarterly* 88/1: 39–73.

Robbins, Lionel 1932. An Essay on the Nature and Significance of Economic Science. London: Macmillan.

Ross, Jacob 2006. Rejecting Ethical Deflationism. *Ethics* 116/4: 742–68.

Sen, Amartya 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Sen, Amartya 1979. Interpersonal Comparisons of Welfare. In *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*, ed. M. J. Boskin, New York: Academic Press.

Sepielli, Andrew 2009. What to Do When You Don't Know What to Do. *Oxford Studies in Metaethics* 4, ed. Russ Shafer-Landau.

Sepielli, Andrew 2010. 'Along an Imperfectly-Lighted Path': Practical Rationality and Normative Uncertainty. PhD thesis. https://rucore.libraries.rutgers.edu/rutgers-lib/26567/.

Sepielli, Andrew 2013. Moral Uncertainty and the Principle of Equity Among Moral Theories. *Philosophy and Phenomenological Research* 86/3: 580–89.

Singer, Peter 1975. *Animal Liberation*. New York: Harper Collins.

Smith, Michael 2002. Evaluation, Uncertainty and Motivation, *Ethical Theory and Moral Practice* 5/3: 305–20.

Staffel, Julia 2019. Expressivism, Normative Uncertainty, and Arguments for Probabilism. *Oxford Studies in Epistemology 6,* ed. T.S. Gendler, J. Hawthorne. New York: Oxford University Press.

Tarsney, Christian 2018. Intertheoretic Value Comparison: a Modest Proposal. *Journal of Moral Philosophy* 15/3: 324–44.

Weatherson, Brian 2014. Running Risks Morally. *Philosophical Studies* 167/1: 141–63.