Fixed-Point Solutions to the Regress Problem in Normative Uncertainty

Philip Trammell

August 29, 2018

Abstract

When we are faced with a choice among acts, but are uncertain about the true state of the world, we may be uncertain about the acts' "choiceworthiness". Decision theories guide our choice by making normative claims about how we should respond to this uncertainty. If we are unsure which decision theory is correct, however, we may remain unsure of what we ought to do. Given this decision-theoretic uncertainty, meta-theories attempt to resolve the conflicts between our decision theories... but we may be unsure which meta-theory is correct as well. This reasoning can launch a regress of ever-higher-order uncertainty, which may leave one forever uncertain about what one ought to do. There is, fortunately, a class of circumstances under which this regress is not a problem. If one holds a cardinal understanding of subjective choiceworthiness, and accepts certain other criteria (which are too weak to specify any particular decision theory), one's hierarchy of metanormative uncertainty ultimately converges to precise definitions of "subjective choiceworthiness" for any finite set of acts. If one allows the metanormative regress to extend to the transfinite ordinals, the convergence criteria can be weakened further. Finally, the structure of these results applies straightforwardly not just to decision-theoretic uncertainty, but also to other varieties of normative uncertainty, such as moral uncertainty.

1 Introduction

People sometimes make claims about how we ought to act in the face of empirical uncertainty. A "decision theory" is a collection of such claims. Because they make demands on our behavior, decision theories are "norms". Moral theories are also norms, for example, because they too are collections of claims about how we "ought" (albeit in another sense) to act. We suffer decision-theoretic uncertainty when we assign positive probability to conflicting decision theories, and, more generally, some form of normative uncertainty whenever we assign positive probability to conflicting norms.

The most widely accepted decision theory, by far, is expected utility theory. Expected utility theory can be interpreted as the claim that we have a cardinal utility function whose value depends on what act we choose and on the state of the world, and that we ought, normatively, to act so as to maximize the expected value of that function, given our uncertainty across states. Alternatively, expected utility theory can be interpreted as the claim that we have an *ordinal* utility function whose value depends on what act we choose and on the state of the world, and that we ought, normatively, to act in such a way as satisfies various assumptions (the von Neumann-Morgenstern axioms, perhaps), which together happen to entail that we will be acting *as if* we were maximizing the expected value of a cardinal utility function. Either way, expected utility theory is a theory about how an ideal agent acts. It is, in other words, a set of normative claims: a set of claims about what we ought to do.

Expected utility theory is not self-evident. Soon after its foundations were laid in the 1950s by John von Neumann, Oskar Morgenstern, and Leonard Savage, others, such as Maurice Allais, raised objections to the claim that expected utility maximization is a model of ideal behavior. It is true that, in the apparent absence of plausible alternatives, expected utility theory came to serve as the unchallenged basis for almost all of economic theory. But the debate over the normative claims that underlie it continued among philosophers, and has recently received new attention with Lara Buchak's 2013 publication of *Risk and Rationality*, which argues that it can be rational to violate the von Neumann-Morgenstern "independence" axiom so as to act on explicit risk preferences. In sum, the normative claims made by expected utility theory—even if modern economists sometimes take them for granted—are claims about which an agent can be uncertain.

What "should" one do in the face of decision-theoretic uncertainty? Is there even a coherent way to interpret that question? The position that there is—that we "ought" to act on the basis of our normative uncertainty in general—has been called "uncertaintism" (or, less frequently, "metanormativism"). The uncertaintist, presumably, must then offer an account of what one should do when one assigns, say, a seventy percent chance to the truth of expected utility theory, and a thirty percent chance to the truth of Buchak's alternative.

As we can see, questions about how to deal with empirical uncertainty give rise to questions about how to deal with decision-theoretic uncertainty. But our regress does not stop there. Just as decision theories are theories about how to act in the face of empirical uncertainty, let us use the term "metanormative theories" for collections of claims about how we ought to act in the face of normative uncertainty. It seems that, just as we can suffer normative uncertainty, we can suffer metanormative uncertainty as well: we can assign positive probability to conflicting metanormative theories. Metametanormative theories, then, are collections of claims about how we ought to act in the face of

metanormative uncertainty. And so on. In the end, it seems that the very existence of normative claims—the very notion that there are, in some sense or another, ways "one ought to behave"—organically gives rise to an infinite hierarchy of metanormative uncertainty, with which an agent may have to contend in the course of making a decision.

Postulating such a hierarchy may seem like a strange and unneccesarily complex solution to a rather small and obscure problem. By analogy, therefore, consider two other structures explored in detail by recent generations of economic theorists and philosophers: belief hierarchies and preference hierarchies. We have beliefs about the world; and when one reflects on the fact that we can also have beliefs about people's beliefs, one can hardly help but document the emergence of a "belief hierarchy", constituted of beliefs about the world, beliefs about beliefs (2^{nd}-order beliefs), beliefs about beliefs about beliefs (3^{rd}-order), and so on. Likewise, we have preferences over features of the world, and the fact we can also have preferences over the contents of people's preferences (2^{nd}-order preferences), and so on, gives rise to the well-established concept of a "preference hierarchy". Decision theories tell us how to act in the face of uncertainty about the true state of the world. The fact that we may have to act in the face of uncertainty about the true decision theory, therefore, seems plausibly to give rise to a "metanormative hierarchy", similar in many respects to the hierarchies above.[1]

Furthermore, to repeat, the problem of decision-theoretic uncertainty is more than hypothetical. I myself, for instance, do not currently know the correct decision theory. Expected utility theory seems highly plausible to me, but I cannot fully rule out Buchak's arguments for some sort of risk-weighted expected utility theory, or arguments for theories that evade "Pascal's muggings" by giving special treatment to low-probability but high-expected-value events, to name just two examples. I am likewise not certain of the correct metatheory, nor of the correct theory at any order of the above hierarchy.

Despite all this uncertainty, however, I am rarely—perhaps never—uncertain of how I subjectively ought (in a rational *or* in a moral sense) to act. Even when I ponder my options, it is generally evident to me that the act of pondering is in fact appropriate to my overall state of uncertainty. Somehow, from my infinite hierarchy of metanormative uncertainty, a certain kind of normative certainty can and often does arise.

The process by which this happens seems potentially interesting. Questions of decision-theoretic uncertainty have received little attention so far, however—

[1] There is another infinite regress, and associated hierarchy concept, closely associated with decision theory. When we are not sure which act would maximize our utility, we may find it useful to ponder our options, even if doing so would come at a cost. Since pondering is in some sense just another option, we must then ponder whether to take one of our first-order options, or to take the option of pondering among them. And so on. Lipman (1991) explores this hierarchy in depth, in terms somewhat similar to those we will use here.

research on normative uncertainty has focused predominantly on moral uncertainty—and even the most recent inquiries into normative uncertainty (again, usually presented in the context of moral uncertainty) have generally brushed such questions aside. Weatherson (2014) avoids the problem by positing that, in essence, the only norms are "first-order"; that there simply is no right or wrong way to deal with normative uncertainty. Sepielli (2014), on the other hand, hopes for "fixed point solutions"—but simply assumes that such fixed points will exist in general, and does not rigorously search for them.[2] Tarsney (2017) observes that such fixed points will not necessarily exist, and concludes that solutions to the regress problem must thus be found altogether elsewhere.[3] Finally, many accept that uncertaintism does indeed require some sort of fixed-point solution to the regress problem—but conclude from this that there simply must be something deeply wrong with uncertaintism. The following sentiment, described in Volume 7 of *Oxford Studies in Normative Ethics* (2017), is typical:

> But a regress problem looms. Let us suppose that I am uncertain among some ordinary moral theories, [and] I ask what to do given the probability distribution over T1...Tn. But I am uncertain as to the answer, assigning some probability to each of U1...Un. This prompts me to ask what I ought to do given *this* probability distribution.... We can imagine this process iterating indefinitely.... The possibility of normative uncertainty all the way up makes the uncertaintist project look pointless.

Does this possibility in fact render the uncertainist project pointless? Or can one accept the possibility of normative uncertainty all the way up, and still be norm-guided in some important classes of circumstances? At least in some circumstances, it seems, the latter. To begin to answer the above questions more precisely, Section 2 presents a formal framework that aims to capture our intuitions about the concept of decision-theoretic uncertainty. Within this framework, Section 3 specifies various conditions under which two similar convergence results follow, as shown in Section 4. Section 5 considers the earlier sections' implications for normative uncertainty more generally.

[2]He writes: "Suppose... that I can do any of mutually exclusive actions A...Z. Perhaps my uncertainty regarding objective normativity will intentionally explain my doing any of A...R.... My hope is to show that, as a general matter, potential actions will be hived off with each stepping-back..., but I'll have to make good on that suspicion elsewhere."

[3]Tarsney (2017) also argues that finite, boundedly rational agents cannot be normatively required to work out the fixed point of an infinite hierarchy, even if it exists. He thus distinguishes between the "ideal regress problem", which an ideal agent with perfect reasoning ability might face, and the "non-ideal regress problem". In order to separate the issue of normative uncertainty from the issue of bounded rationality, here we will consider only what he calls the "ideal regress problem". Note that we are implicitly assuming that normative facts are not logical necessities; if they are, then it is impossible for an agent with perfect reasoning ability to suffer normative uncertainty.

2 Framework

2.1 Choiceworthiness

A finite set $A = \{a_1, ..., a_{|A|}\}$ of "feasible acts" presents itself. There is a finite set of "possible states" $S = \{s_1, ..., s_{|S|}\}$ to which I assign positive probability.[4] I assign utilities to performing each act in each state, as represented by the utility function $u(A, s)$, where the value assigned to each act is not necessarily independent of the alternatives in A.[5] I also find myself in an overall finite epistemic position e, specifying the probabilities I assign to all relevant claims.[6] Let us call $\pi = \langle A, u, e \rangle$ my "choice problem".

Definition 2.1. A choice problem π is a triple of (i) a set of acts A, (ii) an epistemic position e, and (iii) a state-contingent utility function u over A.

We will say that my utility function specifies the "objective choiceworthiness" (or simply "choiceworthiness") of each act, conditional on each state. That is, given s, the choiceworthiness of a_i is $u(A, s)_i$—the i^{th} element of the $|A|$-vector $u(A, s)$. From the probabilities I assign to the states in S, therefore, I also assign probabilities to potential values of the objective choiceworthiness of each act.

Definition 2.2. A (finite) choiceworthiness distribution is a (finite) probability distribution over choiceworthiness values for some (finite) set of acts.

Let $\mathbb{D}^n$ denote the set of all finite probability distributions in $\mathbb{R}^n$, and let some $d(\pi) \in \mathbb{D}^n$ represent the choiceworthiness distribution entailed by π.[7]

[4] Here and elsewhere, we will assume that all credences satisfy the Kolmogorov probability axioms. Note that this implies that all the sets over which I have probability distributions are nonempty.

[5] We will assume that the probability of each state is independent of the chosen act. We will thus bypass the question of how to act in the face of such dependency (i.e. causal decision theory vs. evidential decision theory and other alternatives), and focus entirely on the question of how to act in the face of uncertainty over states (i.e. expected utility theory vs. its alternatives). For an analysis of how to approach uncertainty between causal and evidential decision theory, see MacAskill (2016b).

[6] More precisely, let e specify my probability distribution over the set of [{states of the world} × {decision theories (or, 1-metatheories)} × {2-metatheories} × {3-metatheories} × ...]. The concept of a "k-metatheory" is defined in Section 2.4.

[7] Many other finite probability distributions over $\mathbb{R}^n$ might do just as well as the chosen d at representing my finite choiceworthiness distribution. Exactly which others depends on how much structure is contained in our understanding of "utility". If we understand utility to be a merely "ordinal" quantity, for instance, then any transformation of d that is monotonic in choiceworthiness (and constant in probability) represents the same choiceworthiness distribution. We are here assuming nothing about utility except that it at least partially orders act-state pairs from a given {feasible set × possible set}, and that $\mathbb{R}$ is "rich enough" to capture any potential difference between the choiceworthiness values of particular act-state pairs—that choiceworthiness cannot be lexicographic, for instance. As discussed in Section 2.3, these assumptions about choiceworthiness will follow from similar assumptions made explicitly about subjective choiceworthiness.

2.2 Subjective Choiceworthiness

I am uncertain about acts' choiceworthinesses. Even so, I may know that one act is the most appropriate for me to choose, given my epistemic position. As I write this, for instance, I assign high probability to the event that, if I go to the doctor, I will swiftly be cured of my back injury (an outcome I would prefer immensely to the status quo), and low probability to the roughly complementary event that, if I go to the doctor, I will waste some time and remain injured (an outcome to which I would slightly prefer the status quo). Despite this uncertainty, and all my other uncertainty, I am in fact certain that going to the doctor is the "better choice" for me right now (by far!). There is thus some scale on which the act of going to the doctor scores higher for me than the act of not going—and would score higher for anyone with the same utility function, in the same overall epistemic position, facing the same set of feasible acts. Let us call this scale "subjective chiceworthiness".

It is my intuition that subjective choiceworthiness c, when well-defined, is fundamentally a *cardinal* scale. That is, I would maintain that a representation of acts' subjective choiceworthinesses (for an agent in a given situation) in $\mathbb{R}^{|A|}$ would be unique at least up to affine transformation. If my feasible act set $A = \{a_1, a_2, a_3\}$ consists of going to the doctor (a_1), going to a very slightly less competent doctor (a_2), or not going at all (a_3), then there is some important and foundational sense in which, given my epistemic position and my preferences, the distance between $c(\pi)_1$ and $c(\pi)_2$ is *less than* the distance between $c(\pi)_2$ and $c(\pi)_3$. It might be objected that I will always do whatever winds up being most subjectively choiceworthy; that therefore, in the absence of a specified theory of decision-making under uncertainty, no information is conveyed by postulated differences between the acts not chosen; and that c is therefore better understood as merely an ordering, or perhaps even as a choice relation. To this it might be replied that, under certain circumstances, differences in subjective choiceworthiness could bear some relationship to the subjective probability with which a subjectively sub-optimal act would become optimal upon further reflection. Or that cardinal subjective choiceworthiness takes on a clearer meaning in other situations of normative uncertainty (i.e. one might *not* choose the most subjectively morally choiceworthy act, and might in some sense be more blameworthy the less subjectively morally choiceworthy one's act was)—and that it would be strange for subjective choiceworthiness to be fundamentally cardinal in one of these situations but not the other. Or that our models of the world are generally simpler when we extend our intuitions regarding quantities' cardinality beyond the domains in which they happen to be testable—such as our intuition that temperature is generally cardinal, even on some cold, distant star that we will only discover if its temperature rises above some threshold.

Furthermore, cardinal subjective choiceworthiness allows for the convergence results described below, and less structured interpretations of subjective

choiceworthiness would not. If we are otherwise persuaded that the regress problem must have some solution or other, it is not circular to allow this observation itself to lend credibility to the concept of cardinal subjective choiceworthiness.

In any event, for the purposes of this analysis, we will understand subjective choiceworthiness (again, when well-defined) to be cardinal. We will represent it by a "subjective choiceworthiness function" $c(\pi)$, where c assigns a real number to the subjective choiceworthiness of each of the acts in a feasible set A, for an agent with a utility function u, in epistemic position e.[8]

2.3 Metachoiceworthiness

In general, if I am to translate a choiceworthiness distribution d into a determination of how to act, I must invoke a "decision theory": a collection of claims concerning how to evaluate acts in light of one's choiceworthiness distribution. For example, using this terminology, one decision theory is "Expected Choiceworthiness Theory" (EC). EC is characterized by the fact that, if I am certain that it is the correct decision theory, then each act's subjective choiceworthiness for me is its expected choiceworthiness under d.[9] Another decision theory would be "minimum choiceworthiness"—a theory characterized by the fact that, if I am certain that it is the correct decision theory, then each act's subjective choiceworthiness for me is its minimum possible choiceworthiness under d.

Just as I am uncertain about the true state of the world, I may also be uncertain about the correct decision theory. To come to a determination of how to act, therefore, I may have to invoke a sort of "meta decision theory" (or, "2-metatheory"): a collection of claims concerning how to respond to one's uncertainty over decision theories.

Note that, since this is so, the decision theories (we will awkwardly call these "1-metatheories", for ease of indexing) cannot themselves be claims about subjective choiceworthiness. This is perhaps a surprising claim, so it bears repeating: expected utility theory (for example) is *not*, in this language, a theory about what subjective choiceworthiness is, or even about what it *ought* to be

[8]Note that by having c map into the real numbers, we are assuming that all information about differences in subjective choiceworthiness (and therefore utility) can be captured by ratios of differences in real numbers. We are here explicitly assuming for subjective choiceworthiness what we provisionally assumed above for objective choiceworthiness—that, for instance, it cannot be lexicographic. Like probability theories that let us condition on probability 0 events, utility theories that let us distinguish between acts that differ infinitesimally in choiceworthiness may also be interesting to consider in light of the regress problem. However, we will not touch them here.

[9]Let us distinguish EC from "Maximize Expected Choiceworthiness" (MEC). MEC is the weaker theory characterized only by the fact that, if I am certain that it is correct, then the acts with the *highest* subjective choiceworthiness for me are the acts with the highest expected objective choiceworthiness under d.

"all things considered". It is, rather, a theory about what subjective choiceworthiness "*1-ought*" to be, for someone with a given objective choiceworthiness distribution over his feasible set—or, a theory about what subjective choiceworthiness *is* for someone with a given objective choiceworthiness distribution over his feasible set, *if he knows the true 1-metatheory*.

Suppose, for instance, that I am faced with three feasible acts, that I assign probability to each of two 1-metatheories, t_1 and t_2, and that I am certain of "2-metatheory" m. The theories are such that if I were certain of t_1, the subjective choiceworthinesses of the acts would be ordered $a_1 \succ a_2 \succ a_3$; if I were certain of t_2, the subjective choiceworthinesses of the acts would be ordered $a_3 \succ a_2 \succ a_1$; and, given the probabilities I assign to t_1 and t_2, but my certainty about m, the subjective choiceworthinesses of the acts are in fact ordered $a_2 \succ a_3 \succ a_1$. Although I assign probability $\frac{1}{2}$ to t_1, I assign *no* positive probability to the event that a_1 is more subjectively choiceworthy than a_2 from my epistemic position. The 1-metatheories' claims, therefore, are not claims about the acts' subjective choiceworthinesses given my empirical uncertainty, but about how the acts score on an altogether different scale. Let us call this scale "metachoiceworthiness", or "1-metachoiceworthiness". Of course, metachoiceworthiness must be constructed such that, if I know that an act's 1-metachoiceworthiness is x, then the act's subjective choiceworthiness for me is also x. We might therefore informally think of 1-metachoiceworthiness as "whatever subjective choiceworthiness is, for someone who knows the correct 1-metatheory". But since, again, decision theories are not actually claims about subjective choiceworthiness, let us begin by thinking about 1-metachoiceworthiness on its own terms, and only afterward consider its relationship to subjective choiceworthiness.

In any event, the elusiveness of subjective choiceworthiness is not restricted to "order 1". Just as I may be uncertain as to the correct 1-metatheory, I may be uncertain as to the correct 2-metatheory; I may therefore have to appeal to a "3-metatheory"; and the 2-metatheories are therefore making claims not about acts' subjective choiceworthiness given beliefs about their 1-metachoiceworthiness, but about acts', say, "2-metachoiceworthiness" given beliefs about their 1-metachoiceworthiness. So our regress begins.

2.4 k-Metachoiceworthiness

Let us call choiceworthiness "0-metachoiceworthiness", choiceworthiness distributions "0-metachoiceworthiness distributions", and decision theories "1-metatheories". The concepts of k-choiceworthiness, k-metachoiceworthiness distributions, and k-metatheories can then together be defined recursively.

Definition 2.3. The k-metachoiceworthiness c_k of an act a_i, for an agent facing finite choice problem π, is a_i's subjective choiceworthiness for an agent with the same (k-1)-metachoiceworthiness distribution as that entailed by π, but who knows the correct k-metatheory.

Let us denote acts' relative k-metachoiceworthiness by the two-place relation $\succ_{\pi,k}$.

Definition 2.4. A (finite) k-metachoiceworthiness distribution $d_k \in \mathbb{D}^{|A|}$ is a probability distribution over k-metachoiceworthiness values for some (finite) set of acts A.

Definition 2.5. A k-metatheory, applied to a finite set of acts A, is a function $t_k : \mathbb{D}^{|A|} \to \mathbb{R}^{|A|}$, representing claims about the k-metachoiceworthiness of the acts in A given (k-1)-metachoiceworthiness distribution $d_{k-1} \in \mathbb{D}^{|A|}$.[10]

We can now define a few aditional terms.

Definition 2.6. A k-metatheory distribution d_{t_k} is a probability distribution over k-metatheories.

Definition 2.7. A metatheoretic hierarchy (or simply "hierarchy") T is a collection of k-metatheories t_k with one for each $k \in \mathbb{N}$.

Definition 2.8. A hierarchy distribution d_T is a probability distribution over hierarchies.

Let $|d_{t_k}|$ and $|d_T|$ denote the number of k-metatheories and hierarchies, respectively, to which I assign positive probability.

Let $\vec{c_k} \in \mathbb{R}^{|d_{t_k}||A|}$ represent the claims made by my $|d_{t_k}|$ k-metatheories about the k-metachoiceworthinesses of the $|A|$ acts in A. Let $\vec{p_k} \in \Delta^{|d_{t_k}|-1}$ represent the probabilities I assign to these k-metatheories. We can now represent my k-metachoiceworthiness distribution by $d_k = \langle \vec{c_k}, \vec{p_k} \rangle$.[11]

2.5 The Relationship of k-Metachoiceworthiness to Subjective Choiceworthiness

Upon introducing the cardinal subjective choiceworthiness function $c(\pi)$ above, we placed no restrictions on what it could be. Now that we have documented the emergence of an elaborate web of concepts concerning π, however, we can consider how it relates to c.

Recall that k-metachoiceworthiness claims are defined so that, if I know that an act's k-metachoiceworthiness for me is x, the act's subjective choiceworthiness for me is x. Let us now introduce a compatible, minimally restrictive

[10] Strictly speaking, if we want our k-metatheories to make k-metachoiceworthiness claims over finite act-sets of arbitrary size, we would have to say that a k-metatheory is a family of functions $\{t_k^n\}$ from $\mathbb{D}^n$ to $\mathbb{R}^n$, with one for each $n \in \mathbb{N}$. For simplicity, however, we will take $n = |A|$ as given and interpret our project only as an attempt to find criteria under which the subjective choiceworthinesses of any n acts will be well-defined—with the understanding that identical reasoning would apply to any other n.

[11] This is not to say that a given distribution can only be represented by one particular vector pair. Multiple k-metatheories make make the same k-metachoiceworthiness claims in some situation, for instance.

principle with which one's subjective choiceworthiness function might comply in the face of uncertainty about an act's k-metachoiceworthiness.

Definition 2.9. The Dominance Principle is the principle that

- If $b \geq x \ \forall b \in [\vec{c_k}]_i$, and $b^* > x$ for some $b^* \in [\vec{c_k}]_i$, then $c(a_i) > x$.
- If $b \leq x \ \forall b \in [\vec{c_k}]_i$, and $b^* < x$ for some $b^* \in [\vec{c_k}]_i$, then $c(a_i) < x$.

Note that if I accept the Dominance Principle, it follows immediately that my subjective choiceworthiness for an act a_i is well-defined whenever $|\cap_{k\in\mathbb{N}} [\min([\vec{c_k}]_i), \max([\vec{c_k}]_i)]| = 1$. That is, whenever exactly one number lies in the ranges of "admissible" (not dominated) k-metachoiceworthiness values, across all k, for an act, that number must be the act's subjective choiceworthiness.

Note also that any claim about subjective choiceworthiness itself, such as the Dominance Principle, in some sense takes on both a positive and a normative interpretation. One could interpret the Principle normatively as asserting that one's subjective choiceworthiness always *ought* to obey the above pattern. In this case, if one accepts the Principle, one's subjective choiceworthiness also *does* obey it, since to hold that an act should be ranked highly for someone in your epistemic position is simply another way to say that it is highly subjectively choiceworthy. Alternatively, one could interpret the Principle positively as asserting that, as a matter of fact, subjective choiceworthiness always obeys the above pattern. If one accepts this claim (and that "ought implies can"), one must also accept that subjective choiceworthiness always *ought* to obey the above pattern. Either way, if one accepts the Principle, one cannot assign positive probability to k-metatheories that claim that the k-metachoiceworthiness of an act lies outside the admissible range imposed by one's k'-metachoiceworthiness distribution for the act for lower orders $k' < k$.

Finally, note that the framework outlined here differs from other approaches to subjective choiceworthiness in the following respect. Some other approaches (e.g. that of MacAskill (2016a)) begin with the normative theories in all their diversity; work through problems of intertheoretic comparability; and then try to define subjective choiceworthiness with no more structure than necessary—even if that is nothing but a binary classification of acts into the "permissible" and the "impermissible" (as recommended, for instance, in Barry and Tomlin (2016)). The above approach, by contrast, begins by assuming that subjective choiceworthiness is a cardinal scale, and it characterizes k-metachoiceworthiness claims, and the k-metatheories that make them, in terms of the subjective choiceworthinesses that they would induce if they were known. This approach has the cost of assuming cardinal subjective choiceworthiness, but it has the benefit of immediately giving all my k-metachoiceworthiness claims both unit and level comparability, without requiring any further assumptions.[12]

[12]Thus, from a cardinal definition of subjective choiceworthiness, we also get a cardinal

3 Conditions

3.1 Totality

A "partial k-metatheory" would be one that makes claims about the k-metachoiceworthinesses of some acts under some (k-1)-metachoiceworthiness distributions, but not of all acts under all (k-1)-metachoiceworthiness distributions. A partial decision theory of "strict dominance", for instance, claims that $a_i \succ_{\pi,1} a_j \iff u(A,s)_i > u(A,s)_j \ \forall s \in S$, and makes no other claims at all. That is, it claims that an act a_i is more 1-metachoiceworthy than an act a_j if and only if a_i is more objectively choiceworthy than a_j in all the states to which I assign positive probability, and it is silent about acts' relative 1-metachoiceworthinesses in all other cases.

Conversely,

Definition 3.1. A k-metatheory, applied to a finite set of acts A, is <u>total</u> if it is defined throughout $\mathbb{D}^{|A|}$.

One condition for the results below is that I assign positive probability only to total decision theories. Believing that the true decision theory is total is, I think, reasonably well motivated by the sense that, just as I know acts' 0-metachoiceworthiness (i.e. objective choiceworthiness) if I know the true state, I should be able to know acts' 1-metachoiceworthiness if I know the true decision theory (and so on up the hierarchy). In any event, we will remove partial decision theories from consideration so as to separate the regress problem from the problems of incomparability that can plague normative uncertainty in their own right.[13]

3.2 Continuity

We will say that

Definition 3.2. A k-metatheory t_k is <u>continuous</u> if $\forall \delta > 0 \ \exists \varepsilon > 0 : |\vec{x}| < \varepsilon \implies |t_k(c_{k+1}^{\rightarrow} + \vec{x}, p_{k-1}^{\rightarrow}) - t_k(c_{k-1}^{\rightarrow}, p_{k-1}^{\rightarrow})| < \delta \ (\delta \in \mathbb{R}, \varepsilon \in \mathbb{R}, \vec{x} \in \mathbb{R}^{|d_{t_{k-1}}||A|})$.

definition of utility, without having to assume it explicitly. (By similar reasoning, we also get cardinal definitions of k-metachoiceworthiness for all k.) Note that we are not taking the Von Neumann-Morgenstern approach of defining my utility function so that it represents the choices I would make if I were maximizing expected utility; indeed, our project is to explore how far I can stray from certainty about expected utility theory while still knowing how I subjectively ought to act.

[13] Note that the framework laid out in Section 2 does not allow us to assign positive probability to the "nihilistic decision theory" (one that makes no claims about acts' 1-metachoiceworthinesses under any choiceworthiness distribution). Since a decision theory is a collection of claims determining what acts' subjective choiceworthinesses would be for me if I knew how to respond to my empirical uncertainty, and since my subjective choiceworthiness is already defined in the degenerate case of empirical certainty, all my decision theories at least claim that an act's subjective choiceworthiness is its objective choiceworthiness, when my objective choiceworthiness distribution is degenerate.

That is, we will call t_k "continuous" if slight changes to the individual (k-1)-metachoiceworthiness claims to which one assigns positive probability produce only slight changes to the k-metachoiceworthiness claims made by t_k. (We will not require t_k to respond continuously to the *probability* one assigns to some (k-1)-metachoiceworthiness claim.)

A second condition, necessary for only the first of the results below, is that I assign positive probability only to continuous decision theories.

3.3 The Analog Principle

MacAskill (2014) argues that, when we are facing both empirical and normative uncertainty over a set of acts, there is a sense in which we should treat our empirical and normative uncertainty "analogously". If I am uncertain which act is objectively best, it may seem unlikely that the appropriate response to my uncertainty would depend on the reason (i.e., empirical or normative) for my uncertainty—especially upon considering that I might have uncertainty about how to behave without even knowing the reason for my uncertainty.

In the context of the regress problem, one might likewise argue that we should treat our empirical and k-metatheoretic uncertainty analogously. More formally:

Definition 3.3. Let $t_k^* : \mathbb{D}^{|A|} \to \mathbb{R}^{|A|}$ denote the true k-metatheory. The Analog Principle is the claim that $t_k^* = t_1^* \ \forall k \geq 1$.

A final condition, necessary only for the first of the results below, is that I accept the Analog Principle.[14]

4 Convergence

4.1 Intuition

In the context of the framework above, the commonness of well-defined subjective choiceworthiness is not surprising. If I assign positive probability to a finite number of theories, and they disagree about how subjectively choiceworthy some act should be for me, there will be a minimum and a maximum to that range of values. In the face of that uncertainty, my subjective choiceworthiness should be somewhere in the interior of the range. Where, exactly? I will assign positive probability to different answers, producing a smaller range. And so on. Given a few other assumptions (either the continuity of my theories and my

[14]One might wonder if it matters whether my beliefs about the k-metatheories are correlated across different orders k' (as of course they are—very strongly!—if I accept the Analog Principle), or whether they are correlated my beliefs about the state of the world. In fact, it does not. A k-metatheory is simply a function of *my (k-1)-metachoiceworthiness distribution*; a k-metatheory's output therefore does not depend on the probability that it is the true k-metatheory, nor on its probability conditional on some state or k'-metatheory.

acceptance of the Analog Principle, or the possibility of transfinite hierarchies), this process will not "get stuck" by shrinking the range of potential subjective choiceworthiness values for each act merely from a larger range to a smaller range. Instead, the process is guaranteed ultimately to shrink said range to a single point.

In other words, nothing very counterintuitive falls out of the mathematical setup of the problem. The point of this exercise is simply to formally illustrate a coherent framework whereby our intuitions about normative uncertainty—including about the infinite regress that it threatens—can be reconciled with the understanding that, at the end of the day, we make norm-guided decisions.

With that said, the convergence results can be stated as follows.

4.2 Natural Hierarchies

Theorem 1. *If one assigns positive probability only to a finite set of decision theories all of which are total and continuous, and if one accepts the Dominance Principle and the Analog Principle, then one's subjective choiceworthiness is well-defined over any finite set A of acts.*

Proof: Let d_t represent my probability distribution over decision theories. By the Analog Principle, d_t also represents my probability distribution over k-metatheories, for any k. My probability distribution over the available acts' k-metachoiceworthinesses can then be represented by the pair $\langle \vec{c_k}, \vec{p_0} \rangle$, $\vec{c_k} \in \mathbb{R}^{|d_t||A|}$, $\vec{p_0} \in \Delta^{|d_t|-1}$, for all $k \geq 1$. Note that $\vec{p_0}$ does not depend on k. We can thus let $f : \mathbb{R}^{|d_t||A|} \to \mathbb{R}^{|d_t||A|}$ represent the function, fully specified by my probability distribution over decision theories, from the ordered set of k-metachoiceworthiness claims about A made by my $|d_t|$ k-metatheories to the ordered set of $(k+1)$-metachoiceworthiness claims about A made by my $|d_t|$ $(k+1)$-metatheories.

Let us think of the output of f as an $\mathbb{R}^{|A|}$-valued vector of length $|d_t|$, with one point in $\mathbb{R}^{|A|}$ given by each decision theory to which I assign positive probability. Since all the decision theories to which I assign positive probability are continuous in $\mathbb{R}^{|d_t||A|}$, and since vector-valued functions are continuous if their components are continuous, f is continuous.

Let us now return to thinking of the output of f after k iterations as a vector $\vec{c_k}$ (a $\mathbb{R}^{|d_t|}$-valued vector of length $|A|$, representing the k-metachoiceworthiness assigned to each act by each theory). By the Dominance Principle, for each act a_i there either exists an order $k : \min([\vec{c_k}]_i) = \max([\vec{c_k}]_i)$, or else $\min([\vec{c_{k+1}}]_i) > \min([\vec{c_k}]_i)$ and $\max([\vec{c_{k+1}}]_i) < \max([\vec{c_k}]_i)$ for all k. In the former case, the subjective choiceworthiness of a_i for me is of course well-defined.

In the latter case, the sequence $\{\min([\vec{c_k}]_i)\}$ is monotonically increasing, and the sequence $\{\max([\vec{c_k}]_i)\}$ is monotonically decreasing, in k. Since $\min([\vec{c_k}]_i)$ is bounded above (for example, by $\max([\vec{c_0}]_i)$), each sequence has a limit, by the

Monotone Convergence Theorem. It now follows from the continuity of f that $\lim_{k\to\infty} \min([\vec{c_k}]_i) = \lim_{k\to\infty} \max([\vec{c_k}]_i)$.

To see this, by contradiction let $\{c_j\}$ be a convergent subsequence of $\{c_k\}$ (as must exist, by the boundedness of $\{c_k\}$), and let $\vec{c} = \lim_{k\to\infty} \vec{c_j}$, with $\min([\vec{c}]_i) < \max([\vec{c}]_i)$. Setting $2\delta = \min(f(\vec{c})_i) - \min([\vec{c}]_i)$, we know that $|f(\vec{c}) - \vec{c}| \geq 2\delta$. (Let t be one of the theories assigning the minimum value to a_i under $\vec{c}$. Since t must assign at least the minimum value to a_i under $f(\vec{c})$, 2δ can, by the Triangle Inequality, serve as a lower bound for the difference between $f(\vec{c})$ and $\vec{c}$.) And since $\{c_k\} \to \vec{c}$, we can for any ε choose j large enough that $|\vec{c_j} - \vec{c}| < \varepsilon$. We now have a point $\vec{c}$ and a distance δ such that, for any sufficiently small ε (namely $\varepsilon \leq \delta$), there is a j^* with $|\vec{c_j} - \vec{c}| < \varepsilon$, but $|f(\vec{c_j}) - f(\vec{c})| \geq \delta$, for all $j \geq j^*$. (This follows from the Reverse Triangle Inequality: $|f(\vec{c_j}) - f(\vec{c})| \geq ||f(\vec{c}) - \vec{c}| - |f(\vec{c_j}) - \vec{c}|| = ||f(\vec{c}) - \vec{c}| - |\vec{c_{j+1}} - \vec{c}|| \geq 2\delta - \varepsilon \geq \delta$.) Since there is no ε small enough to ensure that $|\vec{x} - \vec{c}| < \varepsilon \implies |f(\vec{x}) - f(\vec{c})| < \delta$ ($x \in \mathbb{R}^{|d_t||A|}$), f is not continuous in $\mathbb{R}^{|d_t||A|}$.

We have seen that for each act a_i, $\lim_{k\to\infty} \min([\vec{c_k}]_i) = \lim_{k\to\infty} \max([\vec{c_k}]_i)$. It follows that $|\cap_{k\in\mathbb{N}} [\min(d_k(\pi)_i), \max(d_k(\pi)_i)]| = 1$ for each a_i. In other words, the subjective choiceworthiness of each act in A is well-defined. □

Proposition 4.1. If two acts are equally subjectively choiceworthy, this fact will not necessarily be revealed by the iterated application of one's distribution over k-metatheories.

Proof: Consider the following example.

I assign probability $\frac{1}{2}$ to Expected Choiceworthiness Theory, and to the analogous hierarchy (T_1) according to which the (k+1)-metachoiceworthiness of an act is its expected k-metachoiceworthiness. I assign probability $\frac{1}{2}$ to a risk-averse hierarchy of theories (T_2) according to which the (k+1)-metachoiceworthiness of an act is the average of its expected k-metachoiceworthiness and its minimum possible k-metachoiceworthiness.

I am deciding between two acts, a_1 and a_2. I assign probability $\frac{1}{2}$ to a state in which a_1 has objective choiceworthiness 0 and probability $\frac{1}{2}$ to a state in which a_1 has objective choiceworthiness 1. Act a_2 has objective choiceworthiness $\frac{1}{3}$ in both states. The k-metachoiceworthiness of a_2 is $\frac{1}{3}$ for all $k \geq 1$, but the k-metachoiceworthiness of a_1 is $\sum_{n=1}^{k} \frac{1}{4^n}$, according to T_2, and is $\sum_{n=1}^{k} \frac{1}{2^{2n-1}}$, according to T_1.

As it happens, $\sum_{n=1}^{k} \frac{1}{4^n} < \frac{1}{3} \forall k$, and $1 - \sum_{n=1}^{k} \frac{1}{2^{2n-1}} > \frac{1}{3} \forall k$; but $\lim_{k\to\infty} \sum_{n=1}^{k} \frac{1}{4^n} = \lim_{k\to\infty} (1 - \sum_{n=1}^{k} \frac{1}{2^{2n-1}}) = \frac{1}{3}$. Thus the subjective choiceworthiness of each act is well-defined and equal to $\frac{1}{3}$, even though arbitrarily many iterations of "k-metatheoretic deliberation" will not rule out the possibility that one act is subjectively more choiceworthy than the other. □

This is not to say that one might not be able to infer acts' relative subjective choiceworthinesses, from one's choiceworthiness distribution and one's hierarchy

distribution, in finite time. Indeed, since the subjective choiceworthiness of each act is well-defined under the above conditions, and is a logical consequence of one's finite choice problem, we know from the completeness of first-order logic that the value of $c(\pi)_i$ can be determined for all i by some finite proof—such as the one given above. The above example serves only to highlight the observation that there may indeed be facts about subjective choiceworthiness that are not captured by the "iterate a few times and hope my uncertainty is more or less resolved" approach.[15]

4.3 Transfinite Hierarchies

Suppose that the k-metatheories to which I assign positive probability are all total and compatible with the Dominance Principle, but that they are not continuous, or that I reject the Analog Principle. It is then possible that my subjective choiceworthiness for some act a_i does not converge, even after infinite steps. That is, though $\lim_{k\to\infty} \min([\vec{c_k}]_i)$ and $\lim_{k\to\infty} \max([\vec{c_k}]_i)$ do both exist (by the fact that the respective sequences in k are monotonic and bounded), $\lim_{k\to\infty} \min([\vec{c_k}]_i) < \lim_{k\to\infty} \max([\vec{c_k}]_i)$. In such a situation, it seems, someone could still claim that there is a "right way" for me to act. Someone could claim that my subjective choiceworthiness for a_i should be the average of $\lim_{k\to\infty} \min([\vec{c_k}]_i)$ and $\lim_{k\to\infty} \max([\vec{c_k}]_i)$, for example.

If I assign positive probability to competing theories of how to act in situations like the above, I must appeal to a theory about how to act in the face of this uncertainty. So our regress extends beyond the natural numbers, into the transfinite ordinals.

The definitions given in Section 2 and Section 3 can generally be reinterpreted so that k (or, κ) is any ordinal, not just any natural number. Two, however, will require slight tweaks:

Definition 4.1 (Definition 2.3, revised)**.** The κ-metachoiceworthiness c_κ of an act a_i, for an agent facing finite choice problem π, is a_i's subjective choiceworthiness for an agent with the same κ'-metachoiceworthiness distribution as that entailed by π for all $\kappa' < \kappa$, but who knows the correct κ-metatheory.

Definition 4.2 (Definition 2.5, revised)**.** A κ-metatheory, applied to a finite set of acts A, is a function $t_\kappa : \mathbb{D}^{\kappa|A|} \to \mathbb{R}^{|A|}$, representing claims about the κ-metachoiceworthiness of the acts in A given (κ-1)-metachoiceworthiness distributions $d_{k'} \in \mathbb{D}^{|A|}$ for all $k' < k$.

Note that this definition of a higher-order metatheory is strictly more general than the original, even with respect to finite k, because it allows k-metatheories to be functions of one's beliefs not only about $(k-1)$-metachoiceworthiness

[15] In particular, I believe it demonstrates a flaw in the claim that fixed-point solutions to the regress problem must take the form either of convergence after finite k or of monotonic decreases in the set of maximally k-choiceworthy acts (as made by, for example, Tarsney (2017), 239).

but about k'-metachoiceworthiness at all lower orders $k' < k$. The following result thus holds, as Theorem 1 does not, regardless of whether we want to allow for this possibility.

We can now state the following:

Theorem 2. *If one assigns positive probability only to a finite set of total κ-metatheories for each ordinal κ, and one accepts the Dominance Principle, then one's subjective choiceworthiness is well-defined over any finite set A of acts.*

Proof: Choose an act a_i. By the Dominance Principle, $\{\min([\vec{c_\kappa}]_i)\}$ and $\{\max([\vec{c_\kappa}]_i)\}$ must be monotonically increasing (decreasing) transfinite sequences indexed by κ. By the Monotone Convergence Theorem, these sequences have limits; let $\{\min([\vec{c_\kappa}]_i)\} \to x$ and $\{\max([\vec{c_\kappa}]_i)\} \to y$. Consider the set $I_i = \cap_\kappa[\min([\vec{c_\kappa}]_i), \max([\vec{c_\kappa}]_i)]$. By (the transfinite case of) the Nested Interval Theorem, I_i cannot be empty. I_i can only be a point (if $x = y$), in which case the subjective choiceworthiness of a_i is well-defined, or a positive-length interval (if $x < y$), in which case the subjective choiceworthiness of a_i is not well-defined. By contradiction, therefore, suppose $x < y$.

Choose $\varepsilon > 0$. Define the interval $G = [x - \varepsilon, x)$, and divide it into the countably infinite partition given by $G_j = [x - \frac{\varepsilon}{2^j}, x - \frac{\varepsilon}{2^{j+1}}), j \geq 0$. For each G_j, choose an ordinal $\gamma : \min([\vec{c_\gamma}]_j) \in G_j$, if such γ exists; skip G_j if no such γ exists. (Such γ must exist for infinitely many G_j; if $\gamma : \min([\vec{c_\gamma}]_j) \in G_j$ existed for only finitely many G_j, $\{\min([\vec{c_\kappa}]_i)\}$ could not converge to x.) We have thus constructed a countably infinite sequence $\Gamma = \{\gamma_j\}$ of ordinals such that $\{\min([\vec{c_{\gamma_j}}]_i)\} \to x$.

Choose $\gamma^* : \gamma^* > \gamma \ \forall \gamma \in \Gamma$.[16] Since $\sup_{\gamma' < \gamma^*} \min([\vec{c_{\gamma'}}]_i) = x < y \leq \inf_{\gamma' < \gamma^*} \max([\vec{c_{\gamma'}}]_i)$, $t_{\gamma^*}(\pi)_i > x$ for all the γ^*-metatheories t_{γ^*} to which I assign positive probability. So $\min([\vec{c_{\gamma^*}}]_i) > x$, a contradiction. □

4.4 The Infectiousness of Stubbornness

Definition 4.3. The Weak Dominance Principle is the principle that

- If $b \geq x \ \forall b \in [\vec{c_\kappa}]_i$, for some κ, then $c(a_i) \geq x$.
- If $b \leq x \ \forall b \in [\vec{c_\kappa}]_i$, for some κ, then $c(a_i) \leq x$.

Let us call a κ-metatheory "compromising" if it is compatible with the (strong) Dominance Principle, and "stubborn" if it is not. Expected Choiceworthiness Theory and risk-weighted variants of it are examples of "compromising" theories. Minimax Theory, according to which an act's metachoiceworthiness is its minimum possible choiceworthiness, is an example of a "stubborn" theory. However, it is compatible with the Weak Dominance Principle.

[16] For every set M of ordinal numbers, there is an ordinal number $\sigma : \sigma > \mu \ \forall \mu \in M$.

Both the theorems above demonstrate that, when all the decision theories (or κ-metatheories) to which I assign positive probability are compromising (along with some other conditions), the range of potential subjective choiceworthiness values for each act shrinks to point. In both cases, this is demonstrated roughly by the fact that, when the range of potential subjective choiceworthiness values for some act is a non-degenerate interval at some order k (or κ), the application of even higher-order metatheories shrinks this interval by increasing its minimum.

One might notice that, strictly speaking, neither of the proofs requires that *all* my decision theories (or κ-metatheories) be compromising. Suppose, for example, that I reject the Dominance Principle, but accept the Weak Dominance Principle. Suppose further that just *one* decision theory (or *one* κ-metatheory for each κ) to which I assign positive probability is "stubborn"—or, that all the stubborn theories to which I assign positive probability are one-sidedly pessimistic (like Minimax) or optimistic (like Maximax) at each order. Then our proofs can go through with only slight modifications. We just need to shrink our interval exclusively from the top or from the bottom, at a given order, to avoid asking concessions of our stubborn theories.

The plausibility of stubborn theories, however, poses two challenges for this project in general.

First, stubborn theories are "infectious": they can determine our behavior regardless how little positive credence we give them. Suppose I am deciding between two acts, a_1 and a_2. I assign probability 0.99 to a state in which a_1 has objective choiceworthiness 1, and probability 0.01 to a state in which a_1 has objective choiceworthiness 0. Act a_2 has objective choiceworthiness 0.0001 in both states. Furthermore, I assign probability 0.99 to Expected Choiceworthiness, and probability 0.01 to Minimax, at every order of the hierarchy. It would be deeply counterintuitive to conclude that, from such a position, a_2 is *more subjectively choiceworthy* than a_1. It would be perhaps even more counterintuitive to conclude that the acts' subjective choiceworthinesses were *equal*, if a_2's objective choiceworthiness were known to be 0. But of course we do reach both conclusions, as repeated applications of my distribution over decision theories bring a_1's higher-order metachoiceworthiness arbitrarily close to 0—according even to the sequence of "Expected Choiceworthiness"-analogous theories.

Second, stubborn theories can clash with each other. If we are going to give some weight to Maximin, at every order, it seems only fair to give some weight to Maximax at every order as well. But if we do, then each act's range of potential subjective choiceworthiness values never shrinks at all; $\min([\vec{c_\kappa}]_i) = \min([\vec{c_0}]_i)$, and $\max([\vec{c_\kappa}]_i) = \max([\vec{c_0}]_i)$, for all κ.

It does not feel as though I can fully rule out stubborn theories. Thus, despite all our progress, I am still left with the original motivating question: how do I so regularly wind up with well-defined subjective choiceworthiness? One encouraging thought is the observation that, though stubbornness is infectious

in one sense, there is a sense in which compromise is infectious as well. For example, suppose I assign positive probability to stubborn κ-metatheories (or even, *only* to stubborn κ-metatheories) at almost all κ, but assign positive probability only to compromising theories at a relatively sparse class of orders—at the limit ordinals, perhaps. (At least for the proof above, we will still need our class Γ of "all-compromising" orders to be such that, for every set M of ordinal numbers, there is an ordinal number $\gamma \in \Gamma : \gamma > \mu \, \forall \mu \in M$.) Then, even though there is a sense in which I believe in stubborn theories "almost everywhere" up the hierarchy, the scattered all-compromising orders will still force my subjective choiceworthiness range for each act down to a point. Similar reasoning applies to the case of merely natural hierarchies. My hierarchy distribution can handle a lot of stubbornness; as long as an all-compromising order comes along every now and then to shrink my subjective choiceworthiness range for each act, there are reasonable conditions under which subjective choiceworthiness will generally be well-defined.

4.5 Rescaling

MacAskill (2014) offers the following example of undefined subjective choiceworthiness.

Suppose an agent faces a choice problem:

Order 0

	s_1 (Pr. $\frac{18}{23}$)	s_2 (Pr. $\frac{5}{23}$)
a_1	0	4
a_2	1	0

She must choose among acts $A = \{a_1, a_2\}$. She assigns probability $\frac{18}{23}$ to a state s_1 in which a_1 has objective choiceworthiness 0 and a_2 has objective choiceworthiness 1, and probability $\frac{5}{23}$ to a state s_2 in which $c_0(a_1) = 4$ and $c_0(a_2) = 0$.

In evaluating the acts at order 1, she assigns probability $\frac{18}{23}$ to Expected Choiceworthiness Theory (t_1), and probability $\frac{5}{23}$ to "Square Root Theory" (t_2), according to which an act's 1-metachoiceworthiness is the expectation of the square root of the difference between its objective choiceworthiness and the objective choiceworthiness of the least objectively choiceworthy act in A. Thus:

Order 1

	t_1 (Pr. $\frac{18}{23}$)	t_2 (Pr. $\frac{5}{23}$)
a_1	$\frac{20}{23}$	$\frac{10}{23}$
a_2	$\frac{18}{23}$	$\frac{18}{23}$

On MacAskill's reading of the problem, "transformations of each individual choice-worthiness function by an absolute value are permissible, and transfor-

mations of all choice-worthiness functions by a multiplying factor are permissible". Thus he produces

Order 1, rescaled

	s_1 (Pr. $\frac{18}{23}$)	s_2 (Pr. $\frac{5}{23}$)
a_1	1	0
a_2	0	4

As we can see, after rescaling, the 1-metachoiceworthiness distribution of a_1 is precisely what the 0-metachoiceworthiness distribution of a_2 had been, and the 1-metachoiceworthiness distribution of a_2 is precisely what the 0-metachoiceworthiness distribution of a_1 had been. Therefore, if we obey the Analog Principle—that is, if our distribution over k-metatheories is the same at every order—and if we rescale after every step, our k-metachoiceworthiness distribution for the acts will flip forever between that of "Order 0" and that of "Order 1, rescaled", without converging.

To my mind, however, k-metachoiceworthiness claims are characterized by the property that, if one believes them, they define one's subjective choiceworthiness. If an agent faces empirical uncertainty over two acts' objective choiceworthiness as represented above, we want to say that, in the event that she learns the truth of s_2, act a_1's subjective choiceworthiness for her is 4. In precisely the same language, I think, we want to say that in the event that she learns the truth of Expected Choiceworthiness Theory (but does not learn the true state), a_1's subjective choiceworthiness for her is $\frac{20}{23}$—and likewise all up the hierarchy. To keep these claims "in line", the framework of Section 2 permits real-valued representations of the subjective choiceworthiness values and k-metachoiceworthiness distributions associated with a given choice problem to be rescaled only in conjunction, not independently.

Furthermore, if this is the right way to think about k-metachoiceworthiness, then Square Root Theory (SR) is, as stated, incoherent. SR does not specify which real-valued representations of objective choiceworthiness to use as inputs, so its claims should be independent to rescaling. But they are not. Using our agent's 0-metachoiceworthiness distribution as represented above, SR claims that the 1-metachoiceworthiness of a_1 is $\frac{20}{23}$, and EC claims that the 1-metachoiceworthiness of a_1 is lower (just $\frac{10}{23}$). But if we had represented her 0-metachoiceworthiness distribution differently,

Order 0

	s_1 (Pr. $\frac{18}{23}$)	s_2 (Pr. $\frac{5}{23}$)
a_1	0	1
a_2	$\frac{1}{4}$	0

we would conclude that EC claims that the 1-metachoiceworthiness of A is $\frac{5}{23}$ for her, and that SR claims the same.

To ensure that an act's true k-metachoiceworthiness for an agent be independent of the scale she arbitrarily uses to represent her (k-1)-metachoiceworthiness distribution, all our k-metatheories have to be "affine" (unique up to affine transformation). Though this condition closes the door to "Square Root Theory", it permits a wide array of other risk-averse theories, including Buchak's REU Theory and the risk-averse theory presented in Proposition 4.1.

5 Applications to Moral Uncertainty

If I assign positive probability only a finite set of total, cardinal, comparable moral theories—or, if I at least know the right way to represent all my moral theories' choiceworthiness claims on the same cardinal scale—then the results above can be applied almost directly to my moral choice problems under empirical certainty.

Suppose I am certain about the state of the world. I then simply have to swap out our language about objective choiceworthiness being "my utility function, contingent on the true state of the world" for language about objective choiceworthiness being "axiological value, contingent on the true moral theory", and Sections 2–4 apply to cases of moral uncertainty, under empirical certainty, in full.[17] If I face both empirical uncertainty and moral uncertainty, my situation is more complex. One approach would be for me to take "objective choiceworthiness" to be a function of both the true state and the true moral theory, to consider my probability distribution over {states} × {moral theories}, and then to apply my hierarchy distribution. Another approach, however, would be for me first to work out the subjective choiceworthiness of each act, conditional on each state, in light of my distribution over moral theories, and then to apply my hierarchy distribution a second time, to work out the subjective choiceworthiness of each act in light of my distribution over states. A third approach, symmetrical to the second, would be for me first to work out the subjective choiceworthiness of each act, conditional on each moral

[17]We must also assume that my moral theories make claims only about the axiological value of each state of the world. That is, we must say that, for example, among varieties of utilitarianism, I assign positive probability only to those that claim that an act's objective choiceworthiness is (something along the lines of) its objective impact on total utility. In particular, we must suppose that my moral theories do *not* make claims about how I ought to deal with uncertainty; that I assign no positive probability to a variety of utilitarianism that claims that an act's objective choiceworthiness is *my expectation of* its impact on total utility, say. Utilitarianism is so often described as the idea that we ought to maximize the world's *expected utility*—see Parfit (1984), 25-26, for instance—that one might easily come to believe that Expected Choiceworthiness is the only way utilitarians are allowed to deal with uncertainty. In this context, however, we should be careful to separate the unique moral claim of utilitarianism (that value is identified with utility) from the independent decision-theoretic claim (that one ought to maximize expected value). Moral theories that explicitly incorporate such decision-theoretic claims may also be interesting to consider in light of the regress problem, but we will not discuss them here.

theory, in light of my distribution over states, and then to apply my hierarchy distribution a second time, to work out the subjective choiceworthiness of each act in light of my distribution over moral theories. (These approaches have the disadvantage that they would not be able to account for any dependence between my distribution over states and my over moral theories. They have the advantage, however, that they would be able to account for the possibility that my hierarchy distribution over ways of dealing with moral uncertainty differs from my hierarchy distribution over ways of dealing with empirical uncertainty.) And other conceivable approaches abound.

Unfortunately, these approaches will not necessarily all yield the same subjective choiceworthiness values, or even the same act recommendations—not even under decision-theoretic certainty, and not even when I believe that the same theory should be used in the face of emprical uncertainty as in the face of moral uncertainty. Consider the following situation. I assign positive probability to a set of moral theories $M = \{m_1, m_2, m_3\}$ and to a set of states $S = \{s_1, s_2, s_3\}$. I have two feasible acts, a_1 and a_2. Their objective choiceworthinesses, conditional on each state and moral theory, are as follows:

a_1

	m_1	m_2	m_3
s_1	0	2	4
s_2	2	4	6
s_3	4	6	8

a_2

	m_1	m_2	m_3
s_1	3	3	3
s_1	3	3	3
s_1	3	3	3

Furthermore, I am certain that an act's k-metachoiceworthiness is its second-lowest-possible (k-1)-metachoiceworthiness. If I apply this decision theory to my uncertainty over {states} × {moral theories}, I get $c(a_1) = 2$ and $c(a_2) = 3$, so $a_2 \succ a_1$. However, if I apply this decision theory first over states (conditional on each moral theory) and then over moral theories—or, first over moral theories (conditional on each state) and then over states—then I get $c(a_1) = 4$ and $c(a_2) = 3$, so $a_1 \succ a_2$. These complications only worsen when $|M| \neq |S|$, in which case even the "same" decision theory can aggregate across moral theories and across states arbitrarily differently.

There is another way in which decision-theoretic uncertainty can interact with moral uncertainty. It is often argued that morality requires us to make decisions as if from behind a "veil of ignorance" about our own identity among those affected by our actions. If so, the moral choiceworthiness of an act depends directly on its decision-theoretic 1-metachoiceworthiness. Suppose that I ought to act toward a group as if my identity is, in probability, distributed uniformly over the group. Then, if Expected Choiceworthiness is the correct decision theory, the Veil of Ignorance argument points toward classical utilitarianism as the correct moral theory; if Minimax is the correct decision theory, toward Rawls's "maximin criterion"; if some risk-weighted theory is the correct decision

theory, toward the corresponding version of prioritarianism; and so on.[18] But we will not explore this interaction further here.

Finally, the above thoughts about how to integrate the results of Sections 2–4 into situations of moral uncertainty can apply straightforwardly to normative uncertainty in other domains, so long as one assigns positive probability finite set of theories which are in some analogous sense total, cardinal, and intertheoretically comparable. But we will not explore such applications further here.

6 Conclusion

We are often uncertain about the moral and decision-theoretic norms which we believe should guide our behavior. Even when these norms conflict, however, we often have a subjective understanding of whether some act would be rationally or morally permissible for us, from our position of normative uncertainty. "Uncertaintism" might be understood as the project of unraveling how this uncertainty translates into the subjective choiceworthiness on which we ultimately feel justified in acting.

When the uncertaintist tries to specify any particular mechanism for translating the uncertainty over choiceworthiness into an appropriate characterization of subjective choiceworthiness, however, we find that, just as we are not certain of our acts' objective choiceworthinesses, we are not certain of his proposed mechanism either. Nor are we certain about how to how to deal with our uncertainty about such a mechanism. Indeed, our certainty about subjective choiceworthiness seems to stand strangely on its own. In general, when we try to ground our certainty about subjective choiceworthiness in metanormative certainty at some order, we find that the hoped-for ground of certainty does not exist. For some, as cited above, this "possibility of normative uncertainty all the way up makes the uncertaintist project look pointless".

The results presented here demonstrate that, as stated, the quoted worry is not justified. We can reliably have well-defined subjective choiceworthiness without being certain about the correct first-order normative theory or about any higher-order metatheory. We only have to commit to a weaker family of assumptions, such as the Dominance Principle. This observation should lend the "uncertainist project" at least some hope.

But commitment to these weaker assumptions may still be a strong requirement. Certainty about them may never actually obtain, or may obtain only rarely. Ultimately, therefore, it is up to the reader to judge whether this theorizing sheds any light on more realistic cases of normative uncertainty.

In any event, this preliminary investigation has uncovered one class of "fixed-point" solutions to the regress problem. Even if doubts can be cast

[18] Further discussion of the implications of risk-weighted expected utility theory for decisions made on behalf of groups, rather than individuals, can be found in Buchak (2013) 167-8.

on the constraints here imposed in the process, I hope these results have encouraged the reader that solutions along similar lines might more generally be found.

References

[1] Allais, M. (1953). "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole Americaine", *Econometrica* 21(4): 503-546.

[2] Barry, C. and P. Tomlin (2016). "Moral uncertainty and permissibility: Evaluating Option Sets", *Canadian Journal of Philosophy* 46(6): 898-923.

[3] Bostrom, N. (2016). "Pascal's Mugging", *Analysis* 69(3): 443-445.

[4] Buchak, L. (2013). *Risk and Rationality* (Oxford: Oxford University Press).

[5] Lin, H. (2014). "On the Regress Problem of Deciding How to Decide", *Synthese* 191(4): 661-670.

[6] Lipman, B. (1991). "How to Decide How to Decide How to...: Modeling Limited Rationality", *Econometrica* 59(4): 1105-1125.

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

[8] MacAskill, W. (2013). "The Infectiousness of Nihilism", *Ethics* 123(3): 508-520.

[9] MacAskill, W. (2014). "Normative Uncertainty". DPhil thesis, University of Oxford.

[10] MacAskill, W. (2016). "Normative Uncertainty as a Voting Problem", *Mind* 125(500): 967-1004.

[11] MacAskill, W. (2016). "Smokers, Psychos, and Decision-Theoretic Uncertainty", *The Journal of Philosophy* 113(9): 425-445.

[12] von Neumann, J. and O. Morgenstern (1953). *Theory of Games and Economic Behavior* (Princeton: Princeton University Press).

[13] Parfit, D. (1984). *Reasons and Persons* (Oxford: Oxford University Press).

[14] Pittard, J. and A. Worsnip (2017). "Metanormative Contextualism and Normative Uncertainty", *Mind* 126(1): 155-193.

[15] Rawls, J. (1971). *A Theory of Justice* (Cambridge, MA: Harvard University Press).

[16] Savage, L. (1954). *The Foundations of Statistics* (New York: Wiley).

[17] Sepielli, A. (2009). "What to Do When You Don't Know What to Do". In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics*, vol. X (Oxford: Oxford University Press).

[18] Sepielli, A. (2013). "What to Do When You Don't Know What to Do When You Don't Know What to Do...", *Noûs* 47(1): 521-544.

[19] Sepielli, A. (2017). "How Moral Uncertaintism Can Be Both True & Interesting". In Mark Timmons (ed.), *Oxford Studies in Normative Ethics*, vol. VII (Oxford: Oxford University Press).

[20] Smith, H. (2010). "Subjective Rightness", *Social Philosophy and Policy* 27(2): 64-110.

[21] Tarsney, C. (2017). "Rationality and Moral Risk: A Moderate Defense of Hedging". PhD thesis, University of Maryland, College Park.

[22] Weatherson, B. (2014). "Running Risks Morally", *Philosophical Studies* 167(1): 141-163.