Moral uncertainty

Morality vs related concepts

Cross-posted to the EA Forum.

How can you know I'm talking about morality (aka ethics), rather than something else, when I say that I "should" do something, that humanity "ought" to take certain actions, or that something is "good"? What are the borderlines and distinctions between morality and the various potential "something else"s? How do they overlap and interrelate?

In this post, I try to collect together and summarise philosophical concepts that are relevant to the above questions.[1] I hope this will benefit readers by introducing them to some thought-clarifying conceptual distinctions they may not have been aware of, as well as terms and links they can use to find more relevant info. In another post, I similarly discuss how *moral uncertainty* differs from and overlaps with related concepts.

Epistemic status: The concepts covered here are broad, fuzzy, and overlap in various ways, making definitions and distinctions between them almost inevitably debatable. Additionally, I'm not an expert in these topics; indeed, I expect many readers to know more than me about at least some of them, and one reason I wrote this was to help me clarify my own understandings. I'd appreciate feedback or comments in relation to any mistakes, poor phrasings, etc. (and just in general!).

*Also note that my intention here is mostly to summarise existing ideas, rather than to provide **original** ideas or analysis.*

Normativity

A *normative* statement is any statement related to what one *should* do, what one *ought* to do, which of two things are *better*, or similar. "Something is said by philosophers to have 'normativity' when it entails that some action, attitude or mental state of some other kind is justified, an action one ought to do or a state one ought to be in" (Darwall). Normativity is thus the overarching category (*superset*) of which things like morality, prudence (in the sense explained below), and arguably rationality are just subsets.

This matches the usage of "normative" in economics, where normative claims relate to "what ought to be" (e.g., "The government should increase its spending"), while *positive* claims relate to "what is" (including predictions, such as what effects an increase in government spending may have). In linguistics, the equivalent distinction is between *prescriptive* approaches (involving normative claims about "better" or "correct" uses of language) and *descriptive* approaches (which are about how language *is* used).

Prudence

Prudence essentially refers to the subset of normativity that has to do with *one's own* self-interest, happiness, or wellbeing (see here and here). This contrasts with morality,

which may include but isn't limited to one's self-interest (except perhaps for egoist moral theories).

For example (based on MacAskill p. 41), we may have *moral reasons* to give money to GiveWell-recommended charities, but *prudential reasons* to spend the money on ourselves, and both sets of reasons are "normatively relevant" considerations.

(The rest of this section is my own analysis, and may be mistaken.)

I would expect that the significance of prudential reasons, and how they relate to moral reasons, would differ depending on the moral theories one is considering (e.g., depending on which moral theories one has some belief in). Considering moral and prudential reasons separately *does* seem to make sense in relation to moral theories that don't *precisely* mandate *specific* behaviours; for example, moral theories that simply forbid certain behaviours (e.g., violating people's rights) while otherwise letting one choose from a range of options (e.g., donating to charity or not).[2]

In contrast, "maximising" moral theories like classical utilitarianism claim that the *only* action one is permitted to take is the *very best* action, leaving no room for choosing the "prudentially best" action out of a range of "morally acceptable" actions. Thus, in relation to maximising theories, it seems like keeping track of prudential reasons in addition to moral reasons, and sometimes acting based on prudential rather than moral reasons, would mean that one is effectively either:

- using a modified version of the maximising moral theory (rather than the theory itself), or
- acting as if "morally uncertain" between the maximising moral theory and a "moral theory" in which prudence is seen as "intrinsically valuable".

Either way, the boundary between prudence and morality seems to become fuzzier or less meaningful in such cases.[3]

(Instrumental) Rationality

(This section is sort-of my own analysis, and may be mistaken or use terms in unusual ways.)

Bykvist (2017):

> Rationality, in one important sense at least, has to do with what one should do or intend, given one's beliefs and preferences. This is the kind of rationality that decision theory often is seen as invoking. It can be spelled out in different ways. One is to see it as a matter of coherence: It is rational to do or intend what coheres with one's beliefs and preferences (Broome, 2013; for a critic, see Arpaly, 2000).

Using this definition, it seems to me that:

- Rationality can be considered a subset of normativity in which the "should" statements, "ought" statements, etc. follow in a systematic way from one's beliefs and preferences.
- Whether a "should" statement, "ought" statement, etc. is rational is unrelated to the balance of moral or prudential reasons involved. E.g., what I "rationally

> should" do relates only to morality and not prudence if *my preferences* relate only to morality and not prudence, and vice versa. (And situations in between those extremes are also possible, of course).[4]

For example, the statement "Rationally speaking, I should buy a Ferrari" is true if (a) I believe that doing so will result in me possessing a Ferrari, and (b) I value that outcome more than I value continuing to have that money. And it doesn't matter whether the *reason* I value that outcome is:

- Prudential: based on self-interest;
- Moral: e.g., I'm a utilitarian who believes that the best way I can use my money to increase universe-wide utility is to buy myself a Ferrari (perhaps it looks *really* red and shiny and my biases are self-serving the hell out of me);
- Some mixture of the two.

Epistemic rationality

Note that that discussion focused on *instrumental* rationality, but the same basic points could be made in relation to *epistemic* rationality, given that epistemic rationality itself "can be seen as a form of instrumental rationality in which knowledge and truth are goals in themselves" (LW Wiki).

For example, I could say that, from the perspective of epistemic rationality, I "shouldn't" believe that buying that Ferrari will create more utility in expectation than donating the same money to AMF would. This is because holding that belief won't help me meet the goal of having accurate beliefs.

Whether and how this relates to morality would depend on whether the "deeper reasons" why I prefer to have accurate beliefs (assuming I do indeed have that preference) are prudential, moral, or mixed.[5]

Subjective vs objective

Subjective normativity relates to what one should do *based on what one believes*, whereas *objective* normativity relates to what one "actually" should do (i.e., based on the true state of affairs). Greaves and Cotton-Barratt illustrate this distinction with the following example:

> Suppose Alice packs the waterproofs but, as the day turns out, it does not rain. Does it follow that Alice made the wrong decision? In one (objective) sense of "wrong", yes: thanks to that decision, she experienced the mild but unnecessary inconvenience of carrying bulky raingear around all day. But in a second (more subjective) sense, clearly it need not follow that the decision was wrong: if the probability of rain was sufficiently high and Alice sufficiently dislikes getting wet, her decision could easily be the appropriate one to make given her state of ignorance about how the weather would in fact turn out. Normative theories of decision-making under uncertainty aim to capture this second, more subjective, type of evaluation; the standard such account is expected utility theory.[6][7]

This distinction can be applied to each subtype of normativity (i.e., morality, prudence, etc.).

(I discuss this distinction further in my post Moral uncertainty vs related concepts.)

Axiology

The term *axiology* is used in different ways in different ways, but the definition we'll focus on here is from the Stanford Encyclopaedia of Philosophy:

> Traditional axiology seeks to investigate what things are good, how good they are, and how their goodness is related to one another. Whatever we take the "primary bearers" of value to be, one of the central questions of traditional axiology is that of what stuffs are good: what is of value.

The same article also states: "For instance, a traditional question of axiology concerns whether the objects of value are subjective psychological states, or objective states of the world."

Axiology (in this sense) is essentially one aspect of morality/ethics. For example, classical utilitarianism combines:

- the principle that one must take actions which will lead to the outcome with the *highest possible level of value*, rather than just doing things that lead to "good enough" outcomes, or just avoiding violating people's rights
- the axiology that "well-being" is what has intrinsic value

The axiology itself is not a moral theory, but plays a key role in that moral theory.

Thus, one can't have an axiological "should" statement, but one's axiology may *influence/inform* one's moral "should" statements.

Decision theory

(This section is sort-of my own commentary, may be mistaken, and may accidentally deviate from standard uses of terms.)

It seems to me that the way to fit decision theories into this picture is to say that one must *add* a decision theory to one of the "sources of normativity" listed above (e.g., morality) in order to get some form of normative (e.g., moral) statements. However, a decision theory can't "generate" a normative statement by itself.

For example, suppose that I have a moral preference for having more money rather than less, all other things held constant (because I wish to donate it to cost-effective causes). By itself, this can't tell me whether I "should" one-box or two-box in Newcomb's problem. But once I specify my decision theory, I can say whether I "should" one-box or two-box. E.g., if I'm a causal decision theorist, I "should" two-box.

But if I knew *only* that I was a causal decision theorist, it would still be possible that I "should" one-box, if for some reason I preferred to have less money. Thus, as stated, we must specify (or assume) *both* a set of preferences *and* a decision theory in order to arrive at normative statements.

Metaethics

> While normative ethics addresses such questions as "What should I do?", evaluating specific practices and principles of action, meta-ethics addresses questions such as "What is goodness?" and "How can we tell what *is* good from what is bad?", seeking to understand the nature of ethical properties and evaluations. (Wikipedia)

Thus, metaethics is not *directly* normative at all; it isn't about *making* "should", "ought", "better than", or similar statements. Instead, it's about understanding the "nature" of (the moral subset of) such statements, "where they come from", and other such fun/spooky/nonsense/incredibly important matters.

Metanormativity

Metanormativity relates to the "norms that govern how one ought to act that take into account one's fundamental normative uncertainty". Normative uncertainty, in turn, is essentially a generalisation of moral uncertainty that can also account for (uncertainty about) prudential reasons. I will thus discuss the topic of metanormativity in my next post, on Moral uncertainty vs related concepts.

As stated earlier, I hope this usefully added to/clarified the concepts in your mental toolkit, and I'd welcome any feedback or comments!

(In particular, if you think there's another concept whose overlaps with/distinctions from "morality" are worth highlighting, either let me know to add it, or just go ahead and explain it in the comments yourself.)

---

1. This post *won't* attempt to discuss specific debates within metaethics, such as whether or not there are "objective moral facts", and, if there are, whether or not these facts are "natural". Very loosely speaking, I'm not trying to answer questions about what morality itself *actually is*, but rather about the *overlaps and distinctions* between what morality is *meant to be about* and what other topics that involve "should" and "ought" statements are *meant to be about*. ↩

2. Considering moral and prudential reasons separately also seems to make sense for moral theories which see *supererogation* as possible; that is, theories which see some acts as "morally good although not (strictly) required" (SEP). If we only believe in such theories, we may often find ourselves deciding between one act that's morally "good enough" and another (supererogatory) act that's morally better but prudentially worse. (E.g., perhaps, occasionally donating small sums to whichever charity strikes one's fancy, vs donating 10% of one's income to charities recommended by Animal Charity Evaluators.) ↩

3. The boundary seems even fuzzier when you also consider that many moral theories, such as classical or preference utilitarianism, *already* consider one's own happiness or preferences to be morally relevant. This arguably makes *also* considering "prudential reasons" look like simply "double-counting" one's self-interest, or giving it additional "weight". ↩

4. If we instead used a definition of rationality in which preferences must only be based on self-interest, then I believe rationality would become a subset of *prudence specifically*, rather than of normativity as a whole. It would still be the

case that the distinctive feature of rational "should" statements is that they follow in a systematic way from one's beliefs and preferences. ↩

5. Somewhat relevantly, Darwall writes: "Epistemology has an irreducibly normative aspect, in so far as it is concerned with norms for belief." ↩

6. We could further divide subjective normativity up into, roughly, "what one should do based on what one *actually* believes" and "what one should do based on what it *would be reasonable* for one to believe". The following quote is relevant (though doesn't directly address that *exact* distinction):

   > Before moving on, we should distinguish subjective credences, that is, degrees of belief, from epistemic credences, that is, the degree of belief that one is epistemically justified in having, given one's evidence. When I use the term 'credence' I refer to epistemic credences (though much of my discussion could be applied to a parallel discussion involving subjective credences); when I want to refer to subjective credences I use the term 'degrees of belief'.
   >
   > The reason for this is that appropriateness seems to have some sort of normative force: if it is most appropriate for someone to do something, it seems that, other things being equal, they ought, in the relevant sense of 'ought', to do it. But people can have crazy beliefs: a psychopath might think that a killing spree is the most moral thing to do. But there's no sense in which the psychopath ought to go on a killing spree: rather, he ought to revise his beliefs. We can only capture that idea if we talk about epistemic credences, rather than degrees of belief.

   (I found that quote in this comment, where it's attributed to Will MacAskill's BPhil thesis. Unfortunately, I can't seem to access the thesis, including via Wayback Machine.) ↩

7. It also seems to me that this "subjective vs objective" distinction is somewhat related to, but distinct from, *ex ante* vs *ex post* thinking. ↩

Moral uncertainty vs related concepts

Overview

> How important is the well-being of non-human animals compared with the well-being of humans?
>
> How much should we spend on helping strangers in need?
>
> How much should we care about future generations?
>
> How should we weigh reasons of autonomy and respect against reasons of benevolence?
>
> Few could honestly say that they are fully certain about the answers to these pressing moral questions. Part of the reason we feel less than fully certain about the answers has to do with uncertainty about **empirical** facts. We are uncertain about whether fish can feel pain, whether we can really help strangers far away, or what we could do for people in the far future. However, sometimes, the uncertainty is fundamentally **moral**. [...] Even if were to come to know all the relevant non-normative facts, we could still waver about whether it is right to kill an animal for a very small benefit for a human, whether we have strong duties to help strangers in need, and whether future people matter as much as current ones. Fundamental moral uncertainty can also be more general as when we are uncertain about whether a certain moral theory is correct. (Bykvist; emphasis added)[1]

I consider the above quote a great starting point for understanding what moral uncertainty is; it gives clear examples of moral uncertainties, and contrasts these with related empirical uncertainties. From what I've seen, a lot of academic work on moral uncertainty essentially opens with something like the above, then notes that the rational approach to decision-making under *empirical* uncertainty is typically considered to be expected utility theory, then discusses various approaches for decision-making under *moral* uncertainty.

That's fair enough, as no one article can cover everything, but it also leaves open some major questions about **what moral uncertainty actually *is***.[2] These include:

1. How, more precisely, can we *draw lines between moral and empirical uncertainty*?
2. What are the overlaps and distinctions between moral uncertainty and other related concepts, such as *normative*, *metanormative*, *decision-theoretic*, and *metaethical uncertainty*, as well as *value pluralism*?
    - My prior post answers similar questions about how *morality* overlaps with and differs from related concepts, and may be worth reading before this one.
3. Is what we "ought to do" under moral uncertainty an *objective* or *subjective* matter?

4. Is what we "ought to do" under moral uncertainty a matter of *rationality* or *morality*?

5. Are we talking about "*moral risk*" or about "*moral (Knightian) uncertainty*" (if such a distinction is truly meaningful)?

6. What *"types" of moral uncertainty* are meaningful for *moral antirealists and/or subjectivists*?[3]

In this post, I collect and summarise ideas from academic philosophy and the LessWrong and EA communities in an attempt to answer the first two of the above questions (or to at least clarify what the questions *mean*, and what the *most plausible* answers are). My next few posts will do the same for the remaining questions.

I hope this will benefit readers by facilitating clearer thinking and discussion. For example, a better understanding of the nature and types of moral uncertainty may aid in determining how to *resolve* (i.e., reduce or clarify) one's uncertainty, which I'll discuss two posts from now. (How to make decisions *given* moral uncertainty is discussed later in this sequence.)

Epistemic status: The concepts covered here are broad, fuzzy, and overlap in various ways, making definitions and distinctions between them almost inevitably debatable. Additionally, I'm not an expert in these topics (though I have now spent a couple weeks mostly reading about them). I've tried to mostly collect, summarise, and synthesise ***existing*** *ideas. I'd appreciate feedback or comments in relation to any mistakes, unclear phrasings, etc. (and just in general!).*

Empirical uncertainty

In the quote at the start of this post, Bykvist (the author) seemed to imply that it was easy to identify which uncertainties in that example were empirical and which were moral. However, in many cases, the lines aren't so clear. This is perhaps most obvious with regards to, as Christian Tarsney puts it:

> Certain cases of uncertainty about moral considerability (or moral status more generally) [which] turn on *metaphysical* uncertainties that resist easy classification as empirical or moral.
>
> [For example,] In the abortion debate, uncertainty about when in the course of development the fetus/infant comes to count as a *person* is neither straightforwardly empirical nor straightforwardly moral. Likewise for uncertainty in Catholic moral theology about the time of ensoulment, the moment between conception and birth at which God endows the fetus with a human soul [...]. Nevertheless, it seems strange to regard these uncertainties as fundamentally different from more clearly empirical uncertainties about the moral status of the developing fetus (e.g., uncertainty about where in the gestation process complex mental activity, self-awareness, or the capacity to experience pain first emerge), or from more clearly moral uncertainties (e.g., uncertainty, given a certainty that the fetus is a person, whether it is permissible to cause the death of such a person when doing so will result in more total happiness and less total suffering).[4]

And there are also other types of cases in which **it seems hard to find clear, non-arbitrary lines between moral and empirical uncertainties** (some of which

Tarsney [p. 140-146] also discusses).[5] Altogether, I expect drawing such lines will quite often be difficult.

Fortunately, we may not actually *need* to draw such lines anyway. In fact, as I discuss in my post on *making decisions* under both moral and empirical uncertainty, many approaches for handling moral uncertainty were consciously designed by analogy to approaches for handling empirical uncertainty, and it seems to me that they can easily be extended to handle both moral and empirical uncertainty, without having to distinguish between those "types" of uncertainty.[6][7]

The situation is a little less clear when it comes to *resolving* one's uncertainty (rather than just making decisions *given* uncertainty). It seems at first glance that you might need to investigate different "types" of uncertainty in different ways. For example, if I'm uncertain whether fish react to pain in a certain way, I might need to read studies about that, whereas if I'm uncertain what "moral status" fish deserve (even assuming that I know all the relevant empirical facts), then I might need to engage in moral reflection. However, it seems to me that the key difference in such examples is *what the uncertainties are actually* ***about***, rather than specifically whether *a given uncertainty* should be classified as "moral" or "empirical".

(It's also worth quickly noting that the topic of "cluelessness" is only about *empirical* uncertainty - specifically, uncertainty regarding the consequences that one's actions will have. Cluelessness thus won't be addressed in my posts on moral uncertainty, although I do plan to later write about it separately.)

Normative uncertainty

As I noted in my prior post:

> A *normative* statement is any statement related to what one *should* do, what one *ought* to do, which of two things are *better*, or similar. [...] Normativity is thus the overarching category (*superset*) of which things like morality, prudence [essentially meaning the part of normativity that has to do with one's own self-interest, happiness, or wellbeing], and arguably rationality are just subsets.

In the same way, ***normative uncertainty* is a broader concept, of which *moral uncertainty* is just one component**. Other components could include:

- prudential uncertainty
- decision-theoretic uncertainty (covered below)
- metaethical uncertainty (also covered below) - although perhaps it'd make more sense to see metaethical uncertainty as instead just feeding into one's moral uncertainty

Despite this, academic sources seem to commonly either:

- focus only on moral uncertainty, or
- state or imply that essentially the same approaches for decision-making will work for both moral uncertainty in particular and normative uncertainty in general (which seems to me a fairly reasonable assumption).

On this matter, Tarsney writes:

> Fundamentally, the topic of the coming chapters will be the problem of *normative* uncertainty, which can be roughly characterized as uncertainty about one's objective reasons that is not a result of some underlying empirical uncertainty (uncertainty about the state of concretia). However, I will confine myself almost exclusively to questions about *moral* uncertainty: uncertainty about one's objective *moral* reasons that is not a result of etc etc. This is in part merely a matter of vocabulary: "moral uncertainty" is a bit less cumbersome than "normative uncertainty," a consideration that bears some weight when the chosen expression must occur dozens of times per chapter. It is also in part because the vast majority of the literature on normative uncertainty deals specifically with moral uncertainty, and because moral uncertainty provides more than enough difficult problems and interesting examples, so that there is no need to venture outside the moral domain.
>
> Additionally, however, focusing on moral uncertainty is a useful simplification that allows us to avoid difficult questions about the relationship between moral and non-moral reasons (though I am hopeful that the theoretical framework I develop can be applied straightforwardly to normative uncertainties of a non-moral kind). For myself, I have no taste for the moral/non-moral distinction: To put it as crudely and polemically as possible, it seems to me that all objective reasons are moral reasons. But this view depends on substantive normative ethical commitments that it is well beyond the scope of this dissertation to defend. [...]
>
> If one does think that all reasons are moral reasons, or that moral reasons always override non-moral reasons, then a complete account of how agents ought to act under moral uncertainty can be given without any discussion of non-moral reasons (Lockhart, 2000, p. 16). To the extent that one does not share either of these assumptions, theories of choice under moral uncertainty must generally be qualified with "insofar as there are no relevant non-moral considerations."

Somewhat similarly, this sequence will nominally focus on moral uncertainty, even though:

- some of the work I'm drawing on was nominally focused on normative uncertainty (e.g., Will MacAskill's thesis)
- I intend most of what I say to be fairly easily generalisable to normative uncertainty more broadly.

Metanormative uncertainty

In MacAskill's thesis, he writes that *metanormativism* is "the view that there are second-order norms that govern action that are relative to a decision-maker's uncertainty about first-order normative claims. [...] The central metanormative question is [...] about which *option* it's appropriate to choose [when a decision-maker is uncertain about which first-order normative theory to believe in]". MacAskill goes on to write:

> A note on terminology: Metanormativism isn't *about* normativity, in the way that meta-ethics is about ethics, or that a meta-language is about a language. Rather, 'meta' is used in the sense of 'over' or 'beyond'

In essence, ***metanormativism* focuses on what metanormative theories (or "approaches") should be used for making decisions under *normative***

uncertainty.

We can therefore imagine being *metanormatively uncertain*: uncertain about *what metanormative theories to use* for making decisions under normative uncertainty. For example:

- You're *normatively* uncertain if you see multiple ("first-order") moral theories as possible and these give conflicting suggestions.
- You're _meta_normatively uncertain if you're also unsure whether the best approach for deciding what to do given this uncertainty is the "My Favourite Theory" approach or the "Maximising Expected Choice-worthiness" approach (both of which are explained later in this sequence).

This leads inevitably to the following thought:

> It seems that, just as we can suffer [first-order] normative uncertainty, we can suffer [second-order] metanormative uncertainty as well: we can assign positive probability to conflicting [second-order] metanormative theories. [Third-order] Metametanormative theories, then, are collections of claims about how we ought to act in the face of [second-order] 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. (Philip Trammell)

I refer readers interested in this possibility of infinite regress - and potential solutions or reasons not to worry - to Trammell, Tarsney, and MacAskill (p. 217-219). (I won't discuss those matters further here, and I haven't properly read those Trammell or Tarsney papers myself.)

Decision-theoretic uncertainty

(Readers who are unfamiliar with the topic of decision theories may wish to read up on that first, or to skip this section.)

MacAskill writes:

> Given the trenchant disagreement between intelligent and well-informed philosophers, it seems highly plausible that one should not be certain in either causal or evidential decision theory. In light of this fact, Robert Nozick briefly raised an interesting idea: that perhaps one should take decision-theoretic uncertainty into account in one's decision-making.

This is precisely analogous to taking uncertainty about first-order moral theories into account in decision-making. Thus, ***decision-theoretic uncertainty* is just another type of normative uncertainty**. Furthermore, arguably, it can be handled using the same sorts of "metanormative theories" suggested for handling moral uncertainty (which are discussed later in this sequence).

Chapter 6 of MacAskill's thesis is dedicated to discussion of this matter, and I refer interested readers there. For example, he writes:

> metanormativism about decision theory [is] the idea that there is an important sense of 'ought' (though certainly not the only sense of 'ought') according which a decision-maker ought to take decision-theoretic uncertainty into account. I call any metanormative theory that takes decision-theoretic uncertainty into account a type of *meta decision theory* [- in] contrast to a metanormative view according to which there are norms that are relative to moral and prudential uncertainty, but not relative to decision-theoretic uncertainty.[8]

Metaethical uncertainty

> While normative ethics addresses such questions as "What should I do?", evaluating specific practices and principles of action, meta-ethics addresses questions such as "What is goodness?" and "How can we tell what *is* good from what is bad?", seeking to understand the nature of ethical properties and evaluations. (Wikipedia)

To illustrate, *normative* (or "first-order") ethics involves debates such as "Consequentialist or deontological theories?", while _meta_ethics involves debates such as "Moral realism or moral antirealism?" Thus, **in just the same way we could be uncertain about first-order ethics (*morally uncertain*), we could be uncertain about metaethics (*metaethically uncertain*).**

It seems that metaethical uncertainty is rarely discussed; in particular, I've found no detailed treatment of *how to make decisions* under metaethical uncertainty. However, there is one brief comment on the matter in MacAskill's thesis:

> even if one endorsed a meta-ethical view that is inconsistent with the idea that there's value in gaining more moral information [e.g., certain types of moral antirealism], one should not be certain in that meta-ethical view. And it's high-stakes whether that view is true — if there are moral facts out there but one thinks there aren't, that's a big deal! Even for this sort of antirealist, then, there's therefore value in moral information, because there's value in finding out for certain whether that meta-ethical view is correct.

It *seems to me* that, if and when we face metaethical uncertainties that are relevant to the question of what we should *actually do*, we could likely use basically the same approaches that are advised for decision-making under *moral* uncertainty (which I discuss later in this sequence).[9]

Moral pluralism

A different matter that could *appear* similar to moral uncertainty is *moral pluralism* (aka *value pluralism*, aka *pluralistic moral theories*). According to SEP:

> moral pluralism [is] the view that there are many different moral values.
>
> Commonsensically we talk about lots of different values—happiness, liberty, friendship, and so on. The question about pluralism in moral theory is whether these apparently different values are all reducible to one supervalue, or whether we should think that there really are several distinct values.

MacAskill notes that:

> Someone who [takes a particular expected-value-style approach to decision-making] under uncertainty about whether only wellbeing, or both knowledge and wellbeing, are of value looks a lot like someone who is conforming with a first-order moral theory that assigns both wellbeing and knowledge value.

In fact, one may even decide to react to moral uncertainty by just no longer having *any* degree of belief in each of the first-order moral theories they're uncertain over, and **instead having complete belief in a *new* (and *still first-order*) moral theory that *combines* those previously-believed theories**.[10] For example, after discussing two approaches for thinking about the "moral weight" of different animals' experiences, Brian Tomasik writes:

> Both of these approaches strike me as having merit, and not only am I not sure which one I would choose, but I might actually choose them both. In other words, more than merely having moral uncertainty between them, I might adopt a "value pluralism" approach and decide to care about both simultaneously, with some trade ratio between the two.[11]

But it's important to note that **this *really isn't* the same as moral uncertainty**; the difference is not merely verbal or merely a matter of framing. For example, if Alan has complete belief in a pluralistic combination of utilitarianism and Kantianism, rather than uncertainty over the two theories:

1. Alan has no need for a (second-order) metanormative theory for decision-making under moral uncertainty, because he no longer has any moral uncertainty.
    - If instead Alan has less than complete belief in the pluralistic theory, then the moral uncertainty that remains is between *the pluralistic theory* and whatever other theories he has some belief in (rather than between *utilitarianism*, *Kantianism*, and whatever other theories the person has some belief in).
2. We can't represent the idea of Alan updating to believe more strongly in the Kantian theory, or to believe more strongly in the utilitarian theory.[12]
3. Relatedly, we're no longer able to straightforwardly apply the idea of *value of information* to things that may inform Alan degree of belief in each theory.[13]

Closing remarks

I hope this post helped clarify the distinctions and overlaps between moral uncertainty and related concepts. (And as always, I'd welcome any feedback or comments!) In my next post, I'll continue exploring what moral uncertainty actually **is**, this time focusing on the questions:

1. Is what we "ought to do" under moral uncertainty an *objective* or *subjective* matter?
2. Is what we "ought to do" under moral uncertainty a matter of *rationality* or *morality*?

---

1. For another indication of why the topic of moral uncertainty as a whole matters, see this quote from Christian Tarsney's thesis:

   > The most popular method of investigation in contemporary analytic moral philosophy, the method of reflective equilibrium based on heavy appeal to intuitive judgments about cases, has come under concerted attack and is regarded by many philosophers (e.g. Singer (2005), Greene (2008)) as deeply suspect. Additionally, every major theoretical approach to moral philosophy (whether at the level of normative ethics or metaethics) is subject to important and intuitively compelling objections, and the resolution of these objections often turns on delicate and methodologically fraught questions in other areas of philosophy like the metaphysics of consciousness or personal identity (Moller, 2011, pp. 428- 432). Whatever position one takes on these debates, it can hardly be denied that our understanding of morality remains on a much less sound footing than, say, our knowledge of the natural sciences. If, then, we remain deeply and justifiably uncertain about a litany of important questions in physics, astronomy, and biology, we should certainly be at least equally uncertain about moral matters, even when some particular moral judgment is widely shared and stable upon reflection.

   ↩

2. In an earlier post which influenced this one, Kaj_Sotala wrote:

   > I have long been slightly frustrated by the existing discussions about moral uncertainty that I've seen. I suspect that the reason has been that they've been unclear on what exactly they mean when they say that we are "uncertain about which theory is right" - what is uncertainty about moral theories? Furthermore, especially when discussing things in an FAI [Friendly AI] context, it feels like several different senses of moral uncertainty get mixed together.

   ↩

3. In various places in this sequence, I'll use language that may appear to endorse or presume moral realism (e.g., referring to "moral information" or to probability of a particular moral theory being "correct"). But this is essentially just for convenience; I intend this sequence to be as neutral as possible on the matter of moral realism vs antirealism (except when directly focusing on such matters).

   I think that the interpretation and importance of moral uncertainty is clearest for realists, but, as I discuss in this post, I also think that moral uncertainty can still be a meaningful and important topic for many types of moral antirealist. ↩

4. As another example of this sort of case, suppose I want to know whether fish are "conscious". This may seem on the face of it an empirical question. However, I might not yet know precisely what I *mean* by "conscious", and I might in fact only really want to know whether fish are "conscious in a sense I would morally care about". In this case, the seemingly empirical question becomes hard to disentangle from the (seemingly moral) question: "What forms of consciousness are morally important?"

   And in turn, my answers to *that* question may be influenced by empirical discoveries. For example, I may initially believe that avoidance of painful stimuli

demonstrates consciousness in a morally relevant sense, but then revise that belief when I learn that this behaviour can be displayed in a stimulus-response way by certain extremely simple organisms. ↩

5. The boundaries become even fuzzier, and may lose their meaning entirely, if one assumes the metaethical view *moral naturalism*, which:

   > refers to any version of moral realism that is consistent with [...] general philosophical naturalism. Moral realism is the view that there are objective, mind-independent moral facts. For the moral naturalist, then, there are objective moral facts, these facts are facts concerning natural things, and we know about them using empirical methods. (SEP)

   This sounds to me like it would mean that all moral uncertainties are effectively empirical uncertainties, and that there's no difference in how moral vs empirical uncertainties should be resolved or incorporated into decision-making. But note that that's my own claim; I haven't seen it made explicitly by writers on these subjects.

   That said, one quote that seems to suggest something this claim is the following, from Tarsney's thesis:

   > Most generally, naturalistic metaethical views that treat normative ethical theorizing as continuous with natural science will see first-order moral principles as at least epistemically if not metaphysically dependent on features of the empirical world. For instance, on Railton's (1986) view, moral value attaches (roughly) to social conditions that are stable with respect to certain kinds of feedback mechanisms (like the protest of those who object to their treatment under existing social conditions). What sort(s) of social conditions exhibit this stability, given the relevant background facts about human psychology, is an empirical question. For instance, is a social arrangement in which parents can pass down large advantages to their offspring through inheritance, education, etc, more stable or less stable than one in which the state intervenes extensively to prevent such intergenerational perpetuation of advantage? Someone who accepts a Railtonian metaethic and is therefore uncertain about the first-order normative principles that govern such problems of distributive justice, though on essentially empirical grounds, seems to occupy another sort of liminal space between empirical and moral uncertainty.

   Footnote 15 of this post discusses relevant aspects of moral naturalism, though not this specific question. ↩

6. In fact, Tarsney's (p.140-146) discussion of the difficulty of disentangling moral and empirical uncertainties is used to argue for the merits of approaching moral uncertainty analogously to how one approaches empirical uncertainty. ↩

7. An alternative approach that *also* doesn't require determining whether a given uncertainty is moral or empirical is the "worldview diversification" approach used by the Open Philanthropy Project. In this context, a worldview is described as representing "a combination of views, sometimes very difficult to disentangle, such that uncertainty between worldviews is constituted by a mix of empirical uncertainty (uncertainty about facts), normative uncertainty (uncertainty about morality), and methodological uncertainty (e.g. uncertainty about how to handle uncertainty [...])." Open Phil "[puts] significant resources behind *each* worldview

that [they] find highly plausible." This doesn't require treating moral and empirical uncertainty any differently, and thus doesn't require drawing lines between those "types" of uncertainty. ↩

8. As with metanormative uncertainty in general, this can lead to complicated regresses. For example, there's the possibility to construct causal meta decision theories and evidential meta decision theories, and to be uncertain over which of those meta decision theories to endorse, and so on. As above, see Trammell, Tarsney, and MacAskill (p. 217-219) for discussion of such matters. ↩

9. In a good, short post, Ikaxas writes:

   > How should we deal with metaethical uncertainty? [...] One answer is this: insofar as some metaethical issue is relevant for first-order ethical issues, deal with it as you would any other normative uncertainty. And insofar as it is not relevant for first-order ethical issues, ignore it (discounting, of course, intrinsic curiosity and any value knowledge has for its own sake).
   >
   > Some people think that normative ethical issues ought to be completely independent of metaethics: "The whole idea [of my metaethical naturalism] is to hold fixed ordinary normative ideas and try to answer some *further* explanatory questions" (Schroeder [...]). Others [...] believe that metaethical and normative ethical theorizing should inform each other. For the first group, my suggestion in the previous paragraph recommends that *they ignore metaethics entirely* (again, setting aside any intrinsic motivation to study it), while for the second my suggestion recommends pursuing exclusively those areas which are likely to influence conclusions in normative ethics.

   This seems to me like a good extension/application of general ideas from work on the value of information. (I'll apply such ideas to moral uncertainty later in this sequence.)

   Tarsney gives an example of the sort of case in which metaethical uncertainty is relevant to decision-making (though that's not the point he's making with the example):

   > For instance, consider an agent Alex who, like Alice, divides his moral belief between two theories, a hedonistic and a pluralistic version of consequentialism. But suppose that Alex also divides his *metaethical* beliefs between a robust moral realism and a fairly anemic anti-realism, and that his credence in hedonistic consequentialism is mostly or entirely conditioned on his credence in robust realism while his credence in pluralism is mostly or entirely conditioned on his credence in anti-realism. (Suppose he inclines toward a hedonistic view on which certain qualia have intrinsic value or disvalue entirely independent of our beliefs, attitudes, etc, which we are morally required to maximize. But if this view turns out to be wrong, he believes, then morality can only consist in the pursuit of whatever we contingently happen to value in some distinctively moral way, which includes pleasure but also knowledge, aesthetic goods, friendship, etc.)

   ↩

10. Or, more moderately, one could remove just *some* degree of belief in *some subset* of the moral theories that one had some degree of belief in, and place

that amount of belief in a new moral theory that combines *just that subset* of moral theories. E.g., one may initially think utilitarianism, Kantianism, and virtue ethics each have a 33% chance of being "correct", but then switch to believing that a pluralistic combination of utilitarianism and Kantianism is 67% likely to be correct, while virtue ethics is still 33% likely to be correct. ↩

11. Luke Muelhauser also appears to endorse a similar approach, though not explicitly in the context of moral uncertainty. And Kaj Sotala also seems to endorse a similar approach, though without using the term "pluralism" (I'll discuss Kaj's approach two posts from now). Finally, MacAskill quotes Nozick appearing to endorse a similar approach with regards to decision-theoretic uncertainty:

    > I [Nozick] suggest that we go further and say not merely that we are uncertain about which one of these two principles, [CDT] and [EDT], is (all by itself) correct, but that both of these principles are legitimate and each must be given its respective due. The weights, then, are not measures of uncertainty but measures of the legitimate force of each principle. We thus have a normative theory that directs a person to choose an act with maximal decision-value.

    ↩

12. The closest analog would be Alan updating his beliefs *about the pluralistic theory's contents/substance*; for example, coming to believe that a more correct interpretation of the theory would lean more in a Kantian direction. (Although, if we accept that such an update is possible, it may arguably be best to represent Alan as having moral uncertainty between *different versions* of the pluralistic theory, rather than being certain that the pluralistic theory is "correct" but *uncertain about what it says*.) ↩

13. That said, we can still apply value of information analysis to things like Alan reflecting on how best to interpret the pluralistic moral theory (assuming again that we represent Alan as *uncertain about the theory's contents*). A post later in this sequence will be dedicated to how and why to estimate the "value of moral information". ↩

Moral uncertainty: What kind of 'should' is involved?

This post follows on from my prior post; consider reading that post first.

We are often forced to make decisions under conditions of uncertainty. This may be empirical uncertainty (e.g., what is the likelihood that nuclear war would cause human extinction?), or it may be moral uncertainty (e.g., does the wellbeing of future generations matter morally?).

In my prior post, I discussed overlaps with and distinctions between moral uncertainty and related concepts. In this post, I continue my attempt to clarify **what moral uncertainty actually *is*** (rather than how to make decisions when morally uncertain, which is covered later in the sequence). Specifically, here I'll discuss:

1. Is what we "ought to do" (or "should do") under moral uncertainty an *objective* or *subjective* (i.e., *belief-relative*) matter?
2. Is what we "ought to do" (or "should do") under moral uncertainty a matter of *rationality* or *morality*?

An important aim will be simply clarifying the questions and terms themselves. That said, to foreshadow, the tentative "answers" I'll arrive at are:

1. It seems both more intuitive and more action-guiding to say that the "ought" is *subjective*.
2. Whether the "ought" is a rational or a moral one may be a "merely verbal" dispute with no practical significance. But I'm very confident that interpreting the "ought" as a matter of *rationality* works in any case (i.e., whether or not interpreting it as a matter of *morality* does, and whether or not the distinction really matters).

This post *doesn't* explicitly address what types of moral uncertainty would be *meaningful for moral antirealists and/or subjectivists*; I discuss that topic in a separate post.[1]

*Epistemic status: The concepts covered here are broad, fuzzy, and overlap in various ways, making definitions and distinctions between them almost inevitably debatable. Additionally, I'm not an expert in these topics (though I have now spent a couple weeks mostly reading about them). I've tried to mostly collect, summarise, and synthesise **existing** ideas (from academic philosophy and the LessWrong and EA communities). I'd appreciate feedback or comments in relation to any mistakes, unclear phrasings, etc. (and just in general!).*

Objective or subjective?

*(Note: What I discuss here is **not** the same as the objectivism vs subjectivism debate **in metaethics**.)*

As I noted in a prior post:

> *Subjective* normativity relates to what one should do *based on what one believes*, whereas *objective* normativity relates to what one "actually" should do (i.e., based on the true state of affairs).

Hilary Greaves & Owen Cotton-Barratt give an example of this distinction in the context of *empirical* uncertainty:

> Suppose Alice packs the waterproofs but, as the day turns out, it does not rain. Does it follow that Alice made the wrong decision? In one (**objective**) sense of "wrong", yes: thanks to that decision, she experienced the mild but unnecessary inconvenience of carrying bulky raingear around all day. But in a second (more **subjective**) sense, clearly it need not follow that the decision was wrong: if the probability of rain was sufficiently high and Alice sufficiently dislikes getting wet, her decision could easily be the appropriate one to make given her state of ignorance about how the weather would in fact turn out. Normative theories of decision-making under uncertainty aim to capture this second, more subjective, type of evaluation; the standard such account is expected utility theory.

Greaves & Cotton-Barratt then make the analogous distinction for *moral* uncertainty:

> How should one choose, when facing relevant **moral** uncertainty? In one (**objective**) sense, of course, what one should do is simply what the true moral hypothesis says one should do. But it seems there is also a second sense of "should", analogous to the **subjective** "should" for empirical uncertainty, capturing the sense in which it is appropriate for the agent facing moral uncertainty to be guided by her moral credences [i.e., beliefs], whatever the moral facts may be. (emphasis added)

(This objective vs subjective distinction seems to me somewhat similar - though not identical - to the distinction between *ex post* and *ex ante* thinking. We might say that Alice made the right decision *ex ante* - i.e., based on what she knew when she made her decision - even if it *turned out* - *ex post* - that the other decision would've worked out better.)

MacAskill notes that, in both the empirical and moral contexts, "The principal argument for thinking that there must be a subjective sense of 'ought' is because the objective sense of 'ought' is not sufficiently action-guiding." He illustrates this in the case of moral uncertainty with the following example:

> Susan is a doctor, who faces three sick individuals, Greg, Harold and Harry. Greg is a human patient, whereas Harold and Harry are chimpanzees. They all suffer from the same condition. She has a vial of a drug, D. If she administers all of drug D to Greg, he will be completely cured, and if she administers all of drug to the chimpanzees, they will both be completely cured (health 100%). If she splits the drug between the three, then Greg will be almost completely cured (health 99%), and Harold and Harry will be partially cured (health 50%). She is unsure about the value of the welfare of non-human animals: she thinks it is equally likely that chimpanzees' welfare has no moral value and that chimpanzees' welfare has the same moral value as human welfare. And, let us suppose, there is no way that she can improve her epistemic state with respect to the relative value of humans and chimpanzees.
>
> [...]
>
> Her three options are as follows:

A: Give all of the drug to Greg

B: Split the drug

C: Give all of the drug to Harold and Harry

Her decision can be represented in the following table, using numbers to represent how good each outcome would be.

	Chimpanzee welfare is of no moral value – 50%	Chimpanzee welfare is of significant moral value – 50%
A	100	0
B	99	199
C	0	200

Finally, suppose that, according to the true moral theory, chimpanzee welfare is of the same moral value as human welfare and that therefore, she should give all of the drug to Harold and Harry. What should she do?

Clearly, the best *outcome* would occur if Susan does C. But she doesn't *know* that that would cause the best outcome, because she doesn't *know* what the "true moral theory" is. She thus has no way to act on the advice "Just do what is *objectively* morally right." Meanwhile, as MacAskill notes, "it seems it would be morally reckless for Susan **not** to choose option B: given what she knows, she would be risking severe wrongdoing by choosing either option A or option C" (emphasis added).

To capture the intuition the Susan should choose option B, and **to provide *actually followable* guidance for action, we need to accept that there is a *subjective* sense of "should"** (or of "ought") - a sense of "should" that *depends in part on what one believes*. (This could also be called a "belief-relative" or "credence-relative" sense of "should".)[2]

An additional argument in favour of accepting that there's a subjective "should" in relation to moral uncertainty is consistency with how we treat *empirical* uncertainty, where most people accept that there's a subjective "should".[3] This argument is made regularly, including by MacAskill and by Greaves & Cotton-Barratt, and it seems particularly compelling when one considers that it's often difficult to draw clear lines between empirical and moral uncertainty (see my prior post). That is, if it's often hard to say whether an uncertainty is empirical or moral, it seems strange to say we should accept a subjective "should" under empirical uncertainty but *not* under moral uncertainty.

Ultimately, most of what I've read on moral uncertainty is premised on there being a subjective sense of "should", and much of this sequence will rest on that premise also.[4] As far as I can tell, this seems necessary if we are to come up with any meaningful, action-guiding approaches for decision-making under moral uncertainty ("metanormative theories").

But I should note that *some* writers *do* appear to argue that there's only an objective sense of "should" (one example, I think, is Weatherson, though he uses different language and I've only skimmed his paper). Furthermore, while I can't see how this could lead to action-guiding principles for *making decisions under uncertainty*, it does seem to me that it'd still allow for *resolving* one's uncertainty. In other words, if we do recognise only objective "oughts":

- We may be stuck with fairly useless principles for decision-making, such as "Just do what's actually right, even when you don't know what's actually right"
- But (as far as I can tell) we could still be guided to *clarify* and *reduce* our uncertainties, and thereby bring our beliefs more in line with what's actually right.

Rational or moral?

> There is also debate about what precisely kind of "should" is involved [in cases of moral uncertainty]: rational, moral, or something else again. (Greaves & Cotton-Barratt)

For example, in the above example of Susan the doctor, are we wondering what she *rationally* ought to do, given her moral uncertainty about the moral status of chimpanzees, or what she *morally* ought to do?

It may not matter either way

Unfortunately, even after having read up on this, it's not actually clear to me what the distinction is meant to be. In particular, I haven't come across a clear explanation of what it would mean for the "should" or "ought" to be *moral*. I suspect that what that would mean would be partly a matter of interpretation, and that some definitions of a "moral" should could be effectively the same as those for a "rational" should. (But I should note that I didn't look exhaustively for such explanations and definitions.)

Additionally, both Greaves & Cotton-Barratt and MacAskill explicitly avoid the question of whether what one "ought to do" under moral uncertainty is a matter of rationality or morality.[5] This does not seem to at all hold them back from making valuable contributions to the literature on moral uncertainty (and, more specifically, on how to make decisions when morally uncertain).

Together, the above points make me **inclined to believe (though with low confidence) that this may be a "merely verbal" debate with no real, practical implications** (at least while the words involved remain as fuzzy as they are).

However, I still did come to two less-dismissive conclusions:

1. I'm very confident that the project of working out meaningful, action-guiding principles for decision-making under moral uncertainty **makes sense if we see the relevant "should" as a rational one**. (Note: This doesn't mean that I think the "should" *has* to be seen as a rational one.)
2. I'm less sure whether that project would make sense if we see the relevant "should" as a moral one. (Note: This doesn't mean I have any particular reason to believe it *wouldn't* make sense if we see the "should" as a moral one.)

I provide my reasoning behind these conclusions below, though, given my sense that this debate may lack practical significance, **some readers may wish to just skip to the next section**.

A rational "should" likely works

Bykvist writes:

> An alternative way to understand the ought relevant to moral uncertainty is in terms of rationality (MacAskillet al., forthcoming; Sepielli, 2013). Rationality, in one important sense at least, has to do with what one should do or intend, given one's beliefs and preferences. This is the kind of rationality that decision theory often is seen as invoking. It can be spelled out in different ways. One is to see it as a matter of coherence: It is rational to do or intend what coheres with one's beliefs and preferences (Broome, 2013; for a critic, see Arpaly, 2000). Another way to spell it out is to understand it as matter of rational processes: it is rational to do or intend what would be the output of a rational process, which starts with one's beliefs and preferences (Kolodny, 2007).
>
> To apply the general idea to moral uncertainty, we do not need to take stand on which version is correct. We only need to assume that when a conscientious moral agent faces moral uncertainty, she cares about doing right and avoid doing wrong but is uncertain about the moral status of her actions. She prefers doing right to doing wrong and is indifferent between different right doings (at least when the right doings have the same moral value, that is, none is morally supererogatory). She also cares more about serious wrongdoings than minor wrongdoings. The idea is then to apply traditional decision theoretical principles, according to which rational choice is some function of the agent's preferences (utilities) and beliefs (credences). Of course, different decision-theories provide different principles (and require different kinds of utility information). But the plausible ones at least agree on cases where one option dominates another.
>
> Suppose that you are considering only two theories (which is to simplify considerably, but we only need a logically possible case): "business as usual," according to which it is permissible to eat factory-farmed meat and permissible to eat vegetables, and "vegetarianism," according to which it is impermissible to eat factory-farmed meat and permissible to eat vegetables. Suppose further that you have slightly more confidence in "business as usual." The option of eating vegetables will dominate the option of eating meat in terms of your own preferences: No matter which moral theory is true, by eating vegetables, you will ensure an outcome that you weakly [prefer] to the alternative outcome: if "vegetarianism" is true, you prefer the outcome; if "business as usual is true," you are indifferent between the outcomes. The rational thing for you to do is thus to eat vegetables, given your beliefs and preferences. (lines breaks added)

It seems to me that that reasoning makes perfect sense, and that we can have valid, meaningful, action-guiding principles about what one *rationally* (and subjectively) should do given one's moral uncertainty. This seems further supported by the approach Christian Tarsney takes, which seems to be useful and to also treat the relevant "should" as a rational one.

Furthermore, MacAskill seems to suggest that there's a correlation between (a) writers fully engaging with the project of working out action-guiding principles for decision-

making under moral uncertainty and (b) writers considering the relevant "should" to be rational (rather than moral):

> (Lockhart 2000, 24,26), (Sepielli 2009, 10) and (Ross 2006) all take metanormative norms to be norms of rationality. (Weatherson 2014) and (Harman 2014) both understand metanormative norms as moral norms. So there is an odd situation in the literature where the defenders of metanormavism (Lockhart, Ross, and Sepielli) and the critics of the view (Weatherson and Harman) seem to be talking past one another.

A moral "should" may or may not work

I haven't seen any writer (a) *explicitly* state that they understand the relevant "should" to be a moral one, and then (b) go on to fully engage with the project of working out meaningful, action-guiding principles for decision-making under moral uncertainty. Thus, I have an absence of evidence that one can engage in that project while seeing the "should" as moral, and I take this as (very weak) evidence that one can't engage in that project while seeing the "should" that way.

Additionally, as noted above, MacAskill writes that Weatherson and Harman (who seem fairly dismissive of that project) see the relevant "should" as a moral one. Arguably, this is evidence that that project of finding such action-guiding principles won't make sense if we see the "should" as moral (rather than rational). However, I consider this to *also* be very weak evidence, because:

- It's only two data points.
- It's just a correlation anyway.
- I haven't closely investigated the "correlation" myself. That is, I haven't checked whether or not Weatherson and Harman's reasons for dismissiveness seem highly related to them seeing the "should" as moral rather than rational.

Closing remarks

In this post, I've aimed to:

- Clarify what is meant by the question "Is what we "ought to do" under moral uncertainty is an *objective* or *subjective* matter?"
- Clarify what is meant by the question "Is that 'ought' a matter of *rationality* or of *morality*?"
- Argue that it seems both more intuitive and more action-guiding to say that the "ought" is *subjective*.
- Argue that whether the "ought" is a rational or a moral one may be a "merely verbal" dispute with no practical significance (but that interpreting the "ought" as a matter of *rationality* works in any case).

I hope this has helped give readers more clarity on the seemingly neglected matter of what we actually *mean* by moral uncertainty. (And as always, I'd welcome any feedback or comments!)

My next posts will continue in a similar vein, but this time building to the question of whether, when we're talking about moral uncertainty, we're actually talking about *moral risk* rather than about *moral (Knightian) uncertainty* - and whether such a

distinction is truly meaningful. (To do so, I'll first discuss the risk-uncertainty distinction *in general*, and the related matter of unknown unknowns, before applying these ideas in the context of *moral* risk/uncertainty in particular.)

---

1. But the current post is still relevant for many types of moral antirealist. As noted in my last post, this sequence will sometimes use language that may appear to endorse or presume moral realism, but this is essentially just for convenience. ↩

2. We could further divide subjective normativity up into, roughly, "what one should do based on what one *actually* believes" and "what one should do based on what it *would be reasonable* for one to believe". The following quote, while not directly addressing that *exact* distinction, seems relevant:

   > Before moving on, we should distinguish subjective credences, that is, degrees of belief, from epistemic credences, that is, the degree of belief that one is epistemically justified in having, given one's evidence. When I use the term 'credence' I refer to epistemic credences (though much of my discussion could be applied to a parallel discussion involving subjective credences); when I want to refer to subjective credences I use the term 'degrees of belief'.
   >
   > The reason for this is that appropriateness seems to have some sort of normative force: if it is most appropriate for someone to do something, it seems that, other things being equal, they ought, in the relevant sense of 'ought', to do it. But people can have crazy beliefs: a psychopath might think that a killing spree is the most moral thing to do. But there's no sense in which the psychopath ought to go on a killing spree: rather, he ought to revise his beliefs. We can only capture that idea if we talk about epistemic credences, rather than degrees of belief.

   (I found that quote in this comment, where it's attributed to MacAskill's *BPhil* thesis. Unfortunately, I can't seem to access that thesis, including via Wayback Machine.) ↩

3. Though note that Greaves and Cotton-Barratt write:

   > Not everyone does recognise a subjective reading of the moral 'ought', even in the case of empirical uncertainty. One can distinguish between objectivist, (rational-)credence-relative and pluralist views on this matter. According to objectivists (Moore, 1903; Moore, 1912; Ross, 1930, p.32; Thomson, 1986, esp. pp. 177-9; Graham, 2010; Bykvist and Olson, 2011) (respectively, credence-relativists (Prichard, 1933; Ross, 1939; Howard-Snyder, 2005; Zimmermann, 2006; Zimmerman, 2009; Mason, 2013), the "ought" of morality is uniquely an objective (respectively, a credence-relative) one. According to pluralists, "ought" is ambiguous between these two readings (Russell, 1966; Gibbard, 2005; Parfit, 2011; Portmore, 2011; Dorsey, 2012; Olsen, 2017), or varies between the two readings according to context (Kolodny and Macfarlane, 2010).

   ↩

4. In the following quote, Bykvist provides what seems to me (if I'm interpreting it correctly) to be a different way of explaining something similar to the objective vs subjective distinction.

> One possible explanation of why so few philosophers have engaged with moral uncertainty might be serious doubt about whether it makes much sense to ask about what one ought do when one is uncertain about what one ought to do. The obvious answer to this question might be thought to be: "you ought to do what you ought to do, no matter whether or not you are certain about it" (Weatherson, 2002, 2014). However, this assumes the same sense of "ought" throughout.
>
> A better option is to assume that there are different kinds of moral ought. We are asking what we morally ought to do, in one sense of ought, when we are not certain about what we morally ought to do, in another sense of ought. One way to make this idea more precise is to think about the different senses as different levels of moral ought. When we face a moral problem, we are asking what we morally ought to do, at the first level. Standard moral theories, such as utilitarianism, Kantianism, and virtue ethics, provide answers to this question. In a case of moral uncertainty, we are moving up one level and asking about what we ought to do, at the second level, when we are not sure what we ought to do at the first level. At this second level, we take into account our credence in various hypotheses about what we ought to do at the first level and what these hypotheses say about the moral value of each action (MacAskill et al., forthcoming). This second level ought provides a way to cope with the moral uncertainty at the first level. It gives us a verdict of how to best manage the risk of doing first order moral wrongs. That there is such a second-level moral ought of coping with first-order moral risks seems to be supported by the fact that agents are morally criticizable when they, knowing all the relevant empirical facts, do what they think is very likely to be a first-order moral wrong when there is another option that is known not to pose any risk of such wrongdoing.

Yet another (and I think similar) way of framing this sort of distinction could make use of the following two terms: "A criterion of rightness tells us what it takes for an action to be right (if it's actions we're looking at). A decision procedure is something that we use when we're thinking about what to do" (Askell).

Specifically, we might say that the true first-order moral theory provides objective "criteria of rightness", but that we don't have direct access to what these are. As such, we can use a second-order "decision procedure" that attempts to lead us to take actions that are close as possible to the best actions (according to the unknown criteria of rightness). To do so, this decision procedure must make use of our credences (beliefs) in various moral theories, and is thus subjective. ↩

5. Greaves & Cotton-Barratt write: "For the purpose of this article, we will [...] not take a stand on what kind of "should" [is involved in cases of moral uncertainty]. Our question is how the "should" in question behaves in purely extensional terms. Say that an answer to that question is a *metanormative theory*."

   MacAskill writes: "I introduce the technical term 'appropriateness' in order to remain neutral on the issue of whether metanormative norms are rational norms, or some other sort of norms (though noting that they can't be first-order norms provided by first-order normative theories, on pain of inconsistency)." ↩

Value uncertainty

Epistemic status: This is basically a new categorisation scheme for, and analysis of, ideas that other people have proposed previously (both in relation to moral philosophy and in relation to AI alignment). I'm not an expert on the topics I cover here, and I'd appreciate feedback or comments in relation to any mistakes, unclear phrasings, etc. (and just in general!).

We are often forced to make decisions under conditions of uncertainty. This may be empirical uncertainty (e.g., what is the likelihood that nuclear war would cause human extinction?), or it may be moral uncertainty (e.g., is the wellbeing of future generations morally important?).

But what if you don't believe that "morally important" is a coherent concept? What if you're a *moral antirealist* and/or *subjectivist*, and thus reject the idea that there are any (objective) moral facts? Would existing work on moral uncertainty (see my prior posts) still be relevant to you?

I think that a lot of it is, to a large extent, for the reasons discussed in this footnote.[1] But I think that, to directly discuss *how* that work is relevant to antirealists and/or subjectivists, it would help to speak not of *moral* uncertainty but of *value uncertainty* (VU; i.e., uncertainty about what one "values" or "prefers"). Doing so also helps us to categorise different *types* of VU, and potential ways of resolving each of these types of VU. A final benefit is that such an analysis of VUs also has substantial relevance for moral *realists*, and for work on AI alignment.

So this post will:

1. Clarify what I mean by "values" in this context
2. Name, briefly describe, and briefly suggest responses to four types of VU, and two types of situations that actually *aren't* VU, but could look as though they are
3. Return to the question of who and what these ideas are relevant for

Values

I should clarify a few points about what I mean by "values" in this post:

- I mean what a person *actually* values (or would if they knew more, or will in the future, or would if "idealised"), not what a person's *explicitly endorsed* moral theory suggests they *should* value.
- For example, after a bunch of caveats and moral uncertainty, I roughly identify as a classical utilitarian. However, *in reality*, *psychologically speaking*, I also value things other than whatever maximally increases the wellbeing of conscious beings.
    - On the other hand, I don't believe that people *actually* have any specific, neat, ongoing set of values, of the sort that could be used to form a utility function or something like that (see this thread). So this post is basically about how to move towards better understanding whatever's the *closest equivalent* a person *does* have to a specific, neat, ongoing set of values.

- I'll often talk about "a person's" values, which could mean *one's own* values, *someone else's* values, *a group's* values, or *humanity as a whole's* values.

Types of value uncertainty

I'll now name, briefly describe, and briefly suggest responses to four (overlapping) types of VU, and then two types of situations that *aren't* VU but could *appear* to be VU. I hope to later write a post for each type of VU (and the two related situations), where I'll go into more detail and highlight more connections to prior work (e.g., by Kaj Sotala and Justin Shovelain).

Note that this is *not* the only (and perhaps not the best) way to categorise types of VU or frame this sort of discussion. It also may leave out important types. I'm open to feedback about all of this, and in fact one motivation for summarising this categorisation scheme here and then *later* writing more about each type is that doing so allows those later posts to be influenced by feedback on this one.

Present

Description/cause: Present VU is uncertainty about a person's (or group's) *current* values. This occurs when multiple different sets of underlying values could explain the data (i.e., the behaviours you've observed from the person),[2] essentially creating a standard curve fitting problem.

One cause of Present VU is a lack of knowledge about something about the person *other than* their values, such as:

- What decision theory they're using and how rational they are. E.g., does the fact that Alice left studying till the last minute reflect that she values being stressed, or just reflect hyperbolic discounting?
- Their capabilities. E.g., does Kasparov's loss to DeepBlue indicate that he values losing, or just that he didn't *know how* to win?
- Their beliefs. E.g., does this person's continued smoking indicate that they value smoking more than they value additional years of life, or just that they don't know the effects of smoking?

But Present VU can also occur when, even holding constant all of the above factors, different sets of values would lead to the same behaviours (e.g., breathing, or following convergent instrumental goals) in a given circumstance.

This type of VU seems similar to the focus of inverse reinforcement learning. However, with Present VU, the "learner" may not be an AI. (In fact, Present VU may involve *you* trying to learn *your own* values, in which case it could perhaps be thought of as "Introspective" VU.)

Potential ways to resolve this: Gather more data, or do more thinking, regarding the person's decision theory, rationality, capabilities, and/or beliefs.

Think about the assumptions you're making about those factors. Try making different assumptions, and/or "minimal" assumptions. (Similar ideas, and some difficulties with them, have been discussed before by Armstrong and Worley, among others.)

Observe more of the person's behaviours, ideally under different circumstances.

(Similar data and thinking regarding *other people's* behaviours, rationality, etc. may also help to some extent. E.g., evidence about the degree to which *people in general* hyperbolically discount things could help you interpret the behaviour of some other person for whom there is no such data.)

Informational

Description/cause: Informational VU is uncertainty about what a person's (or group's) values *would be* if their knowledge or beliefs *improved*.[3]

We could divide the potential sources of improved knowledge or beliefs into three categories:

- New experiences. E.g., Alan values not eating refried beans, because they look like sewage, but he'd reverse this value if he ever actually *tried* eating them.[4]
- Learning new facts relevant to certain values. E.g., Betty doesn't value the wellbeing of octopi, but would if she learned more about their neurology and behaviour.
- Improved ontologies (related to ontological crises). E.g., Cameron values "his future self's" wellbeing, so he'd have to find a new way to map that value onto the world if he accepted the idea that people don't fundamentally retain the "same identity" over time (relevant discussion here). See also this and this.

Potential ways to resolve this: Think about (or use models/simulations to work out) what new experiences, new facts, or improvements in ontologies would be most likely to affect the person's values, and *how* the person's values would change in response. (This could perhaps be informed by ideas from value of information analysis and sensitivity analysis.)

Try to expose the person to, or teach them about, these (or other) new experiences, new facts, and improvements in ontologies. (Note that this would involve not just predicting but actually *causing* changes in values. This should be done cautiously, if at all.)

For resolving uncertainty about how *your own* values would change if your knowledge or beliefs improved, this might look like just learning a lot, particularly about things you're especially uncertain about and that seem especially relevant to your values.

Predictive

Description/cause: Predictive VU is uncertainty about what a person's (or group's) values *will be* in the future.[5]

This overlaps with Present and Informational VU, because:

- Learning more about a person's *current* values obviously helps you predict what their values will be in the future.
- One source of the changes that *will* occur in a person's values is improvements that *will* occur in the person's knowledge and beliefs. Thus, a subset of what Informational VU is about is also a subset of what Predictive VU is about.

But Predictive VU also includes uncertainty about changes that will occur in a person's values for *other* reasons, such as:

- changes in knowledge and beliefs that *aren't* improvements (e.g., learning misinformation about immigrants)
- being persuaded via the peripheral route (e.g., seeing what attractive or high-status people are doing)
- changes in life circumstances (e.g., nearing the end of life, and thus coming to place greater value on spending quality time with good friends and family, and less value on maintaining large networks)
- what could be called "biological" changes (e.g., changes in hormone levels that lead to changes in libido, risk-taking, etc.)

Potential ways to resolve this: The potential methods for resolving Present and Informational VU are relevant for parts of Predictive VU. E.g., thinking about what a person *will* learn about, and how it *will* affect their values, can help resolve parts of both Informational and Predictive VU.

Also, for all parts of this VU, techniques that are effective for prediction in general (e.g., reference class forecasting) should be useful. E.g., an aspiring effective altruist could predict that their values are likely to shift away from "typical EA values" over time, based on data indicating that that's a common pattern, as well as the more general end-of-history illusion**.**

Idealised

Description/cause: Idealised VU is uncertainty about:

- what values a person's (or group's) "idealised self" or "ideal advisor" would have or advise;
- what values a person would have after "a process of idealisation" or after reaching "reflective equilibrium";
- what values a person would have after "coherent extrapolated volition" (CEV);
- what set of consistent values is closest to a person's current, inconsistent values (see, e.g., this post); or
- something else along these lines.

(Note that it seems to me that it's hard to actually specify those key terms, and I won't properly try to do so here; more details can be found in the links.)

This overlaps with the other types of VU, in that:

- Again, knowing about a person's current values obviously provides a useful starting point when trying to extrapolate out from there.
- The idealisation/extrapolation process would likely involve some improvements in beliefs and knowledge, and perhaps some of the changes that *actually will* occur for the person in the future.

Potential ways to resolve this: This depends substantially on what we mean by the hard-to-specify terms involved. Also, it might be impossible or highly impractical to *actually* work out what values would result from the idealisation or extrapolation process, due to issues such as limited computing power.

But we can perhaps try to *approximate* such an idealisation or extrapolation process, or predict approximately what it *would* result in, using methods like:

- Learning more
- Engaging in more moral reflection
- Trying to forecast (using best practices) what many simulations of the person would say if they'd had lots of time to learn and reflect more (see Muehlhauser)
- Thinking about what apparent "moral progress" in the past has looked like, and what changes from current values might result from similar processes of change
    - E.g., if a person approves of the moral circle expansions that have occurred so far, perhaps we should expect that that values of the person's idealised self would reflect an even more expanded moral circle.

(For details, see the sources linked to at the start of this subsection.)

Situations that could *look* like value uncertainty

I'll now briefly discuss two other types of situations in which a person (or group) *actually isn't* uncertain about their values, but could appear to be, or could even believe themselves to be.

Value *conflict*

Description/cause: Value conflict (VC) is when some or all of the values a person (or group) *actually has* are in conflict with each other. It's like the person has multiple, competing utility functions, or different "parts of themselves" pushing them in different directions.

E.g., Dana is someone whose values include *both* maximising welfare *and* absolutely respecting people's rights; it's not simply that she's uncertain which value she actually has deep down.

In *some* ways, the results of this can be the same as the results of VU (particularly *Present* VU). For this reason, the person's situation may be misdiagnosed as VU by themselves or by others. (E.g., Dana may try to figure out which of those somewhat conflicting values she "really" has, rather than realising that she really has both.)

Potential ways to respond: It seems unclear whether VC is a "problem", as opposed to an acceptable result of the fragility and complexity of our value systems. It thus also seems unclear whether and how one should try to "solve" it. That said, it seems like three of the most obvious *options* for "solving" it are to:

- Engage in, approximate, or estimate the results of "idealisation" with regards to the conflicting values

- (Note that, if one is trying to help *someone else* "solve" *their* VC, one might instead encourage or help *that person* to use this or the following options)
- Use approaches similar to those meant for decision-making under moral uncertainty (see also this), except that here the person is actually certain about their values, so the "weight" given to the values is based on something like how "important" those values feel, rather than degree of belief in them
- Embrace moral pluralism
    - E.g., decide to keep as values each of the conflicting values, and just give them a certain amount of "say" or "weight" in your decision-making.
    - This may not always work, and it's unclear how to decide on how to allocate "say" in any case.

(Related discussion can be found in the "Moral pluralism" section of this post.)

If the goal is just to understand the person or predict their behaviours (rather than helping them to "resolve" their conflict), then one might instead think about, model, or simulate what *would* happen *if* the person used one or more of the above options.

Merely professed VU (or VC)

Description/cause: Merely professed VU (or merely professed VC) is when a person *claims* to be uncertain about their values (or to have multiple, conflicting values), despite this not being the case. They may do this for game-theoretic, signalling, or bargaining reasons.

An example of merely professed *VU*: Eric is certain about his values, and really wants to influence you to have similar values. But he also thinks that, if you believe that he's uncertain and open to changing his mind, you'll be more open to talking about values with him and to changing *your* mind. Thus, he feigns VU.

An example of merely professed *VC*: Fatma actually knows that she values only her own wellbeing, but she wishes to gain resources from altruists. To do so, she claims that there's "part of her" that values benefiting only herself, and "another part of her" that values helping others.

(*Perhaps* this sort of thing could also play out on an unconscious level, so that *the person themselves* genuinely believes that they have VU or VC. But then it seems hard to disentangle this from *actual* VU or VC.)

Potential ways to respond: I haven't thought much about this, and I think how to respond would depend a lot on what one wishes the responses to achieve and on the specific situation. It seems like often one should respond in the same ways that are generally useful when someone may be lying to you or trying to manipulate you.

Who and what are these ideas useful for?

Antirealists and/or subjectivists

As noted in the introduction, one purpose of this post is to explicitly discuss the ways in which (something like) moral uncertainty is relevant for **moral antirealists** and/or **subjectivists**.

Roughly speaking (see Joyce for details), a moral antirealist is someone who accepts one of the three following claims:

- **Noncognitivism**: The position that moral sentences are neither true nor false; they are not beliefs or factual claims. For example, moral sentences might express one's emotions (e.g., "Murder is bad" might mean something like "Murder - boo!").
- **Error theory**: The position that moral sentences are *meant to be* beliefs or factual claims, but are just never true, as there are simply no moral facts. Error theory is similar to "nihilism".
- **Subjectivism** (or **non-objectivism**): "moral facts exist and are mind-dependent (in the relevant sense)" (Joyce). In other words, moral claims *can* be true, but their truth or falsity depends on someone's judgement, rather than being simply an objective fact about the universe.
    - A moral subjectivist *may* be a **moral relativist**; e.g., they could believe that what's morally true for me could be different from what's morally true for you. But they don't *have* to be a relativist; e.g., they could believe that the same things are morally true for everyone, but that moral truth depends on one particular person's judgement, or on the judgement of an "ideal observer".

(In contrast, a **moral realist** is someone who rejects all three of those claims. Thus, moral realists believe that moral sentences *do* (at least sometimes) reflect beliefs or factual claims, that they *can* sometimes be true, and that their truth or falsity *is* objective.[6]

Of these types of antirealism, VU (and VC) is most clearly relevant in the case of subjectivism. For example, many subjectivists think that *their own values* (or their future or idealised values) are at least part of what determines the truth or falsity of moral claims. For these people, resolving uncertainty about their own values should seem very important. Other subjectivists may want to resolve uncertainty about the present, future, or idealised values of their society, of humanity as a whole, of all intelligent life, or something else like that (depending on what they think determines moral truth).

It's less clear what special relevance VU would have for noncognitivists or error theorists (ignoring the argument that they should be metaethically uncertain about those positions). That said:

- It's possible that all work on moral uncertainty should be meaningful to at least some types of noncognitivists (see Sepielli, though see Bykvist and Olson for counterarguments).
- Resolving VU should be important to people who reject the idea of subjective *morality*, as long as they accept that the existence of *some* form of normativity

(see my earlier post for discussion of the overlaps and distinctions between morality, normativity, and related concepts).
- My impression is that most "antirealists" in the LessWrong and EA communities are either moral subjectivists or at least accept the existence of *some* form of normativity (see bmgarfinkel for somewhat relevant discussion)

Moral realists

For moral realists, standard work on moral uncertainty is already clearly relevant. That said, VU still has additional relevance even for moral realists, because:

- For most plausible moral theories, making things go better according to (some) conscious beings' values is morally good. In that case, resolving VUs would help you do morally good things.
- Moral philosophy (whether realist-leaning or not) very often uses intuitions about moral matters as pivotal data when trying to work out what's morally correct. Thus, it *could* be argued that progress in resolving VUs would lead to better moral theories (and thus better behaviours), even by moral realists' lights.
- In order to do morally good things, it will often be useful to be able to predict the *behaviours* of oneself (e.g., should I donate now in case my values drift?) and of others. As values influence behaviours, resolving VU helps you make such predictions.
- Some moral realists may also believe that there's room for prudential (self-interested, non-moral) "should"s, and may therefore want to understand their own values so that they can more effectively do what's "prudentially right".

AI alignment

Ideally, we want our AIs to act in accordance with what we truly value (or what we'd value after some process of idealisation or CEV). Depending on definitions, this may be seen as the core of AI alignment, as one important part of AI alignment, or as at least a nice bonus (e.g., if we use Paul Christiano's definition).

As such, recognising and resolving VUs (and VCs) seems of very clear relevance to AI alignment work. This seems somewhat evidenced by how many VU-related ideas I found in previous alignment-related work (e.g., value learning, inverse reinforcement learning, Stuart Armstrong's research agenda, and CEV). Indeed, a major reason why I'm interested in the topic of VU is its relevance to AI alignment, and I hope that this post can provide useful concepts and framings for others who are also interested in AI alignment.

As mentioned earlier, please do comment if you think there are better categorisations/framings for this topic, better names, additional types worth mentioning, mistakes I've made, or whatever.

My thanks to Justin Shovelain and David Kristoffersson of Convergence Analysis for helpful discussions and feedback on this post.

---

1. Firstly, even someone "convinced" by antirealism and/or subjectivism probably shouldn't be *certain* about those positions. Thus, such people should probably act as if *metaethically* uncertain, and that requires concepts and responses

somewhat similar to those discussed in existing work on moral uncertainty. (See this post's section on "Metaethical uncertainty".)

Relevantly, MacAskill writes: "even if one endorsed a meta-ethical view that is inconsistent with the idea that there's value in gaining more moral information, one should not be certain in that meta-ethical view. And it's high-stakes whether that view is true — if there are moral facts out there but one thinks there aren't, that's a big deal! Even for this sort of antirealist, then, there's therefore value in moral information, because there's value in finding out for certain whether that meta-ethical view is correct."

Secondly, a lot of existing work on moral uncertainty isn't (explicitly) premised on moral realism.

Thirdly, in practice, many similar concepts and principles will be useful for:

- a moral realist who wants to act in accordance with what's "truly right", but who doesn't know what that is
- an antirealist or subjectivist who wants to act in accordance with their "fundamental" or "idealised" values, but who doesn't know what those are.

(This third point has been discussed by, for example, Stuart Armstrong.) ↩

2. Here I use the term "behaviour" very broadly, to include not just our "physical actions" but also what decisions we make, what we say, and what we think (at least on a conscious level). This is because any of these could provide data about underlying values. So some examples of what I'd count as "behaviours" include:

   - going for a run
   - choosing to pursue a career in existential risk reduction
   - saying you love a certain band
   - being unable to stop yourself from thinking about ice cream.

   ↩

3. I haven't listed a specific type of VU for uncertainty about what a person's values would be if their knowledge of beliefs *changed*, *whether for the better or not*. This is because I don't see that as being particularly worth knowing about in its own right, separate from the other types of VU. But the next type of VU (Predictive VU) does incorporate uncertainty about what a person's values *will be* after the changes that *will occur* to their knowledge or beliefs (whether or not these changes are improvements). ↩

4. Note the (somewhat fuzzy) distinction from Present VU:

   - Informational VU is partly *about* what a person's values *would be* if they *did have* certain new experiences
   - Present VU can be partly *resolved* by exposing people to new experiences and seeing how they behave, as this lets you *gather more data* about what their values *already were*.

   ↩

5. One could argue that it doesn't make sense to talk of "a person's values changing". Such arguments could be based on the idea that an "agent" is partly

defined by its values (or utility function, or whatever), or the idea that people don't fundamentally retain the "same identity" over time anyway. For this post, I wish to mostly set aside such complexities, and lean instead on the fact that it's often *useful* to think and speak *as if* "the same person" persists over time (and even despite partial changes in values).

But I do think those complexities are worth acknowledging. One reason is that they can remind us to not take it for granted that a person will or should currently care about "their future self's values" (or "their future self's" ability to act on their values). This applies especially if the person *doesn't* see any reason to care about *other people's* values (or abilities to achieve their values). Ruairi Donnelly discusses similar points, and links this to the concept of *value drift*.

Similar points could also be raised regarding the type of VU I'll cover next: Idealised VU. (See Armstrong for somewhat related ideas.) ↩

6. However, some philosophers classify subjectivists as moral realists instead of as antirealists. To account for this, some sources distinguish between ***minimal* moral realism** (which includes subjectivists) and ***robust* moral realism** (which excludes them). This is why I sometimes write "antirealists *and/or* subjectivists".) ↩

Making decisions under moral uncertainty

Cross-posted to the EA Forum. Updated substantially since initial publication.

Overview/purpose of this sequence

While working on an (upcoming) post about a new way to think about moral uncertainty, I unexpectedly discovered that, as best I could tell:

1. There was no single post on LessWrong or the EA Forum that very explicitly (e.g., with concrete examples) overviewed what seem to be the most prominent approaches to making decisions under moral uncertainty (more specifically, those covered in Will MacAskill's 2014 thesis).[1][2]

2. There was no (easily findable and explicit) write-up of how to handle simultaneous moral and empirical uncertainty. (What I'll propose is arguably quite obvious, but still seems worth writing up explicitly.)

3. There was no (easily findable and sufficiently thorough) write-up of applying sensitivity analysis and value of information analysis to situations of moral uncertainty.

I therefore decided to write a series of three posts, each of which addressed one of those apparent "gaps". My primary aim is to synthesise and make accessible various ideas that are currently mostly buried in the philosophical literature, but I also think it's plausible that some of the ideas in some of the posts (though not this first one) haven't been explicitly explored before.

I expect that these posts are most easily understood if read in order, but each post should also have value if read in isolation, especially for readers who are already familiar with key ideas from work on moral uncertainty.

Epistemic status (for the whole sequence)

I've now spent several days reading about moral uncertainty, but I wouldn't consider myself an actual expert in this topic or in philosophy more broadly. Thus, while I don't expect this sequence to contain any *major*, *central* mistakes, I wouldn't be surprised if it's inaccurate or unclear/misleading in some places.

I welcome feedback of all kinds (on these posts and in general!).

Moral uncertainty

We are often forced to make decisions under conditions of uncertainty. This uncertainty can be empirical (e.g., what is the likelihood that nuclear war would cause human extinction?) or moral (e.g., does the wellbeing of future generations matter morally?).[3][4] The issue of making decisions under empirical uncertainty has been well-studied,

and expected utility theory has emerged as the typical account of how a rational agent should proceed in these situations. The issue of making decisions under *moral* uncertainty appears to have received less attention (though see this list of relevant papers), despite also being of clear importance.

I'll later publish a post on definitions, types, and sources of moral uncertainty. In the present post, I'll instead aim to convey a sense of what moral uncertainty is through various examples. One example (which I'll return to repeatedly) is the following:

Devon's decision

> Suppose Devon assigns a 25% probability to T1, a version of hedonistic utilitarianism in which human "hedons" (a hypothetical unit of pleasure) are worth 10 times more than fish hedons. He also assigns a 75% probability to T2, a different version of hedonistic utilitarianism, which values human hedons just as much as T1 does, but doesn't value fish hedons at all (i.e., it sees fish experiences as having no moral significance). Suppose also that Devon is choosing whether to buy a fish curry or a tofu curry, and that he'd enjoy the fish curry about twice as much. (Finally, let's go out on a limb and assume Devon's humanity.)
>
> According to T1, the choice-worthiness (roughly speaking, the rightness or wrongness of an action) of buying the fish curry is -90 (because it's assumed to cause 1,000 negative fish hedons, valued as -100, but also 10 human hedons due to Devon's enjoyment).[5] In contrast, according to T2, the choice-worthiness of buying the fish curry is 10 (because this theory values Devon's joy as much as T1 does, but doesn't care about the fish's experiences). Meanwhile, the choice-worthiness of the tofu curry is 5 according to both theories (because it causes no harm to fish, and Devon would enjoy it half as much as he'd enjoy the fish curry).

The choice-worthiness of each option according to each theory is summarised in the following table:

	T1 - 25% credence	T2 - 75% credence
Fish curry	-90	10
Tofu curry	5	5

Given this information, what should Devon do?

"My Favourite Theory"

Multiple approaches to handling moral uncertainty have been proposed. The simplest option is the "My Favourite Theory" (MFT) approach, in which we essentially ignore our moral uncertainty, and just do whatever seems best based on the theory in which one has the highest "credence" (belief). In the above situation, MFT would suggest Devon should buy the fish curry, even though doing so is only somewhat better according to T2 (10 - 5 = 5), and is *far* worse (5 - -90 = 95) according to another theory in which he

has substantial (25%) credence. **Indeed, even if Devon had *49%* credence in T1 (vs 51% in T2), and the difference in the choice-worthiness of the options was *a thousand times* as large according to T1 as according to T2, MFT would still ignore the fact the situation is so much "higher stakes" for T1 than T2, refuse to engage in any "moral hedging", and advise Devon proceed with whatever T2 advised.**

On top of generating such counterintuitive results, MFT is subject to other quite damning objections (see pages 20-25 of Will MacAskill's 2014 thesis). Thus, the remainder of this post will focus on other approaches to moral uncertainty, which *do* allow for "moral hedging".

Types of moral theories

Which approach to moral uncertainty should be used depends in part on what types of moral theories are under consideration by the decision-maker - in particular, whether the theories are *cardinally measurable* or only *ordinally measurable*, and, if cardinally measurable, whether or not they're *inter-theoretically comparable*.[6]

Cardinality

Essentially, a theory is cardinally measurable if it can tell you not just which outcome is better than which, but also *by how much*. E.g., it can tell you not just that "X is better than Y which is better than Z", but also that "X is 10 'units' better than Y, which is 5 'units' better than Z". (Some readers may be more familiar with distinctions between ordinal, *interval*, and *ratio* scales; I'm almost certain "cardinal" scales include both interval and ratio scales.)

My understanding is that popular consequentialist theories are typically cardinal, while popular non-consequentialist theories are typically (or at least more often) ordinal. For example, a Kantian theory may simply tell you that lying is worse than not lying, but not by how much, so you cannot directly weigh that "bad" against the goodness/badness of other actions/outcomes (whereas such comparisons are relatively easy under most forms of utilitarianism).

Intertheoretic comparability

Even if a set of theories *are* cardinal, they still may not be *inter-theoretically comparable*. Roughly speaking, two theories are comparable if there's a consistent, non-arbitrary "exchange rate" between the theories' "units of choice-worthiness" (and they're non-comparable if there isn't). MacAskill explains the "problem of intertheoretic comparisons" as follows:

> "even when all theories under consideration give sense to the idea of magnitudes of choice-worthiness, we need to be able to compare these magnitudes of choice-worthiness across different theories. But it seems that we can't always do this. [... Sometimes we don't know] how can we compare the seriousness of the wrongs, according to these different theories[.] For which theory is there more at stake?"

In his own thesis, Tarsney provides useful examples:

> "Consider, for instance, hedonistic and preference utilitarianism, two straightforward maximizing consequentialist theories that agree on every feature of morality, except that hedonistic utilitarianism regards pleasure and pain as the sole non-derivative bearers of moral value while preference utilitarianism regards satisfied and dissatisfied preferences as the sole non-derivative bearers of moral value. Both theories, we may stipulate, have the same cardinal structure. But this structure does not answer the crucial question for expectational reasoning, how the value of a hedon according to hedonic utilitarianism compares to the value of a preference utile according to preference utilitarianism—that is, for an agent who divides her beliefs equally between the two theories and wishes to hedge when they conflict, how much hedonic experience does it take to offset the dissatisfaction of a preference of a given strength (or vice versa)?
>
> Likewise, of course, in trolley problem situations that pit consequentialist and deontological theories against one another, even if we could overcome the apparent structural incompatibility of these rival theories, the thorniest question seems to be: How many net lives must be saved, according to some particular version of consequentialism, to offset the wrongness of killing an innocent person, according to some particular version of deontology?" (line break added)[7]

It's worth noting that similar issues have received attention from, and are relevant to, other fields as well. For example, MacAskill writes: "A similar problem arises in the study of social welfare in economics: it is desirable to be able to compare the strength of preferences of different people, but even if you represent preferences by cardinally measurable utility functions you need more information to make them comparable." Thus, concepts and findings from those fields could illuminate this matter, and vice versa.

Three approaches

In MacAskill's thesis, the approaches to moral uncertainty he argues for are:

1. Maximising Expected Choice-worthiness (MEC), if all theories under consideration by the decision-maker are cardinal and intertheoretically comparable. (This is arguably the “best” situation to be in, as it is the case in which the most information is being provided by the theories.)
2. Variance Voting (VV), a form of what I’ll call “Normalised MEC”, if all theories under consideration are cardinal but *not* intertheoretically comparable.
3. The Borda Rule (BR), if all theories under consideration are ordinal. (This is the situation in which the *least* information is being provided by the theories.)
4. A “Hybrid” procedure, if the theories under consideration differ in whether they’re cardinal or ordinal and/or in whether they’re intertheoretically comparable. (Hybrid procedures will not be discussed in this post; interested readers can refer to pages 117-122 of MacAskill’s thesis.)

I will focus on these approaches (excluding Hybrid procedures), both because these approaches seem to me to be relatively prominent, effective, and intuitive, and because I know less about other approaches. (Potentially promising alternatives include a bargaining-theoretic approach [related presentation slides here], the similar but older and less fleshed-out parliamentary model, and the approaches discussed in Tarsney's thesis.)

Maximising Expected Choice-worthiness (MEC)

MEC is essentially an extension of expected utility theory. MacAskill describes MEC as follows:

> "when all [normative/moral] theories [under consideration by the decision-maker] are cardinally measurable and intertheoretically comparable, the appropriateness of an option is given by its expected choice-worthiness, where the expected choice-worthiness (EC) of an option is as follows:
>
> $$EC(A) = \sum_{i=1}^{n} C(T_i)CW_i(A)$$
>
> The appropriate options are those with the highest expected choice-worthiness."

In this formula, *C(Ti)* represents the decision-maker's credence (belief) in *Ti* (some particular moral theory), while *CWi(A)* represents the "choice-worthiness" (*CW*) of *A* (an "option" or action that the decision-maker can take), according to *Ti*.

To illustrate how MEC works, we will return to the example of Devon deciding whether to buy a fish curry or tofu curry, as summarised in the table of choice-worthiness values from earlier:

	T1 - 25% credence	T2 - 75% credence
Fish curry	-90	10
Tofu curry	5	5

(I've also modelled this example in Guesstimate. In that link, for comparison purposes, this model is followed by a model of the same basic example using traditional expected utility reasoning, and another using MEC-E (an approach I'll explain in my next post).)

Using MEC in this situation, the expected choice-worthiness of buying the fish curry is 0.25 * -90 + 0.75 * 10 = -15, and the expected choice-worthiness of buying the tofu curry is 0.25 * 5 + 0.75 * 5 = 5. Thus, Devon should buy the tofu curry.

This is despite Devon believing that T2 is more likely than T1, and T2 claiming that buying the fish curry is better than purchasing the tofu curry. The reason is that, as discussed earlier, there is far "more at stake" for T1 than for T2 in this example.

To me, this seems like a good, intuitive result for MEC, and shows how it improves upon the “My Favourite Theory” approach.

There are two final things I should note about MEC:

- MEC can be used in exactly the same way when more than two theories are under consideration. (The only reason most examples in this sequence will be ones in which only two moral theories are under consideration is to keep explanations simple.)
- **The basic idea of MEC can also be used as a heuristic, without involving actual numbers.**
    - For example, say Clara believes that there’s a “high chance” utilitarianism is correct, but that some deontological theory, in which lying is deeply wrong, is “plausible”. Clara is considering whether to tell a lie, and has good reason to believe this will lead to a slight net increase in wellbeing. She might still decide not to lie, despite believing it’s likely that lying is the “right” thing to do, because it’d only be *slightly right*, whereas it’s plausible it’s *deeply wrong*.

Another example of applying MEC (which is probably only worth reading if the approach still seems unclear to you) can be found in the following footnote.[8]

Normalised MEC and Variance Voting

(*It's possible I've made mistakes in this section; if you think I have, please let me know.*)

But what about cases in which, despite being cardinal, the theories you have credence in are *not intertheoretically comparable*? (Recall that this essentially means that there's no consistent, non-arbitrary “exchange rate” between the theories' “units of choice-worthiness".)

MacAskill argues that, in such situations, one must first "normalise" the theories in some way (which basically means "adjusting values measured on different scales to a notionally common scale"). MEC can then be applied just as we saw earlier, but now with the new, normalised choice-worthiness scores.

There are multiple ways one could normalise the theories under consideration (e.g., by range), but MacAskill argues for normalising by variance. That is, he argues that we should:

> “[treat] the average of the squared differences in choice-worthiness from the mean choice-worthiness as the same across all theories. Intuitively, the variance is a measure of how spread out choice-worthiness is over different options; normalising at variance is the same as normalising at the difference between the mean choice-worthiness and one standard deviation from the mean choice-worthiness.”

MacAskill uses the term Variance Voting to refer to this process of first normalising by variance and then using the MEC approach.

(Unfortunately, as far as I could tell, none of the three theses/papers I read that referred to normalising moral theories by variance actually provided a clear, worked

example. I've attempted to construct such a worked example based on an extension of the scenario with Devon deciding what meal to buy; that can be found here, and here is a simpler and I think effectively identical method, suggested in a private message.)

In arguing for Variance Voting over its alternatives, MacAskill states that the basic principle normalisation aims to capture is the "*principle of equal say*: the idea, stated imprecisely for now, that we want to give equally likely moral theories equal weight when considering what it's appropriate to do" (emphasis in original). He further writes:

> "To see a specific case of how this could go awry, consider average and total utilitarianism, and assume that they are indeed incomparable. And suppose that, in order to take an expectation over those theories, we choose to treat them as agreeing on the choice-worthiness ordering of options concerning worlds with only one person in them. If we do this, then, for almost all decisions about population ethics, the appropriate action will be in line with what total utilitarianism regards as most choiceworthy because, for almost all decisions, the stakes are huge for total utilitarianism, but not very large for average utilitarianism. So it seems that, if we treat the theories in this way, we are being partisan to total utilitarianism.
>
> In contrast, if we chose to treat the two theories as agreeing on the choice-worthiness differences between options with worlds involving 10^100 people then, for almost all real-world decisions, what it's appropriate to do will be the same as what average utilitarianism regards as most choice-worthy. This is because we're representing average utilitarianism as claiming that, for almost all decisions, the stakes are much higher than for total utilitarianism. In which case, it seems that we are being partisan to average utilitarianism, whereas what we want is to have a way of normalising such that each theory gets equal influence." (line break added)

(Note that it's not a problem for one theory to have much more influence on decisions *due to higher credence in that theory*. The principle of equal say is only violated if additional influence is unrelated to additional credence in a theory, and instead has to do with what are basically *arbitrary/accidental choices about exchange rates* between units of choice-worthiness.)

MacAskill (pages 110-116) provides two arguments that VV is the approach that satisfies the principle of equal say, and Owen Cotton-Barratt similarly argues for the superiority of normalisation by variance over alternative normalisations. (But note that this approach does seem to have its flaws, as discussed in, e.g., pages 222-223 of Tarsney's thesis.)

The Borda Rule (BR)

Finally, what about cases in which all moral theories you have credence in are only ordinal, rather than cardinal (i.e., they say only whether each option is more, equally, or less choice-worthy than each other option, but not by how much)? For such cases, MacAskill recommends a voting method called the Borda Rule (BR; also known as a "Borda count"), with "moral theories play[ing] the part of voters and practical options the part of candidates" (Tarsney). I will first quote MacAskill's formal explanation of BR (which may be somewhat confusing by itself), before quoting an example he gives and showing what applying BR to that looks like:

> "An option *A*'s *Borda Score*, for any theory *Ti*, is equal to the number of options within the option-set that are less choice-worthy than *A* according to theory *Ti*'s

choice-worthiness function, minus the number of options within the option-set that are more choice-worthy than *A* according to *Ti*'s choice-worthiness function.[9]

An option *A*'s *Credence-Weighted Borda Score* is the sum, for all theories *Ti*, of the Borda Score of *A* according to theory *Ti* multiplied by the credence that the decision-maker has in theory *Ti*.

[The *Borda Rule* states that an] option *A* is more appropriate than an option *B* iff [if and only if] *A* has a higher Credence-Weighted Borda Score than *B*; *A* is equally as appropriate as *B* iff *A* and *B* have an equal Credence-Weighted Borda Score."

I will now show, following MacAskill, how this rule applies to an example he gives in his thesis:

"Julia is a judge who is about to pass a verdict on whether Smith is guilty for murder. She is very confident that Smith is innocent. There is a crowd outside, who are desperate to see Smith convicted. Julia has three options:

[G]: Pass a verdict of 'guilty'.

[R]: Call for a retrial.

[I]: Pass a verdict of 'innocent'.

Julia knows that the crowd will riot if Smith is found innocent, causing mayhem on the streets and the deaths of several people. If she calls for a retrial, she knows that he will be found innocent at a later date, and that it is much less likely that the crowd will riot at that later date. If she declares Smith guilty, the crowd will be appeased and go home peacefully. She has credence in three moral theories:

35% credence in a variant of utilitarianism, according to which [G>R>I].

34% credence in a variant of common sense, according to which [R>I>G].

31% credence in a deontological theory, according to which [I>R>G]."

The options' Borda Scores according to each theory, and their Credence-Weighted Borda Scores, are therefore as shown in the following table:

	Utilitarian theory - 35%	Common sense theory - 34%	Deontological theory - 31%	Credence-Weighted Borda Score
Guilty	2 - 0 = **2**	0 - 2 = **-2**	0 - 2 = **-2**	**-0.6**
Retrial	1 - 1 = **0**	2 - 0 = **2**	1 - 1 = **0**	**0.68**
Innocent	0 - 2 = **-2**	1 - 1 = **0**	2 - 0 = **2**	**-0.08**

(For example, G has a score of 2 - 0 = 2 according to utilitarianism because that theory views two options as less choice-worthy than G, and 0 options as more choice-worthy than G.)

The calculations that provided the Credence-Weighted Borda Scores shown in the above table are as follows:

G: 0.35 * 2 + 0.34 * -2 + 0.31 * -2 = -0.6 (this because the utilitarian, common sense, and deontological theories are given credences of 35%, 34%, and 31%, respectively, and these serve as the weightings for the Borda Scores these theories provide)

R: 0.35 * 0 + 0.34 * 2 + 0.31 * 0 = 0.68

I: 0.35 * -2 + 0.34 * 0 + 0.31 * 2 = -0.08

BR would therefore claim that Julia should call for a retrial. **This is the case even though passing a guilty verdict was seen as best by Julia's "favourite theory" (the variant of utilitarianism). Essentially, calling for a retrial is preferred because both passing a guilty verdict and passing an innocent verdict were seen as *least* preferred by some theory Julia has substantial credence in, whereas calling for a retrial is not *least* preferred by any theory.**

MacAskill notes that preferring this sort of a compromise option in a case like this seems intuitively right. He also argues that alternatives to BR fail to give us the sort of answers we'd want in these or other sorts of cases. (Though Tarsney raises some objections to BR which I won't get into.)

Closing remarks

I hope you have found this post a useful, clear summary of key ideas around what moral uncertainty is, why it matters, and how to make decisions when morally uncertain. Personally, I believe that an understanding of moral uncertainty - particularly a sort of heuristic version of MEC - has usefully enriched my thinking, and influenced some of the biggest decisions I've made over the last year.[10]

In the next post, I will discuss (possibly novel, arguably obvious) extensions of each of the three approaches discussed here, in order to allow for modelling *both moral and empirical uncertainty, explicitly and simultaneously*. The post after that will discuss how we can combine the approaches in the first two posts with sensitivity analysis and value of information analysis.[11][12]

---

1. I genuinely mean no disrespect to the several posts on moral uncertainty I did discover (e.g., here, here, and here). All did meet some of those criteria, and I'd say most were well-written but just weren't highly explicit (e.g., didn't include enough concrete examples), and/or didn't cover (in the one post) each of the prominent approaches and the related ideas necessary to understand them. ↩

2. Other terms/concepts that are sometimes used and are similar to "moral uncertainty" are *normative*, *axiological*, and *value* uncertainty. In this sequence, I'll use "moral uncertainty" in a general sense that also incorporates axiological and value uncertainty, and at least a large part of normative uncertainty.

   Also, throughout this sequence, I will use the term "approach" in a way that I believe aligns with MacAskill's use of the term "metanormative theory". ↩

3. It seems to me that there are many cases where it's not entirely clear whether the uncertainty is empirical or moral. For example, I might wonder "Are fish conscious?", which seems on the face of it an empirical question. However, I

might not yet know precisely what I mean by “conscious”, and only really want to know whether fish are “conscious in a sense I would morally care about”. In this case, the seemingly empirical question becomes hard to disentangle from the (seemingly moral) question “What forms of consciousness are morally important?”

(Furthermore, my answers to *that* question in turn may be influenced by empirical discoveries. For example, I may initially believe avoidance of painful stimuli demonstrates consciousness in a morally relevant sense, but then change that belief after learning that this behaviour can be displayed in a stimulus-response way by certain extremely simple organisms.)

In such cases, I believe the approach suggested in the next post of this sequence will still work well, as that approach does not really require empirical and moral uncertainty to be treated fundamentally differently. (Another approach, which presents itself differently but I think is basically the same in effect, is to consider uncertainty over “worldviews”, with those worldviews combining moral and empirical claims.) ↩

4. In various places in this sequence, I will use language that may appear to endorse or presume moral realism (e.g., referring to “moral information” or to probability of a particular moral theory being “true”). But this is essentially just for convenience; I intend this sequence to be neutral on the matter of moral realism vs antirealism, and I believe this post can be useful in mostly similar ways regardless of one’s position on that matter. I discuss the matter of "moral uncertainty for antirealists" in more detail in this separate post. ↩

5. The matter of how to actually assign “units” or “magnitudes” of choice-worthiness to different options, and what these things would even mean, is complex, and I won’t really get into it in this sequence. ↩

6. Christian Tarsney's 2017 thesis thesis (e.g., pages 175-176) explains other ways the "structure" of moral theories can differ, and potential implications of these other differences. These were among the juicy complexities I had to resist cramming in this originally-intended-as-bitesized post (but I may write another post about Tarsney's ideas later; please let me know if you think that'd be worthwhile). ↩

7. It's worth noting that similar issues have received attention from, and are relevant to, other fields as well. For example, MacAskill writes: "A similar problem arises in the study of social welfare in economics: it is desirable to be able to compare the strength of preferences of different people, but even if you represent preferences by cardinally measurable utility functions you need more information to make them comparable." Thus, concepts and findings from those fields could illuminate this matter, and vice versa. ↩

8. Suppose Alice assigns a 60% probability to hedonistic utilitarianism (HU) being true and a 40% probability to preference utilitarianism (PU) being true. Suppose also that Bob *wants* to play video games, but would actually *get slightly more joy* out of a day at the beach. Thus, according to HU, letting Bob play video games has a CW of 5, and taking him to the beach has a CW of 6; while according to PU, letting Bob play video games has a CW of 15, and taking him to the beach has a CW of -20.

Under these conditions, the expected choice-worthiness of letting Bob play video games is 0.6 * 5 + 0.4 * 15 = 9, and the expected choice-worthiness of taking Bob to the beach is 0.6 * 6 + 0.4 * -20 = -4.4. Therefore, Alice should let Bob play video games.

Analogously to the situation with the Devon example, this is despite Alice believing HU is more likely than PU, and despite HU positing that taking Bob to the beach being better than letting him play video games. As before, the reason is that there is "more at stake" in this decision for the less-believed theory than for the more-believed theory; HU considers there to only be a very small difference between the choice-worthiness of the options, while PU considers there to be a large difference. ↩

9. MacAskill later notes that a simpler method (which doesn't subtract the number of options that are more choice-worthy) can be used when there are no ties. His calculations for the example I quote and work through in this post use that simpler method. But in this post, I'll stick to the method MacAskill describes in this quote (which is guaranteed to give the same final answer in this example anyway). ↩

10. However, these concepts are of course not an instant fix or cure-all. In a (readable and interesting) 2019 paper, MacAskill writes "so far, the implications for practical ethics have been drawn too simplistically [by some philosophers.] First, the implications of moral uncertainty for normative ethics are far more wide-ranging than has been noted so far. Second, one can't straightforwardly argue from moral uncertainty to particular conclusions in practical ethics, both because of 'interaction' effects between moral issues, and because of the variety of different possible intertheoretic comparisons that one can reasonably endorse."

    For a personal example, a heuristic version of MEC still leaves me unsure whether I should move from being a vegetarian-flirting-with-veganism to a strict vegan, or even whether I should spend much time making that decision, because that might trade off to some extent with time and money I could put towards longtermist efforts (which seem more choice-worthy according to other moral theories I have some credence in). I suspect any quantitative modelling simple enough to be done in a reasonable amount of time would still leave me unsure.

    That said, I, like MacAskill (in the same paper), "do believe, however, that consideration of moral uncertainty should have major impacts for how practical ethics is conducted. [...] It would be surprising if the conclusions [of approaches taking moral uncertainty into account] were the same as those that practical ethicists typically draw."

    In particular, I'd note that considering moral uncertainty can reveal some "low-hanging fruit": some "trades" between moral theories that are relatively clearly advantageous, due to large differences in the "stakes" different moral theories see the situation as having. (Personally, cases of apparent low-hanging fruit of this kind have included becoming at least vegetarian, switching my career aims to longtermist ones, and yet engaging in global-poverty-related movement-building when an unusual opportunity arose and it wouldn't take up too much of my time.) ↩

11. To foreshadow: Basically, my idea is that, once you've made explicit your degree of belief in various moral theories and how good/bad outcomes appear to each of

those theories, you can work out which updates to your beliefs in moral theories or to your understandings of those moral theories are most likely to change your decisions, and thus which "moral learning" to prioritise and how much resources to expend on it. ↩

12. I'm also considering later adding posts on:

    - Different types and sources of moral uncertainty (drawing on these posts).
    - The idea of ignoring even very high credence in nihilism, because it's never decision-relevant.
    - Whether it could make sense to give moral realism disproportionate (compared to antirealism) influence over our decisions, based on the idea that realism might view there as "more at stake" than antirealism does.

    I'd be interested in hearing whether people think those threads are likely to be worth pursuing. ↩

Making decisions when both morally and empirically uncertain

Cross-posted to the EA Forum. For an epistemic status statement and an outline of the purpose of this sequence of posts, please see the top of my prior post. There are also some explanations and caveats in that post which I won't repeat - or will repeat only briefly - in this post.

Purpose of this post

In my prior post, I wrote:

> We are often forced to make decisions under conditions of uncertainty. This uncertainty can be empirical (e.g., what is the likelihood that nuclear war would cause human extinction?) or moral (e.g., does the wellbeing of future generations matter morally?). The issue of making decisions under empirical uncertainty has been well-studied, and expected utility theory has emerged as the typical account of how a rational agent should proceed in these situations. The issue of making decisions under *moral* uncertainty appears to have received less attention (though see this list of relevant papers), despite also being of clear importance.

I then went on to describe three prominent approaches for dealing with moral uncertainty (based on Will MacAskill's 2014 thesis):

1. Maximising Expected Choice-worthiness (MEC), if all theories under consideration by the decision-maker are cardinal and intertheoretically comparable.[1]
2. Variance Voting (VV), a form of what I'll call "Normalised MEC", if all theories under consideration are cardinal but *not* intertheoretically comparable.[2]
3. The Borda Rule (BR), if all theories under consideration are ordinal.

But I was surprised to discover that I couldn't find any *very explicit* write-up of how to handle moral and empirical uncertainty *at the same time*. I assume this is because most people writing on relevant topics consider the approach I will propose in this post to be quite obvious (at least when using MEC with cardinal, intertheoretically comparable, consequentialist theories). Indeed, many existing models from EAs/rationalists (and likely from other communities) already effectively use something very much like the first approach I discuss here ("MEC-E"; explained below), just without explicitly noting that this is an integration of approaches for dealing with moral and empirical uncertainty.[3]

But it still seemed worth explicitly spelling out the approach I propose, which is, in a nutshell, using exactly the regular approaches to moral uncertainty mentioned above, but on *outcomes* rather than on *actions*, and combining that with consideration of the likelihood of each action leading to each outcome. My aim for this post is both to make this approach "obvious" to a broader set of people and to explore how it can work with non-comparable, ordinal, and/or non-consequentialist theories (which may be less obvious).

(Additionally, as a side-benefit, readers who are wondering what on earth all this "modelling" business some EAs and rationalists love talking about is, or who are only somewhat familiar with modelling, may find this post to provide useful examples and explanations.)

I'd be interested in any comments or feedback you might have on anything I discuss here!

MEC under empirical uncertainty

To briefly review regular MEC: MacAskill argues that, when all moral theories under consideration are cardinal and intertheoretically comparable, a decision-maker should choose the "option" that has the highest *expected choice-worthiness*. Expected choice-worthiness is given by the following formula:

$$EC(A) = \sum_{i=1}^{n} C(T_i)CW_i(A)$$

In this formula, *C(Ti)* represents the decision-maker's credence (belief) in *Ti* (some particular moral theory), while *CWi(A)* represents the "choice-worthiness" (*CW*) of *A* (an "option" or action that the decision-maker can choose) according to *Ti*. In my prior post, I illustrated how this works with this example:

> Suppose Devon assigns a 25% probability to T1, a version of hedonistic utilitarianism in which human "hedons" (a hypothetical unit of pleasure) are worth 10 times more than fish hedons. He also assigns a 75% probability to T2, a different version of hedonistic utilitarianism, which values human hedons just as much as T1 does, but doesn't value fish hedons at all (i.e., it sees fish experiences as having no moral significance). Suppose also that Devon is choosing whether to buy a fish curry or a tofu curry, and that he'd enjoy the fish curry about twice as much. (Finally, let's go out on a limb and assume Devon's humanity.)
>
> According to T1, the choice-worthiness (roughly speaking, the rightness or wrongness of an action) of buying the fish curry is -90 (because it's assumed to cause 1,000 negative fish hedons, valued as -100, but also 10 human hedons due to Devon's enjoyment). In contrast, according to T2, the choice-worthiness of buying the fish curry is 10 (because this theory values Devon's joy as much as T1 does, but doesn't care about the fish's experiences). Meanwhile, the choice-worthiness of the tofu curry is 5 according to both theories (because it causes no harm to fish, and Devon would enjoy it half as much as he'd enjoy the fish curry).
>
> [...] Using MEC in this situation, the expected choice-worthiness of buying the fish curry is 0.25 * -90 + 0.75 * 10 = -15, and the expected choice-worthiness of buying the tofu curry is 0.25 * 5 + 0.75 * 5 = 5. Thus, Devon should buy the tofu curry.

But can Devon really be *sure* that buying the fish curry will lead to that much fish suffering? What if this demand signal doesn't lead to increased fish farming/capture? What if the additional fish farming/capture is more humane than expected? What if fish can't suffer because they aren't actually conscious (empirically, rather than as a result of what sorts of consciousness our moral theory considers relevant)? We could likewise question Devon's apparent certainty that buying the tofu curry *definitely won't* have any unintended consequences for fish suffering, and his apparent certainty regarding precisely how much he'd enjoy each meal.

These are all empirical rather than moral questions, but they still seem very important for Devon's ultimate decision. This is because **T1 and T2 don't "*intrinsically* care" about whether someone buys fish curry or buys tofu curry; these theories assign no *terminal value* to which curry is bought. Instead, these theories "care" about some of the *outcomes* which those actions may or may not cause.**[4]

More generally, I expect that, in all realistic decision situations, we'll have *both* moral *and* empirical uncertainty, and that it'll often be important to *explicitly consider both types of uncertainties*. For example, GiveWell's models consider both how likely insecticide-treated bednets are to save the life of a child, and how that outcome would compare to doubling the income of someone in extreme poverty. However, typical discussions of MEC seem to assume that we already know for sure what the outcomes of our actions will be, just as typical discussions of expected value reasoning seem to assume that we already know for sure how valuable a given outcome is.

Luckily, it seems to me that MEC and traditional (empirical) expected value reasoning can be very easily and neatly integrated in a way that resolves those issues. (This is perhaps partly due to that fact that, if I understand MacAskill's thesis correctly, MEC was very consciously developed by analogy to expected value reasoning.) Here is my formula for this integration, which I'll call *Maximising Expected Choice-worthiness, accounting for Empirical uncertainty (MEC-E)*, and which I'll explain and provide an example for below:

$$EC(A) = \sum_{i=1}^{n} \sum_{j=1}^{n} P(O_j \mid A) CW_i(O_j) C(T_i)$$

Here, all symbols mean the same things they did in the earlier formula from MacAskill's thesis, with two exceptions:

- I've added *Oj*, to refer to each "outcome": each consequence that an action may lead to, which at least one moral theory under consideration intrinsically values/disvalues. (E.g., a fish suffering; a person being made happy; rights being violated.)
- Related to that, I'd like to be more explicit that *A* refers *only* to the "actions" that the decision-maker can directly choose (e.g., purchasing a fish meal, imprisoning someone), rather than the outcomes of those actions.[5]

(I also re-ordered the choice-worthiness term and the credence term, which makes no actual difference to any results, and was just because I think this ordering is slightly more intuitive.)

Stated verbally (and slightly imprecisely[6]), MEC-E claims that:

> One should choose the action which maximises expected choice-worthiness, accounting for empirical uncertainty. To calculate the expected choice-worthiness of each action, you first, for each potential outcome of the action and each moral theory under consideration, find the product of 1) the probability of that outcome given that that action is taken, 2) the choice-worthiness of that outcome according to that theory, and 3) the credence given to that theory. Second, for each action, you sum together all of those products.

To illustrate, I have modelled in Guesstimate an extension of the example of Devon deciding what meal to buy to also incorporate empirical uncertainty.[7] In the text here, I will only state the information that was not in the earlier version of the example, and the resulting calculations, rather than walking through all the details.

Suppose Devon believes there's an 80% chance that buying a fish curry will lead to "fish being harmed" (modelled as 1000 negative fish hedons, with a choice-worthiness of -100 according to T1 and 0 according to T2), and a 10% chance that buying a tofu curry will lead to that same outcome. He also believes there's a 95% chance that buying a fish curry will lead to "Devon enjoying a meal a lot" (modelled as 10 human hedons), and a 50% chance that buying a tofu curry will lead to that.

The expected choice-worthiness of buying a fish curry would therefore be:

(0.8 * -100 * 0.25) + (0.8 * 0 * 0.75) + (0.95 * 10 * 0.25) + (0.95 * 10 * 0.75) = -10.5

Meanwhile, the expected choice-worthiness of buying a tofu curry would be:

(0.1 * -100 * 0.25) + (0.1 * 0 * 0.75) + (0.5 * 10 * 0.25) + (0.5 * 10 * 0.75) = 2.5

As before, the tofu curry appears the better choice, despite seeming somewhat worse according to the theory (T2) assigned higher credence, because the other theory (T1) sees the tofu curry as *much* better.

In the final section of this post, I discuss potential extensions of these approaches, such as how it can handle probability distributions (rather than point estimates) and non-consequentialist theories.

The last thing I'll note about MEC-E in this section is that MEC-E can be used as a heuristic, without involving actual numbers, in exactly the same way MEC or traditional expected value reasoning can. For example, without knowing or estimating any actual numbers, Devon might reason that, compared to buying the tofu curry, buying the fish curry is "much" more likely to lead to fish suffering and only "somewhat" more likely to lead to him enjoying his meal a lot. He may further reason that, in the "unlikely but plausible" event that fish experiences *do* matter, the badness of a large amount of fish suffering is "much" greater than the goodness of him enjoying a meal. He may thus ultimately decide to purchase the tofu curry.

(Indeed, my impression is that many effective altruists have arrived at vegetarianism/veganism through reasoning very much like that, without any actual numbers being required.)

Normalised MEC under empirical uncertainty

(From here onwards, I've had to go a bit further beyond what's clearly implied by existing academic work, so the odds I'll make some mistakes go up a bit. Please let me know if you spot any errors.)

To briefly review regular *Normalised MEC*: Sometimes, despite being cardinal, the moral theories we have credence in are *not intertheoretically comparable* (basically meaning that there's no consistent, non-arbitrary "exchange rate" between the theories' "units of choice-worthiness"). MacAskill argues that, in such situations, one must first "normalise" the theories in some way (i.e., "[adjust] values measured on different scales to a notionally common scale"), and then apply MEC to the new, normalised choice-worthiness scores. He recommends Variance Voting, in which the normalisation is by variance (rather than, e.g., by range), meaning that we:

> "[treat] the average of the squared differences in choice-worthiness from the mean choice-worthiness as the same across all theories. Intuitively, the variance is a measure of how spread out choice-worthiness is over different options; normalising at variance is the same as normalising at the difference between the mean choice-worthiness and one standard deviation from the mean choice-worthiness."

(I provide a worked example here, based on an extension of the scenario with Devon deciding what meal to buy, but it's possible I've made mistakes.)

My proposal for *Normalised MEC, accounting for Empirical Uncertainty (Normalised MEC-E)* is just to combine the ideas of non-empirical Normalised MEC and non-normalised MEC-E in a fairly intuitive way. The steps involved (which may be worth reading alongside this worked example and/or the earlier explanations of Normalised MEC and MEC-E) are as follows:

1. Work out expected choice-worthiness just as with regular MEC, except that here one is working out the expected choice-worthiness of *outcomes*, not actions. I.e., for each outcome, multiply that outcome's choice-worthiness according to each theory by your credence in that theory, and then add up the resulting products.
    - You could also think of this as using the MEC-E formula, except with "Probability of outcome given action" removed for now.
2. Normalise these expected choice-worthiness scores by variance, just as MacAskill advises in the quote above.
3. Find the "expected value" of each action in the traditional way, with these normalised expected choice-worthiness scores serving as the "value" for each potential outcome. I.e., for each *action*, multiply the probability it leads to each outcome by the normalised expected choice-worthiness of that outcome (from step 2), and then add up the resulting products.
    - You could think of this as bringing "Probability of outcome given action" back into the MEC-E formula.
4. Choose the action with the maximum score from step 3 (which we could call *normalised expected choice-worthiness, accounting for empirical uncertainty*, or *expected value, accounting for normalised moral uncertainty*).[8]

BR under empirical uncertainty

The final approach MacAskill recommends in his thesis is the Borda Rule (BR; also known as *Borda counting*). This is used when the moral theories we have credence in are merely *ordinal* (i.e., they don't say "how much" more choice-worthy one option is compared to another). In my prior post, I provided the following quote of MacAskill's formal explanation of BR (here with "options" replaced by "actions"):

> "An [action] *A*'s *Borda Score*, for any theory *Ti*, is equal to the number of [actions] within the [action]-set that are less choice-worthy than *A* according to theory *Ti*'s choice-worthiness function, minus the number of [actions] within the [action]-set that are more choice-worthy than *A* according to *Ti*'s choice-worthiness function.
>
> An [action] *A*'s *Credence-Weighted Borda Score* is the sum, for all theories *Ti*, of the Borda Score of *A* according to theory *Ti* multiplied by the credence that the decision-maker has in theory *Ti*.
>
> [The *Borda Rule* states that an action] *A* is more appropriate than an [action] *B* iff [if and only if] *A* has a higher Credence-Weighted Borda Score than *B*; *A* is equally as appropriate as *B* iff *A* and *B* have an equal Credence-Weighted Borda Score."

To apply BR when one is also *empirically* uncertain, **I propose just explicitly considering/modelling one's empirical uncertainties, and then figuring out each action's Borda Score with those empirical uncertainties in mind**. (That is, we don't change the method at all on a mathematical level; we just make sure each moral theory's preference rankings over actions - which is used as input into the Borda Rule - takes into account our empirical uncertainty about what outcomes each action may lead to.)

I'll illustrate how this works with reference to the same example from MacAskill's thesis that I quoted in my prior post, but now with slight modifications (shown in bold).

> "Julia is a judge who is about to pass a verdict on whether Smith is guilty for murder. She is very confident that Smith is innocent. There is a crowd outside, who are desperate to see Smith convicted. Julia has three options:
>
> [G]: Pass a verdict of 'guilty'.
>
> [R]: Call for a retrial.
>
> [I]: Pass a verdict of 'innocent'.
>
> She thinks there's a 0% chance of M if she passes a verdict of guilty, a 30% chance if she calls for a retrial (there may mayhem due to the lack of a guilty verdict, or later due to a later innocent verdict), and a 70% chance if she passes a verdict of innocent.
>
> There's obviously a 100% chance of C if she passes a verdict of guilty and a 0% chance if she passes a verdict of innocent. She thinks there's also a 20% chance of C happening later if she calls for a retrial.
>
> **Julia believes the crowd is very likely (~90% chance) to** riot if Smith is found innocent, causing mayhem on the streets and the deaths of several people. If she calls for a retrial, she **believes it's almost certain (~95% chance)** that he will be found innocent at a later date, and that it is much less likely **(only ~30% chance)** that the crowd will riot at that later date **if he is found innocent then**. If she declares Smith guilty, the crowd will **certainly (~100%)** be appeased and go home peacefully. She has credence in three moral theories**, which, *when taking the preceding probabilities into account*, provide the following choice-worthiness orderings**:
>
> 35% credence in a variant of utilitarianism, according to which [G>**I>R**].
>
> 34% credence in a variant of common sense, according to which [**I>R**>G].
>
> 31% credence in a deontological theory, according to which [I>R>G]."

This leads to the Borda Scores and Credence-Weighted Borda Scores shown in the table below, and thus to the recommendation that Julia declare Smith innocent.

	Utilitarian theory - 35%	Common sense theory - 34%	Deontological theory - 31%	Credence-Weighted Borda Score
Preference ordering	G>**I>R**	**I>R**>G	I>R>G	
Score for Guilty	2 - 0 = **2**	0 - 2 = **-2**	0 - 2 = **-2**	**-0.6**
Score for Retrial	0 - 2 = **-2**	1 - 1 = **0**	1 - 1 = **0**	**0.7**
Score for Innocent	1 - 1 = **0**	2 - 0 = **2**	2 - 0 = **2**	**1.3**

(More info on how that was worked out can be found in the following footnote, along with the corresponding table based on the moral theories' preference orderings in my prior post, when empirical uncertainty *wasn't* taken into account.[9])

In the original example, both the utilitarian theory and the common sense theory preferred a retrial to a verdict of innocent (in order to avoid a riot), which resulted in calling for a retrial having the highest Credence-Weighted Borda Score.

However, I'm now imagining that **Julia is no longer assuming each action 100% guarantees a certain outcome will occur, and paying attention to her empirical uncertainty has changed her conclusions**.

In particular, I'm imagining that she realises she'd initially been essentially "rounding up" (to 100%) the likelihood of a riot if she provides a verdict of innocent, and "rounding down" (to 0%) the likelihood of the crowd rioting at a later date. However, with more realistic probabilities in mind, utilitarianism and common sense would both actually prefer an innocent verdict to a retrial (because the innocent verdict seems less risky, and the retrial more risky, than she'd initially thought, while an innocent verdict still frees this innocent person sooner and with more certainty). This changes each action's Borda Score, and gives the result that she should declare Smith innocent.[10]

Potential extensions of these approaches

Does this approach presume/privilege consequentialism?

A central idea of this post has been making a clear distinction between "actions" (which one can directly choose to take) and their "outcomes" (which are often what moral theories "intrinsically care about"). This clearly makes sense when the moral theories one has credence in are consequentialist. However, other moral theories may "intrinsically care" about actions themselves. For example, many deontological theories would consider lying to be wrong *in and of itself*, regardless of what it leads to. Can the approaches I've proposed handle such theories?

Yes - and very simply! For example, suppose I wish to use MEC-E (or Normalised MEC-E), and I have credence in a (cardinal) deontological theory that assigns very low choice-worthiness to lying (regardless of outcomes that action leads to). We can still calculate expected choice-worthiness using the formulas shown above; in this case, we find the product of (multiply) "probability me lying leads to me having lied" (which we'd set to 1), "choice-worthiness of me having lied, according to this deontological theory", and "credence in this deontological theory".

Thus, cases where a theory cares intrinsically about the action and not its consequences can be seen as a "special case" in which the approaches discussed in this post just collapse back to the corresponding approaches discussed in MacAskill's thesis (which these approaches are the "generalised" versions of). This is because there's effectively no empirical uncertainty in these cases; we can be sure that taking an action would lead to us having taken that action. Thus, in these and other cases of no relevant empirical uncertainty, accounting for empirical uncertainty is unnecessary, but creates no problems.[11][12]

I'd therefore argue that a policy of using the generalised approaches by default is likely wise. This is especially the case because:

- One will typically have at least *some* credence in consequentialist theories.
- My impression is that even most "non-consequentialist" theories still do care at least *somewhat* about consequences. For example, they'd likely say lying is in fact "right" if the negative consequences of not doing so are "large enough" (and one should often be empirically uncertain about whether they would be).

Factoring things out further

In this post, I modified examples (from my prior post) in which we had only one moral uncertainty into examples in which we had one moral and one empirical uncertainty. We could think of this as "factoring out" what originally appeared to be only moral uncertainty into its "factors": empirical uncertainty about whether an action will lead to an outcome, and moral uncertainty about the value of that outcome. By doing this, we're more closely approximating (modelling) our actual understandings and uncertainties about the situation at hand.

But we're still far from a full approximation of our understandings and uncertainties. For example, in the case of Julia and the innocent Smith, Julia may also be uncertain how big the riot would be, how many people would die, whether these people would be rioters or uninvolved bystanders, whether there's a moral difference between a rioter vs a bystanders dying from the riot (and if so, how big this difference is), etc.[13]

A benefit of the approaches shown here is that they can very simply be extended, with typical modelling methods, to incorporate additional uncertainties like these. You simply disaggregate the relevant variables into the "factors" you believe they're composed of, assign them numbers, and multiply them as appropriate.[14][15]

Need to determine whether uncertainties are moral or empirical?

In the examples given just above, you may have wondered whether I was considering certain variables to represent moral uncertainties or empirical ones. I suspect this ambiguity will be common in practice (and I plan to discuss it further in a later post). Is this an issue for the approaches I've suggested?

I'm a bit unsure about this, but I think the answer is essentially "no". I don't think there's any need to treat moral and empirical uncertainty in *fundamentally* different ways for the sake of models/calculations using these approaches. Instead, I think that, ultimately, the important thing is just to "factor out" variables in the way that makes the most sense, given the situation and what the moral theories under consideration "intrinsically care about". (An example of the sort of thing I mean can be found in footnote 14, in a case where the uncertainty is actually empirical but has different moral implications for different theories.)

Probability distributions instead of point estimates

You may have also thought that a lot of variables in the examples I've given should be represented by probability distributions (e.g., representing 90% confidence intervals), rather than point estimates. For example, why would Devon estimate the probability of "fish being harmed", as if it's a binary variable whose moral significance switches suddenly from 0 to -100 (according to T1) when a certain level of harm is reached? Wouldn't it make more sense for him to estimate the *amount* of harm to fish that is likely, given that that better aligns both with his understanding of reality and with what T1 cares about?

If you were thinking this, I wholeheartedly agree! Further, I can't see any reason why the approaches I've discussed *couldn't* use probability distributions and model variables as continuous rather than binary (the only reason I haven't modelled things in that way so far was to keep explanations and examples simple). For readers interested in an illustration of

how this can be done, I've provided a modified model of the Devon example in this Guesstimate model. (Existing models like this one also take essentially this approach.)

Closing remarks

I hope you've found this post useful, whether to inform your heuristic use of moral uncertainty and expected value reasoning, to help you build actual models taking into account both moral and empirical uncertainty, or to give you a bit more clarity on "modelling" in general.

In the next post, I'll discuss how we can combine the approaches discussed in this and my prior post with sensitivity analysis and value of information analysis, to work out what specific moral or empirical learning would be most decision-relevant and when we should vs shouldn't postpone decisions until we've done such learning.

---

1. What "choice-worthiness", "cardinal" (vs "ordinal"), and "intertheoretically comparable" mean is explained in the previous post. To quickly review, roughly speaking:

    - *Choice-worthiness* is the rightness or wrongness of an action, according to a particular moral theory.
    - A moral theory is *ordinal* if it tells you only which options are better than which other options, whereas a theory is *cardinal* if it tells you *how big a difference* in choice-worthiness there is between each option.
    - A pair of moral theories can be cardinal and yet still *not intertheoretically comparable* if we cannot meaningfully compare the sizes of the "differences in choice-worthiness" between the theories; basically, if there's no consistent, non-arbitrary "exchange rate" between different theories' "units of choice-worthiness".

    ↩

2. MacAskill also discusses a "Hybrid" procedure, if the theories under consideration differ in whether they're cardinal or ordinal and/or whether they're intertheoretically comparable; readers interested in more information on that can refer to pages 117-122 MacAskill's thesis. An alternative approach to such situations is Christian Tarsney's (pages 187-195) "multi-stage aggregation procedure", which I may write a post about later (please let me know if you think this'd be valuable). ↩

3. Examples of models that effectively use something like the "MEC-E" approach include GiveWell's cost-effectiveness models and this model of the cost effectiveness of "alternative foods".

    And some of the academic moral uncertainty work I've read seemed to indicate the authors may be perceiving as obvious something like the approaches I propose in this post.

    But I think the closest thing I found to an explicit write-up of this sort of way of considering moral and empirical uncertainty at the same time (expressed in those terms) was this post from 2010, which states: "Under Robin's approach to value uncertainty, we would (I presume) combine these two utility functions into one linearly, by weighing each with its probability, so we get EU(x) = 0.99 EU1(x) + 0.01 EU2(x)". ↩

4. Some readers may be thinking the "empirical" uncertainty about fish consciousness is inextricable from moral uncertainties, and/or that the above paragraph implicitly presumes/privileges consequentialism. If you're one of those readers, 10 points to you

for being extra switched-on! However, I believe these are not really issues for the approaches outlined in this post, for reasons outlined in the final section. ↩

5. Note that my usage of “actions” can include “doing nothing”, or failing to do some specific thing; I don’t mean “actions” to be distinct from “omissions” in this context. MacAskill and other writers sometimes refer to “options” to mean what I mean by “actions”. I chose the term “actions” both to make it more obvious what the A and O terms in the formula stand for, and because it seems to me that the distinction between “options” and “outcomes” would be less immediately obvious. ↩

6. My university education wasn’t highly quantitative, so it’s very possible I’ll phrase certain things like this in clunky or unusual ways. If you notice such issues and/or have better phrasing ideas, please let me know. ↩

7. In that link, the model using MEC-E follows a similar model using regular MEC (and thus considering only moral uncertainty) and another similar model using more traditional expected value reasoning (and thus considering only empirical uncertainty); readers can compare these against the MEC-E model. ↩

8. Before I tried to actually model an example, I came up with a slightly different proposal for integrating the ideas of MEC-E and Normalised MEC. Then I realised the proposal outlined above might make more sense, and it does seem to work (though I’m not 100% certain), so I didn’t further pursue my original proposal. I therefore don't know for sure whether my original proposal would work or not (and, if it does work, whether it’s somehow better than what I proposed above). My original proposal was as follows:

    1. Work out expected choice-worthiness just as with regular MEC-E; i.e., follow the formula from above to incorporate consideration of the probabilities of each action leading to each outcome, the choice-worthiness of each outcome according to each moral theory, and the credence one has in each theory. (But don’t yet pick the action with the maximum expected choice-worthiness score.)
    2. Normalise these expected choice-worthiness scores by variance, just as MacAskill advises in the quote above. (The fact that these scores incorporate consideration of empirical uncertainty has no impact on how to normalise by variance.)
    3. *Now* pick the action with the maximum *normalised* expected choice-worthiness score.

    ↩

9. G (for example) has a Borda Scoreof 2 - 0 = 2 according to utilitarianism because that theory views two options as less choice-worthy than G, and 0 options as more choice-worthy than G.

    To fill in the final column, you take a credence-weighted average of the relevant action’s Borda Scores.

    What follows is the corresponding table based on the moral theories' preference orderings in my prior post, when empirical uncertainty *wasn't* taken into account:

	Utilitarian theory - 35%	Common sense theory - 34%	Deontological theory - 31%	Credence-Weighted Borda Score
Preference ordering	G>**R>I**	**R>I**>G	I>R>G	
Score for Guilty	2 - 0 = **2**	0 - 2 = **-2**	0 - 2 = **-2**	**-0.6**
Score for Retrial	1 - 1 = **0**	2 - 0 = **2**	1 - 1 = **0**	**0.68**
Score for Innocent	0 - 2 = **-2**	1 - 1 = **0**	2 - 0 = **2**	**-0.08**

↩

10. It's also entirely possible for paying attention to empirical uncertainty to not change any moral theory's preference orderings in a particular situation, or for some preference orderings to change without this affecting which action ends up with the highest Credence-Weighted Borda Score. This is a feature, not a bug.

    Another perk is that paying attention to both moral and empirical uncertainty also provides more clarity on what the decision-maker should think or learn more about. This will be the subject of my next post. For now, a quick example is that Julia may realise that a lot hangs on what each moral theory's preference ordering should actually be, or on how likely the crowd actually is to riot if she passes a verdict or innocent or calls for a retrial, and it may be worth postponing her decision in order to learn more about these things. ↩

11. Arguably, the additional complexity in the model is a cost in itself. But this is only a problem only in the same way this is a problem for any time one decides to model something in more detail or with more accuracy at the cost of adding complexity and computations. Sometimes it'll be worth doing so, while other times it'll be worth keeping things simpler (whether by considering only moral uncertainty, by considering only empirical uncertainty, or by considering only certain parts of one's moral/empirical uncertainties). ↩

12. The approaches discussed in this post can also deal with theories that "intrinsically care" about other things, like a decision-maker's intentions or motivations. You can simply add in a factor for "probability that, if I take X, it'd be due to motivation Y rather than motivation Z" (or something along those lines). It may often be reasonable to round this to 1 or 0, in which case these approaches didn't necessarily "add value" (though they still worked). But often we may genuinely be (empirically) uncertain about our own motivations (e.g., are we just providing high-minded rationalisations for doing something we wanted to do anyway for our own self-interest?), in which case explicitly modelling that empirical uncertainty may be useful. ↩

13. For another example, in the case of Devon choosing a meal, he may also be uncertain how many of each type of fish will be killed, the way in which they'd be killed, whether each type of fish has certain biological and behavioural features thought to indicate consciousness, whether those features do indeed indicate consciousness, whether the consciousness they indicate is morally relevant, whether creatures with consciousness like that deserve the same "moral weight" as humans or somewhat lesser weight, etc. ↩

14. For example, Devon might replace "Probability that purchasing a fish meal leads to "fish being harmed"" with ("Probability that purchasing a fish meal leads to fish being killed" * "Probability fish who were killed would be killed in a non-humane way" *

“Probability any fish killed in these ways would be conscious enough that this can count as “harming” them”). This whole term would then be in calculations used wherever ““Probability that purchasing a fish meal leads to "fish being harmed"” was originally used.

For another example, Julia might replace “Probability the crowd riots if Julia finds Smith innocent” with “Probability the crowd riots if Julia finds Smith innocent” * “Probability a riot would lead to at least one death” * “Probability that, if at least one death occurs, there’s at least one death of a bystander (rather than of one of the rioters themselves)” (as shown in this partial Guesstimate model). She can then keep in mind *this more specific final outcome, and its more clearly modelled probability*, as she tries to work out what choice-worthiness ordering each moral theory she has credence in would give to the actions she’s considering.

Note that, sometimes, it might make sense to “factor out” variables in different ways for the purposes of different moral theories’ evaluations, depending on what the moral theories under consideration “intrinsically care about”. In the case of Julia, it definitely seems to me to make sense to replace “Probability the crowd riots if Julia finds Smith innocent” with “Probability the crowd riots if Julia finds Smith innocent” * “Probability a riot would lead to at least one death”. This is because all moral theories under consideration probably care far more about potential deaths from a riot than about any other consequences of the riot. This can therefore be considered an “empirical uncertainty”, because its influence on the ultimate choice-worthiness “flows through” the same “moral outcome” (a death) for all moral theories under consideration.

However, it might only make sense to further multiply that term by “Probability that, if at least one death occurs, there’s at least one death of a bystander (rather than of one of the rioters themselves)” for the sake of the common sense theory’s evaluation of the choice-worthiness order, not for the utilitarian theory’s evaluation. This would be the case if the utilitarian theory cared not at all (or at least much less) about the distinction between the death of a rioter and the death of a bystander, while common sense does. (The Guesstimate model should help illustrate what I mean by this.) ↩

15. Additionally, the process of factoring things out in this way could by itself provide a clearer understanding of the situation at hand, and what the stakes really are for each moral theory one has credence in. (E.g., Julia may realise that passing a verdict of innocent is much less bad than she thought, as, even if a riot does occur, there’s only a fairly small chance it leads to the death of a bystander.) It also helps one realise what uncertainties are most worth thinking/learning more about (more on this in my next post). ↩