Original title: Introduction to moral uncertainty

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 me the most important concepts/takeaways from work on moral uncertainty.[1][2]
    - (My view on what the most important concepts/takeaways are is based primarily on reading Will MacAskill's 2014 thesis, which is also my main source for this post.)
2. There was no (easily findable and explicit) write-up of how to handle simultaneous moral and empirical uncertainty.
    - I assume this is because most writers on the topic consider the approach I propose to be quite obvious (at least in the case of cardinal, intertheoretically comparable theories; what this means is explained below). But it still seemed worth explicitly spelling out the approach, and noting how it can also work in cases of non-cardinal or non-comparable theories.
3. There was no (easily findable) 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". I expect that the 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 spent a large portion of the few days before writing these posts reading about moral uncertainty, but I certainly wouldn't consider myself an *expert*, and my only formal background in

---

[1] I genuinely mean no disrespect to the several relevant posts 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 to help guide readers through the concepts), and/or didn't cover (in the one post) all of what I would consider the key points about moral uncertainty.

[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.

philosophy is one undergraduate unit. Thus, I wouldn't be surprised if this sequence is inaccurate or unclear/misleading in some places (but I *would* be surprised if there were *major*, *central* mistakes).

I also expect a decent portion of readers already know much of what I cover in this sequence. But I hope the posts serve as useful introductions/overviews for many readers, and I think it's plausible some of the ideas in the second and third posts haven't been explicitly, publicly explored before.

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

Introduction

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.

---

[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 may 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. (E.g., an antirealist may still have a meaningful conception of "moral learning" in terms of gaining conceptual clarity, encountering new arguments that change what the antirealist values, simplifying their collection of intuitions into a more elegant theory, etc.)

Multiple approaches to handling moral uncertainty have been proposed. Which approach should be used depends in part on what type 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*.

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".[5]

My impression is that popular consequentialist theories are typically cardinal, while popular non-consequentialist theories are typically 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).

Comparability

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 [roughly speaking, the rightness or wrongness of an action], 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?"

Chapter 3, Section I of his thesis provides more explanation and examples relevant to this point. These are hard to summarise, but they very roughly boil down to situations where one cannot find a consistent/non-arbitrary "exchange rate" between different theories' "units of choice-worthiness" (despite all theories being cardinal). (This post's section on Variance Voting provides an example and further discussion.)

Three approaches

In his 2014 thesis, the approaches to moral uncertainty MacAskill argues for are:

[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.

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. 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.)
3. Variance Voting (VV), if all theories under consideration are cardinal but *not* intertheoretically comparable.
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 MacAskill's thesis.)

I will focus on the approaches MacAskill argues for (excluding Hybrid procedures), both because these approaches seem to me to be prominent, effective, and (relatively) intuitive, and because I know much less about any other approaches. (Examples of alternatives include the parliamentary model and a bargaining-theoretic approach.)

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, $C(T_i)$ represents the decision-maker's credence (belief) in T_i (some particular moral theory), while $CW_i(A)$ represents the "choice-worthiness" (CW) of A (an "option" or action that the decision-maker can take), according to T_i.

How MEC works may be best illustrated by an example. (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 explained in my next post).)

Suppose Devon assigns a 25% probability to T_1, 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 T_2, a different version of hedonistic utilitarianism, which values human hedons just as much as T_1 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 have a fish meal or a plant-based meal, and that he'd enjoy the fish meal about twice as much. (Finally, let's go out on a limb and assume Devon's humanity.)

According to T_1, the choice-worthiness of the fish meal 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 T_2, the choice-worthiness of the fish meal is 10 (because this theory values Devon's joy as much as T_1 does, but doesn't care about the fish's experiences). Meanwhile, the choice-worthiness of the plant-based meal 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 meal).

So the expected choice-worthiness (EC) of purchasing the fish meal is 0.25 * -90 + 0.75 * 10 = -15, and the EC of purchasing the plant-based meal is 0.25 * 5 + 0.75 * 5 = 5. Thus, Devon should purchase the plant-based meal.

This is despite Devon believing that T_2 is more likely than T_1, and T_2 positing that purchasing the fish meal is better than purchasing the plant-based meal. The reason is that there is "more at stake" for T_1 than for T_2 in this example. That is, T_2 considers there to only be a small difference (10 - 5 = 5) between the choice-worthiness of the two available options, while T_1 considers there to be a large difference (5 - -90 = 95). To me, this seems like a good, intuitive result for MEC, and shows how it improves upon the "My Favourite Theory" approach (in which the decision-maker simply does what seems best according to the theory they have the highest credence in, ignoring any uncertainty and any differences in the "stakes" for different theories).

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.[6]

The Borda Rule (BR)

The following applies somewhat to BR, and more so to VV:

- This post will describe these approaches relatively briefly, perhaps leaving some readers still somewhat confused.
- These approaches seem harder to get an intuition for than MEC.
- MEC seems to be more widely discussed than BR or VV
- I think the heuristic usefulness of an understanding of moral uncertainty for most people (rather than, e.g., professional moral philosophers, or people attempting to implement related ideas in AI) can be captured just by knowing about MEC, without also knowing about BR and VV.
- My primary motivation for describing these approaches (rather than just MEC) in this post is to set the scene for discussing (in the next post) how to modify these approaches to allow one to also explicitly account for empirical uncertainty (instead of just moral uncertainty).

I leave it to readers to decide whether to skip these sections entirely, read them, or read Chapters 2 and 3 of MacAskill's thesis instead/as well.

Now, for those noble souls who stuck around: In his thesis, MacAskill recommends using BR when all moral theories under consideration 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). I will first quote MacAskill's formal explanation of BR (which may be somewhat

---

[6] 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.

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 T_i, is equal to the number of options within the option-set that are less choice-worthy than *A* according to theory T_i's choice-worthiness function, minus the number of options within the option-set that are more choice-worthy than *A* according to T_i's choice-worthiness function.[7]
>
> An option *A*'s *Credence-Weighted Borda Score* is the sum, for all theories T_i, of the Borda Score of *A* according to theory T_i multiplied by the credence that the decision-maker has in theory T_i.
>
> [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]."

So, according to the variant of utilitarianism, the options' Borda Scores are as follows:

> G: 2 - 0 = 2 (this is because, according to this theory, there are two options that are less choice-worthy than G, and 0 options that are more choice-worthy than G)

[7] 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).

R: 1 - 1 = 0
I: 0 - 2 = -2

Whereas, according to the variant of common sense, the Borda Scores are:

G: 0 - 2 = -2
R: 2 - 0 = 2
I: 1 - 1 = 0

Finally, according to the deontological theory, the Borda Scores are:

G: 0 - 2 = -2
R: 1 - 1 = 0
I: 2 - 0 = 2

The option's Credence-Weighted Borda Scores are therefore:

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.

Variance Voting (VV)

In his thesis, MacAskill recommends using VV when the moral theories under consideration *are* cardinally measurable, but *aren't* intertheoretically comparable. The basic principle this approach 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). MacAskill 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 provides two arguments that VV is the approach that satisfies the principle of equal say.

In explaining what VV actually is, he writes that it:

> "treats 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."

My understanding is that, basically, VV involves:

- First, normalising the choice-worthiness function of each theory at its variance. (This means changing *how much* more choice-worthy options are, relative to each other, according to each moral theory, and doing this based on the variance of choice-worthiness according to each theory. This is meant to ensure that no theory is given "too much say".)
- Second, applying MEC, using these normalised choice-worthiness functions.[8]

[8] Please let me know if you think it'd be worth me investing more time, and/or adding more words, to try to make this section - or the prior one - clearer. (Also, just let me know about any other feedback you might have!)

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.[9]

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.[10][11]

---

[9] 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.)

[10] 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.

[11] 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.