# Normative uncertainty

Paul Christiano Follow
Oct 1, 2015 · 6 min read

Suppose that I am an AI trying to satisfy a user's preferences, and I'm uncertain about what those preferences are. I face the question: how do I take a probability distribution over preferences and turn it into a single preference relation?

This post suggests a simple answer to that question, which seems likely to be good enough for many applications. I'm mostly optimistic not because I think this proposal is really great, but because I think that it can avoid doing anything really wacky or surprising—I'm optimistic that the practical demands on a system for handling normative uncertainty probably won't go beyond that.

Roughly speaking, the proposal is to use a given preference relation to evaluate the goodness of an outcome in terms of the value of a marginal dollar, and then to use the expected $ valuations to decide which outcomes are best. We could also use hours or some other flexible resources as our common unit.

## Preliminaries

For the rest of the post I'll assume that each preference relation can be represented by a utility function U.

One natural approach is to simply take the aggregate utility function E[U], where the expectation is taken over our uncertainty about U itself.

But before even thinking about whether this is the right thing to do, we should notice that this approach is underdetermined: U and xU represent the same preferences for any x > 0, and there is no obvious canonical way to translate preferences into a real-valued function that could be aggregated in this way.

On the flip side, it's easy to argue that any sensible aggregate preferences should be representable by *some* linear combination of these representative utility functions U. So our entire question is reduced to one of finding adequate weights, or of picking common units for different possible preferences.

Much of the analysis in this post will apply in principle to the case of moral uncertainty as well, though I will be thinking about the value learning case when evaluating possible solutions and don't want to make a strong claim about the moral uncertainty case.

# The unit of caring

Suppose that I'm considering two possible sets of preferences, A and B. In order to combine them, I'll imagine three possible experiments:

1. Life continues as normal.
2. I receive $1 and immediately (and frictionlessly) spend it in the way recommended by A.
3. I receive $1 and immediately (and frictionlessly) spend it in the way recommended by B.

Presumably A prefers case 2 to case 1. How much more? I think it is a reasonable convention to say that A prefers case 2 by $1.

We can represent A by a utility function $U^a$ that assigns $0 to case 1 and $1 to case 2. Similarly, we can represent B by a utility function $U^b$ that assigns $0 to case 1 and $1 to case 3.

Under this normalization, I think it is reasonable to use the aggregate preferences $p^aU^a + p^bU^b$, where the p's are the respective probabilities of A and B.

## Why?

The most obvious virtue of the system is that it leads to sensible betting behavior: an agent using it will be willing to make bets about the user's preferences at odds given by that agent's credences.

This property seems to be a very intuitive notion of "fairness," and a failure of this property could be quite surprising. For example, it would seem like a bug if our system inferred that the user had a 99% chance of having preferences X, but nevertheless made a 99 : 1 bet *against* the user having preferences X.

This property uniquely pins down the weighting given above, modulo details about what currency the bets are in and when they pay out (which will be discussed in the next section—I don't think dollars are the right answer).

Of course the behavior of the agent on actual bets is not very relevant, but analogous failures would occur in more realistic situations. For example, suppose that the agent is going to make a guess about what the user wants, and if the agent guesses wrong the user will have to spend 10 seconds correcting the agent. If the agent is willing to make bad bets about the user's preferences, then it will also take actions that systematically waste the user's time in expectation.

Alternative weightings would justify this weird behavior by reasoning like "yes, this will probably waste the user's time, but the user's time matters more if they care about A than if they care about B." For the most part I think that we don't want their tools to make inferences like this.

There are cases where similar reasoning is appropriate. For example, it seems correct for an agent to be especially careful to avoid bothering the user in the case where they are currently especially short on time, even if doing so would save the user time in expectation (so e.g. an agent may conservatively assume that the user is busy even if they think they probably are not). But this kind of reasoning *would* be incorporated into our system, since the value of time can fluctuate over time, and our benchmark is tied to the moment when our thought experiment occurs (see the next section).

# Details

## Currency

The key fact about dollars is that they are very flexible, and can be converted to most other resources we could consider; any other flexible resource would work just as well and I don't think that dollars are the very best option. This scheme is only potentially reasonable if it is not too sensitive to the choice of units.

A better candidate may be the user's time. For example, we can consider experiments in which the user gets a free hour of work to spend pursuing either A or B.

Better yet is a basket is of different resources, whose value will tend to be more stable—because we are considering ratios [value of outcome] / [value of resource], the mixture of two equally stable units (with the same sign) will tend to be more stable than either one alone.

## Exchange rates

This proposal is most likely to behave strangely if the exchange rates between different resources vary in a way that depends on the user's unobserved preferences. In general I think that these quantities are likely to be relatively robust, since the AI can observe the exchange rates at which the user appears to trade off these resources (and so their uncertainty will be concentrated on possible preferences that have roughly those exchange rates).

This is a consideration in favor of using resources that the user behaves sensibly about, that the AI can reason about, and that the AI can observe the user's choices about.

## Timing

Whatever resource we use, if we want to be consistent over time we should fix a particular benchmark moment when the resource is to be received. It seems safest to choose a time *before* the agent begins to operate—doing otherwise can lead to somewhat surprising behavior, though I'm not totally clear on whether that behavior is problematic.

## Size

Our definition considers small windfalls of resources like $1 or 1 hour. This is intended to keep the counterfactuals close to the real world so that they can be reasonably informative about the user's preferences in the real world. We could also consider much larger windfalls. For example, one standard proposal is to normalize (best case – expected case), which essentially corresponds to normalizing the value of an infinite windfall. I don't find these proposals very persuasive precisely because they depend on counterfactuals so distant from reality.

Note that the windfall may be traded if 10x more resources would be more than 10x better. For example, if the agent gets 1 hour to spend optimally pursuing values A, the best thing to do may be to make a bargain the rest of the agent where it spends 10 hours with probability 10% pursuing A.

## Planning

The value of a dollar depends on what I do with it. I'm imagining that we choose a plan with a reasonable investment of effort, but that the cost of investing that effort is not included in the goodness of the counterfactual—we imagine that it is invested in some parallel world. The chosen plan may be to set aside the money in order to use it later in the service of the chosen goal.