Week 11 Summary: Causal, evidential and functional decision theory.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

May 18, 2019

Contents

1 Meta

This week I'm trying for a more reflective and personal rather than descriptive summary. As such, I'm writing what I think and how I interpret things rather than summarizing. I'm referring to the text less as an exercise. I quite possibly will get things wrong.

2 Causal Decision Theory (CDT)

See acts and outcomes as causally independent always. Do **not** acknowledge that the DM's estimated probability distribution does not reflect the objective, true probability distribution.

The problem here is obvious - this assumption that your probability should not be influenced by what you *would* do in some cases is nonsense.

3 Evidential Decision Theory (EDT)

See acts and outcomes as causally dependent (at least sometimes). This is due essentially to incomplete information. See this in a Bayesian way - the very choice of an act may reveal more information about the probability of an outcome (e.g. the psychopath example).

Distinguishing itself from CDT, this does indeed acknowledge that the probability distribution in the initial decision model is the subjective one, not the underlying objective one.

Peterson mentioned that this comes in many flavors, but only went into detail about the Bayesian one. Just based off the name, I don't know what others there could be.

3.1 Problems

In my opinion, none of these problems are very strong. This seems to be the most rational approach.

The crux of the issues lies with the denominator here:

$$P(X|Y) = \frac{P(X \wedge Y)}{P(Y)}$$

Where X is the outcome and Y is the act, this implies that you know your own probability of performing an act $P(Y)$ (in the denominator).

I don't know about this. It's not strictly necessary to have the denominator (as we've seen in Bayesian inferencing). We always know the denominator is between 0 and 1, so we just add multiply the learning rate by some constant. But this is not a strong counter-argument - there's the need to choose the constant, and how on earth do you do that?

I'm dissatisfied with Peterson's counterarguments to this problem. They seem to imply that it is indeed not possible for an agent to estimate their own probability distribution.

Maybe a stronger counter-argument is that of course we can model our own choices probabilistically, as we very clearly have models of ourselves in our own heads. We do. This is the basis of self-reflection and consciousness in the first place. It may or may not be an accurate model of ourselves, but it is odd to deny that we have one. Thus, whatever argument gets flung against having models of ourselves, it is simply descriptively true that we *do* have these self-models.

Another possible line of argument is that probability is an overloaded term (use the ecumenical approach from last week's philosophy of probability).

Also, one of the largest issues is simply that we will infinitely recurse upon and update our subjective probabilities in order to maximize our expected utility. This seems strange to me. This converges upon a final model of ourselves that decides with probability 1 to perform the action with the maximal expected utility. This means you've got a model of yourself as the rational self. It's rational. What's the issue? I feel like I'm misunderstanding something here, but I don't know what it is.

4 Functional Decision Theory (FDT)

I'm having a lot of trouble understanding this. As far as I can tell, it's useful for scenarios where the outcome depends on a skilled predictor's prediction of the decision maker's decision.

Yes, that's quite a head-warpey timey-wimey approach, so let me try to deconstruct the character of it.

1. Assume the predictor will always predict correctly (for simplicity)
2. Assume that the prediction is set in stone

At this stage, you have the ability to choose. While the predictor has no ability to change the past, the predictor had (in my simplification) perfect knowledge of your future choice at the time of its prediction.

Thus you choose to behave in the way the predictor said would have led to maximum utility. E.g. you one-box. Even in the transparent Newcomb version, you *always* one-box.

4.1 (My) Analysis

4.1.1 Rewarding Irrationality

In my opinion, this is irrational. As Yudkowsky mentions, there's a criticism that these types of decision problems (the paradoxes FDT resolves better) reward irrationality. I agree with that criticism. Which in turn might mean that this better performance suggests that FDT is an irrational decision theory.

4.1.2 Miscellaneous Thoughts

This sort of metacognition-esque method makes me think about those attempts by an AI to model a human's model of the AI itself. This sort of thing is useful for making te AI give the human more predictable output (easier for the AI to do if it's trying to understand what you would expect). The way the human thinks the AI will perform becomes a sort of implicit control mechanism. The difference here is that the control mechanism is explicit - you one-box because Newcomb's predictor explicitly lets you know that if you one-box, you will be more likely to get the huge reward.

So the use for this might simply be in controlling agents. It doesn't have to be about rational decision making.

Another way of putting this is that it ensures trust, which is probably useful for protracted games.

5 Some Paradoxes

5.1 Newcomb's Box (predictor)

- 2 boxes, choice to none, both, box 1 or box 2
- Box 1 always has $1000 in it
- Box 2 has either $1M or $0

- Reliable predictor has already predicted a 99% chance that, should you pick ONLY box 2, box 2 contains $1M

Should you take box 2, or both boxes? Obviously, the predictor cannot affect the boxes anymore.

5.2 Psychopath Killer Button

- Belief that anyone who would press the button is a psychopath
- Press the button, killing all psychopaths. You get +ve reward
- If you die, you get more -ve reward than the magnitude of +ve reward from killing all psychopaths

What do you do? Obviously, based on your own belief about pressing the button implying psychopathy, you should not do it

5.3 Death in Damascus (`no stable acts`)

- You meet Death accidentally
- He is surprised to see you and mentions he will visit you tomorrow
- You want to escape death and have the option of moving cities. Do you stay or move?

It is said there is no stable act here, simply because whatever you choose, death will have predicted it and thus you should have done the opposite.

There's clearly no rational act here. Seriously. Damned if you do, damned if you don't. I agree with Peterson's stated criticism - this lies outside the bounds of rationality.

5.4 Psychological Twin Prisoner's Dilemma

- Prisoner's Dilemma, but, of course, the other person is your psychological twin
- Clearly, it is prudent to assume the other person does what you will do

This breaks CDT as it assumes act-state independence. The challenge is this: how do you get the agents to cooperate (and therefore maximize reward) in a strategy that is rational for all games, not just the prisoner's dilemma?