Indirect decision theory

Paul Christiano Follow
Mar 2, 2015 · 3 min read

In which I argue that understanding decision theory can be delegated to AI.

Indirect normativity.

My preferences can probably described by a utility function U : [Possible worlds] → $\mathbb{R}$. But U is likely to lack a simple specification, and even if it has one, I certainly don't know it. So if I wanted to describe my preferences, I might define the utility of a world w as:

> *U*(w) = *"How good I would judge the world* w *to be, after an idealized process of reflection."*

Intuitively, if there were a powerful AI around, I'd like it perform an action a such that $\mathbb{E}[U(w)|do(a)]$ is as large as possible.

Indirect decision theory

But what does $\mathbb{E}[U(w)|do(a)]$ mean anyway? We haven't given a prescription for interpreting "|do(a)," and we haven't specified a distribution over possible worlds.

Really, I'd like to leave these questions up to an AI. That is, whatever work I would do in order to answer these questions, an AI should be able to do just as well or better. And it should behave sensibly in the interim, just like I would.

To this end, consider the definition of a map U' : [Possible actions] → $\mathbb{R}$:

> *U'*(a) = *"How good I would judge the action* a *to be, after an idealized process of reflection."*

Now we'd just like to build an "agent" that takes the action a maximizing $\mathbb{E}[U'(a)]$. Rather than defining our decision theory or our beliefs right now, we will instead come up with some answer during the "idealized process of reflection." And as long as an AI is uncertain

about what we'd come up with, it will behave sensibly in light of its uncertainty.

This feels like a bit of a cheat. But I think the feeling is an illusion. More precisely:

A successful AI will need to be able to reason about quantities like $\mathbb{E}[U(w)|do(a)]$, and we can't dodge this algorithmic problem with a sleight of hand. But a sleight of hand might dodge the philosophical hazard of committing ourselves to a particular definition of $\mathbb{E}[U(w)|do(a)]$.

U' doesn't seem any harder to define than U. Indeed it may be easier—possible worlds are complex and massive objects, and to evaluate them we might have to think long and hard and become very different people than we are today. But actions are close to home.

And U' seems every bit as actionable as U: if a program can calculate $\mathbb{E}[U(w)|do(a)]$ (whatever we mean by that), it can probably just as well calculate $\mathbb{E}[U'(a)]$.

It may be that this approach isn't tenable. But I think that is necessarily a question about the internal structure of an AI.

Possible problems

Is "idealized reflection" up to it?

In order to evaluate how good an action is, we will often want to understand its consequences. This puts an additional requirement "idealized process of reflection:" it needs to be powerful enough to understand the consequences of each possible action.

I don't think this is a big deal:

1. In order for U' to guide an AI's decisions, U' just needs to be as wise as the AI itself. It doesn't matter if we would like an action because of some hard-to-anticipate consequences, unless the AI can anticipate that we'll like it.
2. The bar for an "idealized process of reflection" into whose hands we would entrust the entire future seems much *higher* than the bar for a process of reflection that can determine the consequences of actions today.

A final wrinkle

Computing $\mathbb{E}[U'(a)]$ doesn't seem any more or less complicated than computing $\mathbb{E}[U(w)|do(a)]$.

But, it seems unlikely that superintelligent AI systems will simply compute $\mathbb{E}[U(w)|do(a)]$ for each possible action a and then do the best one. For example, they may have to think about how to think; more broadly, it seems hard to predict what a successful AI system of the future will look like.

It may well be that the internal structure of AI systems favors rational agents over other designs. For example, "maximize $\mathbb{E}[U(w)]$" might be a really useful invariant to organize a system around, and if so it's not clear whether maximizing $\mathbb{E}[U'(a)]$ is a satisfactory alternative. (I discuss this issue inconclusively here.) It's plausible that an understanding of decision theory will help us see how the global goal-directed behavior of a system emerges from a combination of heuristics and goal-directed components; but for now we don't have a very clear picture.

Conclusion

I think this final wrinkle gives us our best reason to study decision theory today. But I think the case is weaker and more subtle than is often assumed, and I am certainly not yet convinced that we can't delegate decision theory to an AI.