An introduction to decision theory for AI safety

Richard Ngo

Overview: two approaches to AI safety

OpenAI, DeepMind, Centre for Human-Compatible AI: safe machine learning

- Reinforcement learning and inverse reinforcement learning
- Safe exploration and scalable oversight
- Adversarial training
- Reward engineering

Machine Intelligence Research Institute (MIRI): “agent foundations”

- Naturalised induction
- Logical and Vingean uncertainty
- Ontology identification and value specification
- Decision theory

Basic framework in decision theory

“There are two types of open problem in AI. One is figuring how to solve in practice problems that we know how to solve in principle. The other is figuring out how to solve in principle problems that we don’t even know how to brute force yet.”

An agent face choices between various options.

Each option is associated with a probability distribution over outcomes.

Agents have utility functions which assign each possible outcome a number.

We ask what “rational” agents who want to maximise expected utility would do.

Descriptions of decision problems are certain and contain all relevant information.

Two standard positions

Causal decision theory (CDT)

- Most widely-accepted position
- Take the action which causally leads to the best consequences

Evidential decision theory (EDT)

- Another mainstream position
- Take the action which you would be happiest to learn that you had taken
- Equivalently: the utility of an outcome is your expected utility when you (Bayesian) conditionalise on taking that action

Illustrating the difference: Newcomb’s problem

A superintelligent, perfectly honest entity Omega offers you two boxes.

Box A is open, and has $1000 inside. Box B is closed.

You can either take both boxes, or just box B.

If Omega predicted you would take just box B, it already put $1000000 in box B.

If it predicted you would take both boxes, then it left box B empty.

Assume that Omega is a perfect predictor, and you only want to maximise the expected amount of money that you get.

What do you do?

Objections and responses

Objection: There's no such thing as a perfect predictor.
Response: CDT and EDT's responses differ even if the predictor is imperfect.

Objection: Newcomb's problem is just biased against rational people.
Response: Omega only decides based on what you will do, not how you think.

Objection: CDT would one-box if given the option to precommit.
Response: Not if Omega predicts before the CDT agent can precommit

Objection: These fantastical cases are totally irrelevant to how people should actually reason in the real world.
Response: See later examples

Functional decision theory

Act as if you determine the output of all implementations of your decision function.

On Newcomb's problem: your action "determines" what Omega predicted, because to predict you Omega needs to implement your decision function.

Unlike CDT/EDT, FDT agents wouldn't self-modify/precommit if possible.

Unlike CDT, FDT agents don't "regret being rational".

Unlike CDT, FDT agents don't get confused by their own predictions.

Unlike EDT, FDT agents wouldn't pay to avoid learning true information.

See Yudkowsky and Soares, 2017, https://arxiv.org/pdf/1710.05060.pdf

Slight variations on Newcomb’s problem

Iterated Newcomb’s problem

- You are locked in a room for 100 days, with 100 pairs of boxes; you can open 1 pair a day. Each Box A contains a sandwich; each Box B may contain a full meal. If you don’t ever get a full meal, you will starve with 90% probability.

Twin prisoner’s dilemma

- You are playing the Prisoner’s dilemma against a perfect clone of yourself.
- Outcomes, from best to worst: (D,C), (C,C), (D,D), (C,D)

More relevant puzzles

Relevance for humans in social situations: Parfit's hitchhiker

- You are stranded in the desert with no money. A driver stops and offers you a ride if you promise to pay them $10000 when you get to town. The driver is very good at reading body language, and will probably (90%) know if you lie.
- You know that once you get to town, you will have no reason to pay them.
- You promise to pay. They believe you. Once you get into town, do you pay?

Relevance for artificial agents

- You are an AI, and you are being predicted by someone who has a copy of your source code, so can model you with very high accuracy.

Failure of EDT: XOR Blackmail

An accurate predictor with a reputation for honesty sends you a message:

“Either you will soon discover you have cancer, xor you will soon pay me $1000.”

EDT reasoning: P(Cancer | Pay) is close to 0. But P(Cancer | ~Pay) is close to 1.

So EDT pays them! But surely this is irrational: they can’t give you cancer!

What do FDT and CDT do?

Failure of CDT: Death in Damascus

Death is a perfectly accurate predictor, and is planning to meet you tomorrow. Death has already written where he will be tomorrow in his appointment book.

You are in Damascus, but you can pay $1000 to flee to Aleppo.

CDT gets very confused. As soon as a CDT agent makes a decision, they realise that Death has predicted that decision, and so they change their mind.

If someone offers to sell you for $1 a coin that cannot be predicted by Death, CDT refuses because it can’t model the logical connection between you and Death.

What do FDT and EDT do?

What is going on?

Diagnosis: Basically, the difference is treatment of counterfactuals.

EDT: all statistical counterfactuals. “If I hadn’t paid, I would have gotten cancer.”

CDT: only causal counterfactuals. “Whether or not I have cancer doesn’t depend on whether or not I pay.” “If I had picked Aleppo, I would have lived.”

FDT: also logical counterfactuals. “If I hadn’t paid, they wouldn’t have blackmailed me in the first place.”
“If I had picked Aleppo, DEATH would have already predicted that.”

One helpful intuition: your decision function should be trying to “make bad outcomes inconsistent”, so that you never end up in that situation in the first place.

And yet...

We have no satisfactory account of either causal or logical counterfactuals!

“If my action causes the good outcomes” - how can this be well-defined?

“If my decision function output a different answer” - but how can we magically insert this change into the world? Relevantly similar to “What if 1+1 = 3”?

What if no satisfactory account exists? Do we just use EDT plus precommitment to not pay anyone who “seems suspicious”? But if you learn that the gene for altruism and the gene for heart disease are statistically correlated, then EDT says you should act selfishly. So the problem occurs outside the “blackmail” framing.

If we want a general theory of how EDT agents should precommit, then we’re back to the problem of finding an account of counterfactuals.

Free will?

A common intuition: it feels like these problems arise from assumption of free will, and the assumptions of perfect predictors undermine that.

But they still work without perfect predictors! As long as you can't "outthink" the predictor by figuring out what they predicted.

This is a special case of Vingean uncertainty - uncertainty about the behaviour of agents which are more intelligent than you.

If we think about them purely in terms of deterministic algorithms, then maybe that's less confusing.

Thanks for listening.