Index of some decision theory posts

Tsvi Benson-Tilsen

1 What this is

An index of posts outlining some ideas in decision theory. I plan to have them available both on the forum and on github as pdfs. I might turn this into a general index for agent-foundations-related decision theory research depending on interest and my time.

2 Index

2.1 Index of some decision theory posts

As advertised.

Forum post: `https://agentfoundations.org/item?id=1026`

Github pdf: `https://github.com/tsvibt/public-pdfs/blob/master/decision-theory/index/main.pdf`

2.2 Notation for induction and decision theory

A reference for notation that might be useful for using (universal) Garrabrant inductors as models for bounded reasoning, and some notation for modelling agents. (Not on the forum because tables.)

Github pdf: `https://github.com/tsvibt/public-pdfs/blob/master/decision-theory/notation/main.pdf`

2.3 Desiderata for decision theory

A list of desiderata for a theory of optimal decision-making for bounded rational agents in general environments.

Forum post: `https://agentfoundations.org/item?id=1053`

Github pdf: `https://github.com/tsvibt/public-pdfs/blob/master/decision-theory/desiderata/main.pdf`

2.4 An inductive setting for decision theory

A discussion of the appropriate setting for studying decision theory.

Forum post: todo

Github pdf: todo

2.5 Training a universal Garrabrant inductor to predict counterfactuals

A proposal for training UGIs to predict action- and policy-counterfactuals by learning from the consequences of actions taken by similar ("logically previous") agents.

Forum post: `https://agentfoundations.org/item?id=1054`

Github pdf: `https://github.com/tsvibt/public-pdfs/blob/master/decision-theory/training-counterfactuals/main.pdf`

2.6 Open problem: very thin logical priors

An open problem relevant to decision theory and to understanding bounded reasoning: is there a very easily computable prior over logical facts that, when updated on the results of computations, performs well in some sense?

Forum post: todo

Github pdf: todo