Chapter 1

Examples

1.1 Wrong Variables.

Suppose worlds are parameterized by three variables A, B, C, each of which can take on two values: variable A can take on either a or $\bar{a}$, B can take on b or $\bar{b}$, and C is either c or $\bar{c}$.

Suppose further that the original agent cannot observe these variables directly, but rather observes variables X and Y, which are logically defined as

$$\begin{aligned} X &:= A \wedge B \\ Y &:= B \vee C \end{aligned}$$

1.2 New Constraint

Chapter 2

Unification

2.1 The Emotion / Reason Dichotomy

Morals, preferences, goals, utilities, rewards/punishments, and desires all have something in common: the behavior that characterizes them is an optimization. They answer a question about *why* something was done, in a way that is compatible with planning and rational, well-thought-out behavior. If one sees a person repeatedly doing something, say walking their dog, it is reasonable to conclude that either they hold that this is a morally good thing to do, have a preference or goal / sub-goal of walking their dog, enjoy it, or have times when they want to do it. Moreover, these are considered explanations of *why* behavior happens. They are descriptions of the things that agents optimize.

The second feature they share is a subjectivity: anyone can have any preferences or utilities or rewards (perhaps subject to certain constraints if you don't want to be manipulated, want to behave robustly in the presence of adversaries, and so on). Once the space has been constrained, having different preferences is merely... a matter of preference. The theories we use for modeling agents do not take into account the mechanisms by which one might obtain preferences, which

More egregiously still, standard utility / reward maximization picture has nothing to say about the possibility of these quantities changing over time.

All of this is to be contrasted with two other concepts:

1. Empirical analysis which determines that some behavior (as opposed to another one) is occurring.
2. Theories of belief, rationality, and *how* to optimize.

All of these have been postulated as reasons why people do things, and more generally, as theories about ways to shape behavior of agents.

The general idea is to consider two descriptions of motivation equivalent if they necessarily result in indistinguishable behavior.

2.1.1 Utilities and Preferences

To simplify things, economists and decision theorists start with the case when the set of possibilities is small and finite. If O is the set of things one could choose between, i.e., the set of outcomes, then we can formalize a preference as a pre-order $(O, \preccurlyeq)$ on O.

$$\forall x \in O.\ x \preccurlyeq x \qquad \text{(Reflexivity)}$$
$$\forall x, y, z \in O.\ (x \preccurlyeq y)\ \wedge\ (y \preccurlyeq z) \Rightarrow\ (x \preccurlyeq z) \qquad \text{(Transitivity)}$$

Rather than dealing with the partial order directly, we might like to have an embedding into something we have more intuition for, such as natural numbers or a continuous space. The problem with this, of course, is that the space may have some structure which is incompatible with the partial order.

of these results in a ranking,

However, there is often a lot more structure on outcomes

2.1.2 Utilities and Rewards

Both utilities and rewards are real-valued functions from something in the world to a one-dimensional notion of 'good-ness' represented by $\mathbb{R}$, and hence are sometimes thought of as equivalent; here we will do some of the work to explore in what sense they might be equivalent, and the sorts of issues that might come up if conflating the two without any thought.

Utilities are over outcomes, so a utility function $U : Z \to \mathbb{R}$ must be

$$\pi^*(x) = \operatorname*{arg\,max}_{y:Y} \mathop{\mathbb{E}}_{z:Z} \left[U(z) \;\middle|\; y, x \right]$$

Determinism

If $\mathcal{W}$ is the set of all possible world states, and $\tau : \mathcal{W} \to \mathcal{W}$ is a deterministic function describing the evolution of the world, then having a preference over futures starting at w_0 and consistent with τ is meaningless, as there is only one such sequence of worlds. Therefore, pure determinism only makes sense with agents with imperfect information.

Let X be the space of world representations for an agent, and let $\eta : \mathcal{W} \to X$ be some function, thought of as perception, which generates a world representation. This gives us an equivalence class $[w] := \eta^{-1}(\eta(w))$ of information sets of the agent, which may not be stable under τ, and so it may be the case that $w \underset{\eta}{\sim} w'$ but not $\tau(w) \underset{\eta}{\sim} \tau(w')$

The fact that one gets more information over time suggests that the optimal policy, even with infinite computation, in general will change with additional samples. After two steps, the policy becomes

$$\begin{aligned}
\pi^*(\eta \circ \tau(x)) &= \operatorname*{arg\,max}_{y:Y} \mathop{\mathbb{E}}_{z:Z} \left[U(z) \;\middle|\; y, \eta\tau x \right] \\
&= \operatorname*{arg\,max}_{y:Y} \int\limits_{\mathrm{d}Z} \Pr\left[z \;\middle|\; \left(Y^{(t)} = y\right) \wedge \left(X^{(t)} = \eta \circ \tau(x)\right) \wedge \left(X^{(t-1)} = x\right) \right] U(z)
\end{aligned}$$

Nondeterminism

Once again, suppose $\mathcal{W}$ is the set of possible worlds, with now $\tau : \mathcal{W} \to 2^{\mathcal{W}}$, a non-deterministic version of the transition function in the deterministic setting from before.

In this more general case, we can still relate the two. To begin, suppose that we have a complete preferences over possible futures.

2.2 Fields in which a related problem has been addressed

2.2.1 Art History

2.2.2 Pedagogy

2.2.3

2.3 Applications

2.3.1 Content Recommendation

The way that these assumptions of static preferences manifest themselves in content recommendation systems is the

2.3.2 AI Safety

2.3.3 Better Inverse Reinforcement Learning

2.3.4 Robotics: Life-long Learning

2.3.5 Meta Ethics

2.4 As a Learning Problem