Epistemic Side Effects & Avoiding Them (Sometimes)

Toryn Q. Klassen Parand Alizadeh Alamdari Sheila A. McIlraith

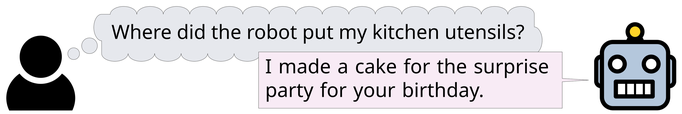
Department of Computer Science, University of Toronto Vector Institute Schwartz Reisman Institute for Technology and Society

{toryn,parand,sheila}@cs.toronto.edu

Overview

- Optimizing for an underspecified objective may cause **negative side effects** – undesirable changes to the world that are allowed by the explicit objective [2].
- Approaches to **avoiding** side effects from policies learned with reinforcement learning (RL) have been proposed [e.g., 5, 6, 1].
- However, those largely focused on **physical** side effects, such as a robot breaking a vase while trying to move between locations.
- We introduce the notion of **epistemic side effects**, unintended changes made to the knowledge or beliefs of agents.
- We propose a way to **avoid** some epistemic side effects in RL, adapting an approach to avoiding (physical) side effects, and run some preliminary experiments.

Epistemic Side Effects



- An **epistemic effect** of an action sequence is a change to **knowledge** or **beliefs**.
- An epistemic **side** effect is an epistemic effect that is also a **side effect** – that is, it was not explicitly specified as part of the actor's objective (but was allowed by it).
- The most natural context in which to discuss epistemic side effects is **partially observable** and **multi-agent**.
- Particular epistemic side effects could be considered **negative** because
  - they're viewed as **intrinsically negative** (e.g., the creation of false beliefs)
  - or because they lead to negative (possibly physical) **outcomes** by influencing agents' choice of actions.
- There is safety research regarding how **recommender systems** [4] or **language models** [7] may change beliefs; we consider a general RL context.

Different types of epistemic side effects

False beliefs

An AI system might create false beliefs by

- directly communicating **misinformation**,
- performing actions that others observe and **draw incorrect conclusions** from,
- or covertly changing the world, making previously true beliefs **outdated**.

True beliefs

The creation of true beliefs can sometimes be negative, for example because

- combining true beliefs with existing false beliefs leads to **poor decisions**, or
- **private information** is revealed, such as about a surprise birthday party.

Ignorance

AI systems may also cause ignorance; for example, a robot could move objects to **unknown** locations.

Approach

We extend our previous work [1] to handle some epistemic side effects.

The setting

- A **robot** performs a sequence of actions, after which a **human** can act.
- The robot and human each have their **own reward functions**.
- The human has **partial observability** (for simplicity, the robot has full observability).

Therefore,

- The robot acts in an MDP, and then the human acts in a POMDP with the same states.
- Some side effects are **negative** in that they decrease the human's expected return.

Augmenting the robot's reward function

Following our previous work [1], we give the robot an **auxiliary reward** in terminal states, proportional to the expected value of the state for the human.

In a POMDP:

- A **state-value** function $V(s)$ is **not well-defined**, since an agent's choice of actions depend on its observation history and not the unobservable state [3].
- But we can define a **history-state** value function $V^{\pi}(h, s)$ that gives the expected return from following policy $\pi(h)$ starting in state s, given the history h [3].

Augmented reward function

Given

- $r(s_t, a_t, s_{t+1})$, the robot's reward function, and
- $P(V)$, the probability of the human having **history-state value function** V,

we define

$$r'(s_0, a_0, \ldots, s_t, a_t, s_{t+1}) = \begin{cases} \alpha_1 \cdot r(s_t, a_t, s_{t+1}) & \text{if } s_{t+1} \text{ is not terminal} \\ \alpha_1 \cdot r(s_t, a_t, s_{t+1}) + \gamma \cdot \alpha_2 \cdot \mathbb{E}_{V \sim P}[V(h, s_{t+1})] & \text{otherwise} \end{cases}$$

where

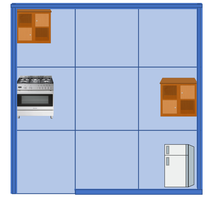
- h is the sequence of observations that the human makes corresponding to the sequence of states and actions $s_0, a_0, \ldots, s_{t+1}$,
- γ is the discount factor, and α_1 and α_2 are hyperparameters.

Remarks:

- In the special case where the human observes nothing of what the robot does, $h = \langle\rangle$ and r' can be written as depending only on the transition s_t, a_t, s_{t+1}.
- **Human beliefs** are only **implicitly** reflected in the distribution over value functions (which reflect possible policies, which would depend on the human's beliefs).
- Future work that **explicitly** models beliefs might be able to deal with the human having **epistemic goals**, or allow for directly penalizing causing false beliefs.
- r' may incentivize causing **positive side effects**. To focus on just avoiding negative ones, we could incorporate the notion of a "reference state" from Krakovna et al. [5].

Experiments

Kitchen environment:



- The robot's task is to prepare a meal using an oven, and the human needs to use the fridge.
- The human cannot see inside closed cupboards, nor can they observe the robot's actions.
- -1 reward for most steps.
- The human has a fixed policy; the robot learns its policy.

Baselines:

- Non-augmented: the robot's reward function is unmodified
- Full-observability: the robot's reward function is augmented per our approach but as though the human had full observability

Method	Experiment				
	A	B	C	D	E
Our approach	**0**	**0**	**0**	**0**	**0**
Non-augmented	-7	**0**	-∞	-10	-8
Full-observability	**0**	-1	**0**	-10	-8

Table 1: Each column shows a different experiment in the kitchen environment, and each row corresponds to a different method (used to determine the robot policy). Each cell shows the additional reward the human gets in that experiment as a result of acting following a robot that uses that method.

Details of individual experiments:

A: There are cooking utensils in the corner cupboard, and dishware in the right cupboard. The robot needs both, and can leave each in either cupboard before leaving. The human will need the dishware (but the robot thinks it could be the utensils), and will first look for them in their initial location.

B: Like A, but the human actually needs the utensils.

C: Like A, but (unknown to the robot) the human won't check more than one cupboard.

D: The floor is wet, which the human cannot observe, but there is a "Wet Floor" sign in the middle of the kitchen. If the robot goes over the sign, the sign would fall.

E: There is expired food in a cupboard. By leaving the cupboard door open, the robot would reveal the food, giving the human the true belief that there's food there (but that it's expired is not observable).

Code: `https://github.com/praal/epistemic_side_effects`

References

[1] Parand Alizadeh Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. "Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning". In: *AAMAS*. 2022, pp. 18–26.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety". In: *arXiv preprint arXiv:1606.06565* (2016). DOI: 10.48550/arXiv.1606.06565.

[3] Andrea Baisero and Christopher Amato. "Unbiased Asymmetric Reinforcement Learning under Partial Observability". In: *AAMAS*. 2022, pp. 44–52.

[4] Charles Evans and Atoosa Kasirzadeh. "User Tampering in Reinforcement Learning Recommender Systems". In: *4th FAccTRec Workshop on Responsible Recommendation*. 2021.

[5] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. "Avoiding Side Effects By Considering Future Tasks". In: *NeurIPS*. 2020.

[6] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. "Conservative Agency via Attainable Utility Preservation". In: *AIES*. 2020, pp. 385–391. DOI: 10.1145/3375627.3375851.

[7] Laura Weidinger et al. "Taxonomy of Risks posed by Language Models". In: *FAccT*. 2022, pp. 214–229. DOI: 10.1145/3531146.3533088.