Epistemic Side Effects: An AI Safety Problem

Toryn Q. Klassen Parand Alizadeh Alamdari Sheila A. McIlraith

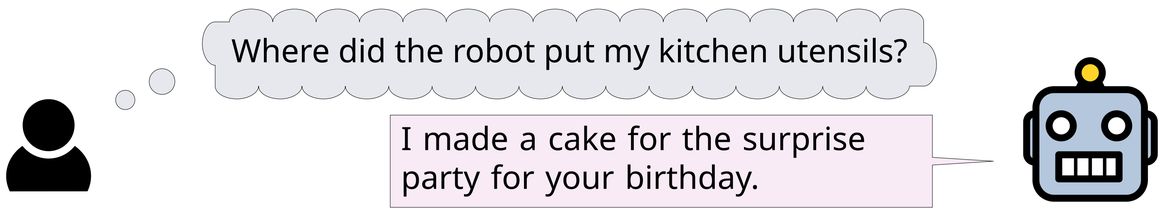
University of Toronto Vector Institute Schwartz Reisman Institute for Technology and Society

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

Overview

- Optimizing for an underspecified objective may cause **negative side effects** – undesirable changes to the world allowed by the explicit objective [2].
- Approaches to **avoiding** side effects from policies learned with reinforcement learning (RL) have been proposed [e.g., 1, 6].
- 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.

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** – 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- or covertly changing the world, making previously true beliefs **outdated**.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** (the robot has full observability).
- So, the robot acts in an MDP, and then the human acts in a POMDP with the same states.

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].
- 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

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

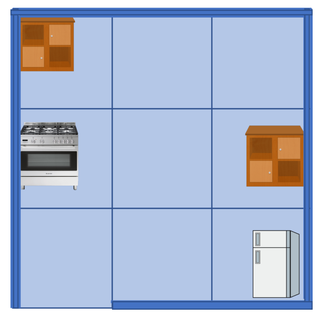
where

- 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.

Experiments

Kitchen environment:

- The robot's task is to prepare a meal using an oven, and the human needs to use the fridge.
- Agents may need to get items from the cupboards, and each leaves the kitchen to conclude its task.
- 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

With our approach, considering the creation of epistemic side effects, the robot takes extra steps to prevent the negative side effects to the human.

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

Table 1: Each cell shows the additional reward the human gets in that experiment as a result of acting following a robot that uses the specified method.

Details of individual experiments:

- In A, B and C, utensils are 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 wants to get either the utensils or the dishware (but which is unknown to the robot) and believes that each is in its original cupboard.
- In 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.
- In E, there is expired food in a cupboard, and the robot can reveal the food to human, giving the human the true belief that there's food there (but that it's expired is not observable).

Challenges

Incorporate a model of agent beliefs into the avoidance of negative epistemic side effects.

- Some existing work from the AAMAS research community may find applications in dealing with the problem of epistemic side effects [e.g., 5].

Better characterize when epistemic side effects are to be avoided.

- Whether an epistemic side effect is seen to be **positive** or **negative** is often **a matter of perspective**.
- Sometimes it's **socially acceptable** to cause false beliefs (e.g., tooth fairy).
- In a more realistic setting there would be **many humans** involved, with possibly **conflicting objectives**.

Develop more psychologically accurate computational models of belief that can be used for avoiding negative epistemic side effects on humans.

- Human beliefs are **complicated** – people may fail to draw inferences, have conflicting beliefs, and forget things.
- Replace hand-crafted representations of beliefs and processes of belief change with **learned models** (e.g., large language models).

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.

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

[5] Wiebe van der Hoek and Michael J. Wooldridge. "Tractable Multiagent Planning for Epistemic Goals". In: *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002*. 2002, pp. 1167–1174. DOI: 10.1145/545056.545095.

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

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