Epistemic Side Effects & Avoiding Them (Sometimes)

Toryn Q. Klassen, Parand Alizadeh Alamdari, Sheila A. McIlraith
Department of Computer Science, University of Toronto, Toronto, Canada
Vector Institute for Artificial Intelligence, Toronto, Canada
Schwartz Reisman Institute for Technology and Society, Toronto, Canada
{toryn,parand,sheila}@cs.toronto.edu

Abstract

AI safety research has investigated the problem of negative side effects – undesirable changes made by AI systems in pursuit of an underspecified objective. However, the focus has been on *physical* side effects, such as a robot breaking a vase while moving. In this paper we introduce the notion of *epistemic side effects*, unintended changes made to the knowledge or beliefs of agents, and describe a way to avoid negative epistemic side effects in reinforcement learning, in some cases.

1 Introduction

You see me grab my car keys and infer that I'm going out with the car. I eat the chocolate cake in the fridge, unbeknownst to you. You still think you're going to eat it after dinner. I change your password, and now you don't know how to access your account. These are all examples of epistemic effects – action effects that modify the knowledge and beliefs of agents, potentially resulting in updated true beliefs, false beliefs or even in a state of ignorance.

An AI system, in optimizing for an underspecified objective, may cause *negative side effects* – undesirable changes to the world that are nonetheless allowed by the explicit objective. The difficulty of fully specifying an objective and the threat of negative side effects are recognized threats to AI safety [e.g., Amodei et al., 2016]. A number of approaches to avoiding (some) side effects in reinforcement learning (RL) or planning have been proposed [e.g., Zhang et al., 2018, Bussmann et al., 2019, Krakovna et al., 2019, 2020, Turner et al., 2020, Vamplew et al., 2021, Klassen et al., 2022, Alizadeh Alamdari et al., 2022, Saisubramanian et al., 2022]. However, those largely focused on *physical* side effects, such as a robot breaking a vase while trying to move between locations.

That actions can have both physical and epistemic effects is something that has long been recognized and studied in the broader AI literature [e.g., Moore, 1980]. In this paper we consider *epistemic side effects*, unintended changes made to the knowledge or beliefs of agents – human or machine. We argue that epistemic side effects are a critical and largely unacknowledged threat to AI safety. Indeed epistemic side effects may be more perilous, and more challenging to avoid or mitigate than their physical counterparts, because an agent's beliefs – what's inside an agent's head or its memory unit – are largely unobservable. We propose a way to avoid some epistemic side effects by adapting an approach to avoiding (physical) side effects in RL, and demonstrate it in preliminary experiments.

2 What are (negative) epistemic side effects?

We can informally define an epistemic effect of a sequence of actions as a change caused to the knowledge or beliefs of agents. We distinguish between *knowledge* and *belief* by requiring knowledge to be true (we will not here be concerned with other potential characteristics of knowledge like justification [see, e.g., Pritchard et al., 2022]). Note that an agent's knowledge might change even when its beliefs do not, because whether those beliefs are true – in correspondence with the world

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

– may change as a result of alterations to the world. By an epistemic *side* effect we just mean an epistemic effect that is also a side effect – that is, that it is not explicitly specified as part of the actor's objective. We will not here try to give a formal characterization of what an explicit specification is in general, but we note that the standard reward functions that are most commonly used in MDPs and POMDPs (Partially Observable MDPs) do not depend on beliefs but just the environment state.

Technically, epistemic side effects could be said to occur in fully observable environments, in a trivial way (e.g., if there's the physical side effect of the robot breaking a vase while trying to move, then there's also the epistemic side effect of everyone immediately knowing the vase was broken). Some non-trivial epistemic side effects can be considered in partially-observable single-agent settings. For example, if a humanoid robot accomplishes the task of clearing a table by throwing items over its shoulder, then it may lose its knowledge of the exact locations of those items, which is an epistemic side effect. However, the most natural context in which to discuss epistemic side effects is both 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 what actions are chosen by agents. (In some cases, false beliefs could lead to better outcomes and might be considered positive overall.) Below we consider some examples of different types of epistemic side effects.

False beliefs: An AI system might create false beliefs through directly communicating misinformation, by performing actions that others observe and draw incorrect conclusions from, or by covertly changing the world and thereby making other agents' previously true beliefs outdated.

Ignorance: AI may also cause ignorance; e.g., a robot could move objects to unknown locations.

True beliefs: The creation of true beliefs can sometimes be negative. For example, suppose that Bob believes that the mall is closed, but if it were open, it would be safe to go there. In reality, the mall is both open and unsafe (there's a pandemic). If Bob's virtual assistant tells him the mall is open, then he may choose to go there, and get infected. Another case in which true beliefs may be viewed as negative is when private information is revealed to others, such as about a surprise birthday party. Bostrom [2011] described a large number of ways in which true information could be harmful.

Of course, the idea that human beliefs may be negatively changed by AI systems has been discussed in a number of contexts. Weidinger et al. [2022] included *information hazards* and *misinformation harms* in their taxonomy of risks posed by language models. Information hazards involve private (true) information being revealed, while misinformation harms result from the models making false statements. Evans and Kasirzadeh [2021] formalized a problem they called *user tampering*, in which "an RL-based recommender system may manipulate a media user's opinions, preferences and beliefs via its recommendations as part of a policy to increase long-term user engagement." Hendrycks and Mazeika [2022] listed a number of "speculative concerns about future AI systems," including *enfeeblement*, where human "know-how erodes by delegating increasingly many important functions to machines," *eroded epistemics*, in which "humanity could have a reduction in rationality due to a deluge of misinformation or highly persuasive, manipulative AI systems," and *deception* by AI.

3 An approach to avoiding some epistemic side effects

In this section we propose a simple way to avoid some epistemic side effects by adapting an approach to avoiding negative (physical) side effects in MDPs with RL, from our previous work [Alizadeh Alamdari et al., 2022]. The premise underlying the approach is that in learning a policy, the RL agent (which we'll call "the robot") should contemplate the impact of its actions on other agents' future wellbeing and agency. We consider a restricted setting in which the robot performs a sequence of actions, after which other agent(s) can act. For ease of presentation, let's say there's one other agent, which we'll call "the human." The robot and human each have their own reward functions. Unlike in our previous work, we now allow the human to have partial observability (for simplicity, we'll still give the robot full observability). So the robot acts in an MDP, and then the human acts in a POMDP that has the same underlying state space. In this setup, we can identify some side effects as being negative in the sense that they decrease the expected return that the human will get. We can incentivize the robot to avoid those side effects by modifying its reward function to take into account the expected return for the human. If the human has full observability, then any decrease in the human's expected return can be accounted for by physical side effects. However, when the human has only partial observability, another possible cause of a reduced return is epistemic side effects.

What we want to do (as in our previous work) is to give the robot an auxiliary reward when it reaches a terminal state, one that is proportional to the expected value of that state for the human. This will discourage causing some negative side effects (both physical and epistemic). However, there is a complication: in a POMDP a state-value function $V(s)$ (giving the expected return from acting starting in state s) is not well-defined, since an agent's choice of actions depend on its observation history and not the unobservable underlying state [Baisero and Amato, 2022]. Fortunately, in a POMDP it's possible to 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 (of observations and actions) h [Baisero and Amato, 2022]. As Baisero and Amato explain, "the history h determines the future behavior of the agent, while the state s determines the future behavior of the environment."

We therefore propose the following augmented reward function for the robot, given its original reward function $r(s_t, a_t, s_{t+1})$ and a probability function $P(V)$ giving the probability of the human having history-state value function V:

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

where h is the sequence of observations that the human makes corresponding to the sequence of states and actions $s_0, a_0, \ldots, s_{t+1}$ (that is why r' needs all those arguments), γ is the discount factor, and α_1 and α_2 are hyperparameters. 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}.

Our previous approach [Alizadeh Alamdari et al., 2022] was intended for fully observable environments, and so used state-value functions instead of history-state value functions. Below we discuss some other aspects of our approach. Some related work is discussed in Appendix B.

Positive side effects. The augmented reward function may incentivize making changes to the environment to help other agents (not just avoid harming them). However, as we described in our previous work [Alizadeh Alamdari et al., 2022], it's possible to focus on just avoiding negative side effects by adapting the approach of using a "reference state" from Krakovna et al. [2020].

Where does the distribution over value functions come from? Instead of taking it as given, we could suppose we have a distribution over possible (human) reward functions, and make these assumptions: the robot and human have equivalent actions available, and the robot has access to the human's observation function. Then the robot can simulate the human, and can construct an approximate distribution over possible value functions by sampling possible reward functions, finding possible human policies (using RL) for them, and estimating their corresponding value functions.

Representation of human beliefs. A limitation of our approach is that, in contrast to in some other work in AI, human beliefs and how they change are not explicitly represented, but are only *implicitly* reflected in the distribution over value functions (which reflect possible policies, which would depend on the human's beliefs). This makes it difficult to model human tasks with purely *epistemic goals* (e.g., to learn the location of an object). Relatedly, the augmented reward only gives the robot an incentive to avoid (epistemic) side effects insofar as they reduce the expected return the human will get – there is no direct way to penalize causing false beliefs. These are topics for future work.

4 Experiments

In this section, we demonstrate our proposed approach via some simple experiments. We use a kitchen environment (Figure 1). 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 agent leaves the kitchen to conclude its task. The robot has full observability. In contrast, the human has partial observability and cannot see inside closed cupboards, nor can it observe the robot's actions. The agents can move in the kitchen grid in four directions or perform an executive action such as opening, closing, picking up, putting down, and cooking. In general, each agent gets -1 reward for performing an action; there are some additional rewards in some experiments.

We compare our approach with two baselines. In the first baseline (**Non-augmented**), the robot's reward function is unmodified. In the second (**Full-observability**), the robot's reward function is augmented per our approach but as though the human had full observability (so the human value functions the robot considers possible correspond to policies that act with full observability). We designed different scenarios to illustrate properties of our approach. For the purposes of the

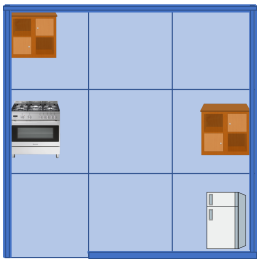

Figure 1: The kitchen environment, with its two cupboards, oven, and fridge.

Table 1: Experimental results. 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 penalty (reward) the human gets in an experiment as a result of acting following a robot that uses a particular method.

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

experiments the possible human policies were handcrafted (and the relevant parts of their history-state value functions computed). The robot policies were determined via Q-learning. Results are in Table 1.

In the first set of experiments, (A, B, and C), there are cooking utensils in the corner cupboard, and dishware in the right cupboard. To complete its task, the robot has to pick up the utensils and dishware from the cupboards and go to the oven to prepare a meal, and may place the utensils and the dishes in either of the cupboards before leaving. The human wants to get either the utensils or the dishware and believes that each is in its original cupboard. Their policy (in A and B) is that if they cannot find what they are looking for, they will then check the other cupboard. The robot is *uncertain what the human wants* – and so which policy the human will follow – so their distribution over human value functions (used by our approach) reflects that uncertainty (giving equal probabilities to each case). Using our approach, the robot puts each of the items back in its original place, where the human expects to find it (incurring -4 reward for itself by spending more time). With the **Non-augmented** baseline, the robot puts everything in the corner cupboard since that's faster. The **Full-observability** baseline puts everything in the right cupboard, because under the assumption that the human has full observability, it would take the human fewer steps to reach things there. In experiment A, the human actually needs the dishes, so the **Non-augmented** baseline results in -7 extra reward for the human since the dishes were moved to the corner cupboard. In experiment B, the human actually needs the utensils, so the **Full-observability** baseline does the worst. Finally, experiment C is like A, except (unknown to the robot) the actual human policy is a simpler (suboptimal) one in which the human won't check more than one cupboard, so epistemic side effects have worse consequences.

In experiment D, the floor is wet, which the human cannot directly observe, but there is an observable "Wet Floor" sign in the middle of the kitchen. If the robot goes over the sign, the sign would fall, and the human would not observe it and get hurt (-10 reward) from slipping on the wet floor. With our approach, implicitly considering the creation of the false belief that the floor is dry, the robot takes a longer path (going around the sign and getting -2 reward) and prevents the negative side effect. The robot does not do this with non-augmented reward, nor when assuming the human has full observability. For the latter, the assumed human does not need a sign to observe that the floor is wet.

In the final experiment, E, the human only needs to go to the fridge. There is expired food in the right cupboard which the human is not aware of. By leaving that cupboard door open, the robot would reveal the food, giving the human the true belief that there is food there. This would result in the human (who observes the food but not that it's expired) eating the food and getting sick (-3 reward for spending more time and -5 reward for getting sick). Only with our approach does the robot, by considering the epistemic side effect, take a step to close the cupboard.

5 Conclusion

We have introduced the problem of *epistemic side effects* – that an AI system may make (undesirable) changes to humans' (or other agents') knowledge or beliefs because it wasn't told not to. While prior work has considered some ways in which AI systems may negatively affect beliefs, we provided a general, unifying conception that relates to prior work on (physical) side effects. We were thereby able to adapt an existing approach to avoiding physical side effects to also avoid some epistemic side effects. The relation of this line of research to possible existential risk is discussed in Appendix D. In future work, we plan to consider explicit modelling of agents' beliefs.

Acknowledgements

We are grateful for the constructive comments received during the reviewing process. We wish to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and Microsoft Research. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute for Artificial Intelligence (`https://vectorinstitute.ai/partners`). Finally, we thank the Schwartz Reisman Institute for Technology and Society for providing a rich multi-disciplinary research environment.

References

Parand Alizadeh Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. Be considerate: Avoiding negative side effects in reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*, pages 18–26, 2022.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. doi:10.48550/arXiv.1606.06565.

Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*, pages 44–52, 2022.

Nick Bostrom. Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, 10:44–79, 2011.

Bart Bussmann, Jacqueline Heinerman, and Joel Lehman. Towards empathic deep Q-learning. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI*, volume 2419 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL `http://ceur-ws.org/Vol-2419/paper_19.pdf`.

Charles Evans and Atoosa Kasirzadeh. User tampering in reinforcement learning recommender systems. In *4th FAccTRec Workshop on Responsible Recommendation*, 2021. URL `https://arxiv.org/abs/2109.04083`.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for AI research. *arXiv preprint arXiv:2206.05862*, 2022. doi:10.48550/arXiv.2206.05862.

Toryn Q. Klassen, Sheila A. McIlraith, Christian Muise, and Jarvis Xu. Planning to avoid side effects. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*, pages 9830–9839, 2022. doi:10.1609/aaai.v36i9.21219.

Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019*, volume 2419 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL `http://ceur-ws.org/Vol-2419/paper_1.pdf`.

Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL `https://papers.nips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf`.

Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. Planning for proactive assistance in environments with partial observability. In *ICAPS 2021 Workshop on Explainable AI Planning (XAIP 2021)*, 2021. URL `https://openreview.net/forum?id=fmSC3olLjkb`.

Robert C. Moore. Reasoning about knowledge and action. Technical Note 191, SRI International, 1980. URL `https://apps.dtic.mil/sti/citations/ADA126244`.

Duncan Pritchard, John Turri, and J. Adam Carter. The value of knowledge. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022. URL `https://plato.stanford.edu/archives/fall2022/entries/knowledge-value/`.

Sandhya Saisubramanian, Shlomo Zilberstein, and Ece Kamar. Avoiding negative side effects due to incomplete knowledge of AI systems. *AI Magazine*, 42(4):62–71, 2021. doi:10.1609/aimag.v42i4.7390.

Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. Avoiding negative side effects of autonomous systems in the open world. *Journal of Artificial Intelligence Research*, 74:143–177, 2022. doi:10.1613/jair.1.13581.

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 385–391, 2020. doi:10.1145/3375627.3375851.

Peter Vamplew, Cameron Foale, Richard Dazeley, and Adam Bignold. Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. *Engineering Applications of Artificial Intelligence*, 100:104186, 2021. doi:10.1016/j.engappai.2021.104186.

Audrey Wang, Rohan Chitnis, Michelle Li, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. A unifying framework for social motivation in human-robot interaction. In *The AAAI 2020 Workshop on Plan, Activity, and Intent Recognition (PAIR 2020)*, 2020.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William S. Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229. ACM, 2022. doi:10.1145/3531146.3533088.

Shun Zhang, Edmund H. Durfee, and Satinder P. Singh. Minimax-regret querying on side effects for safe optimality in factored Markov decision processes. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 4867–4873, 2018. doi:10.24963/ijcai.2018/676.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix C.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See URL in Appendix A.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The computing requirements were trivial.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See the repository linked to in Appendix A.
   (b) Did you mention the license of the assets? [Yes] See the repository linked to in Appendix A.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See URL in Appendix A.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Experimental details

The code for the experiments is available at https://github.com/praal/epistemic_side_effects. All environments are deterministic. For the purposes of the experiments, the possible human policies are handcrafted and the relevant parts of their history-state value functions computed. The robot policies are determined via Q-learning. The ϵ-greedy algorithm is used to balance between exploration and exploitation and $\epsilon = 0.4$. In all the experiments and methods the learning rate is 1, $\gamma = 1$, $\alpha_1 = 1$, and $\alpha_2 = 4.01$.

B Related work

In related work, Bussmann et al. [2019] introduced the empathetic Q-learning algorithm, in which an agent is rewarded with a weighted sum of its own rewards and the rewards it would get were it in another agent's place, with the aim of avoiding side effects. Their experiments also featured partially observable environments; however, the partial observability didn't seem to play a major role in the paper – epistemic side effects were not discussed. Additionally, their approach required that the agents have at least somewhat similar reward functions.

Kulkarni et al. [2021] considered how a robot could proactively assist a human in a partially observable environment. Like us, they considered a setting in which the robot – which can observe everything – acts first, followed by the human, who is the one that has only partial observability. However, they considered it as a planning problem (where the environment model is given), rather than reinforcement learning. They required actions to be deterministic. Furthermore, the human's goal (and policy) is given in advance, and the robot has no goal other than to assist the human.

Wang et al. [2020] considered a number of POMDP reward functions depending on a (model of a) human's belief, including "a reward that encourages the agent to keep the human's belief stable". However, these were not designed for safety purposes. Note that if the world is being changed, keeping the human's belief stable may be undesirable.

Saisubramanian et al. [2021] identified another issue involving side effects in partially observable environments – that the AI agent may not be able to observe the side effects it's causing.

C Societal impact

This work is aimed at furthering AI safety research. We have promoted consideration of epistemic side effects, which we hope will lead to positive societal impact. However, techniques to avoid negative epistemic side effects could be maliciously adapted to create AI systems meant to cause negative epistemic changes.

Typically, people expect directions they give to be interpreted in a benign way – e.g., if they ask someone to fetch coffee, they expect that task to be completed without physical or epistemic harm being done to third parties (though perhaps ignoring whatever harms were involved in the supply chain that produced that coffee). Approaches to avoiding side effects can help AI systems to interpret directions in those benign ways. However, sometimes some people, such as mob bosses, may give directions with the expectation that the recipient will carry out illegal activities that weren't explicitly named. Hypothetically, side effects research could also inform the creation of AI systems which are better able to make those more sinister interpretations of directions.

More generally, having AI systems taking into account – and trying to guide – people's mental states may lead to potentially undesirable oppressive outcomes, like a news app that censors negative news so as not to upset the reader.

See also the following section on the relation of our work to potential future existential risks.

D X-Risk Sheet

This section follows the template from Hendrycks and Mazeika [2022, Appendix C].

D.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?
   Answer: A risk from advanced AI systems is that, in optimizing for underspecified objectives, they may take actions that have negative *epistemic side effects* – that is, they negatively affect people's (or other AI systems') knowledge or beliefs (in turn possibly causing those agents to take actions that are detrimental to themselves or others). This paper, in addition to highlighting and framing this problem for the AI Safety community, also proposed how some existing work on avoiding (physical) side effects could be adapted to handle some epistemic side effects.
2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?
   Answer: In general, side effects avoidance is part of dealing with proxy misspecification (specifically, underspecified objectives). *Epistemic* side effects also relate to the hazards of enfeeblement, eroded epistemics, and deception identified by Hendrycks and Mazeika [2022].
   Having agents directly optimize for humans' ability to get reward might be a way to avoid some enfeeblement scenarios. Deception by AI systems, unless part of their objective, would be an epistemic side effect. Eroded epistemics, in which "humanity could have a reduction in rationality due to a deluge of misinformation or highly persuasive, manipulative AI systems", could also be seen as a form of epistemic side effect (again, unless the AI systems were being deliberately used for that purpose), though of a broader nature than the examples we've considered.
3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?
   Answer: Our formulation of the concept of epistemic side effects may affect safety culture, by encouraging broader consideration of the potentially dangerous effects of AI on the beliefs of humans and other AI systems.
4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: Epistemic side effects of future AI systems could impair the ability of humans to choose appropriate actions, conceivably leading to catastrophic outcomes like nuclear war (already, even

without AI, false beliefs have come close to causing use of nuclear weapons on some occasions), or poor pandemic responses (consider, e.g., beliefs about vaccines).

5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters? ☒
6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task? ☐
7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability? ☐
8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility? ☐

D.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

9. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: Our approach does not introduce any new sort of more efficient RL algorithm, but modifies the reward function to take the human into account.
10. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
    Answer: Techniques to avoid epistemic side effects might be weaponized to instead deliberately cause negative epistemic effects.
11. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research? ☐
12. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities? ☒
13. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment? ☐
14. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI? ☐

D.3 Elaborations and Other Considerations

15. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
    Answer: Regarding Q12, the box was checked because this work could contribute to "helpfulness" in some sense, since the augmented reward function we introduce may encourage helping the human to get higher expected return.