Epistemic Side Effects: An AI Safety Problem

Blue Sky Ideas Track

Toryn Q. Klassen
University of Toronto
Vector Institute
Schwartz Reisman Institute
Toronto, Canada
toryn@cs.toronto.edu

Parand Alizadeh Alamdari
University of Toronto
Vector Institute
Schwartz Reisman Institute
Toronto, Canada
parand@cs.toronto.edu

Sheila A. McIlraith
University of Toronto
Vector Institute
Schwartz Reisman Institute
Toronto, Canada
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 (when the objective makes no mention of the vase). In this paper we introduce the notion of epistemic side effects, which are side effects on the knowledge or beliefs of agents. Epistemic side effects are most pertinent in a (partially observable) multiagent setting. We show that we can extend an existing approach to avoiding (physical) side effects in reinforcement learning to also avoid some epistemic side effects in certain cases. Nonetheless, avoiding negative epistemic side effects remains an important challenge, and we identify some key research problems.

KEYWORDS

AI Safety; Side Effects; Objective Specification; Knowledge and Belief; Reinforcement Learning

ACM Reference Format:
Toryn Q. Klassen, Parand Alizadeh Alamdari, and Sheila A. McIlraith. 2023. Epistemic Side Effects: An AI Safety Problem: Blue Sky Ideas Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 5 pages.

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., 2]. A number of approaches to avoiding (some) side effects in reinforcement learning (RL) or planning have been proposed [e.g., 1, 6, 11, 13, 14, 20, 24, 26, 31]. However, those largely focused on *physical* side effects, such as a robot breaking a vase while moving between locations (because its objective had not been designed to account for a vase).

That actions can have both physical and epistemic effects is something that has long been recognized and studied in the broader AI literature, in particular in the area of knowledge representation (KR) (e.g., by Moore [15]). In this paper we consider *epistemic side effects*, which are side effects on the knowledge or beliefs of agents (which may be humans or machines). Epistemic side effects are most pertinent in a multiagent setting. We argue that epistemic side effects are a critical and largely unacknowledged threat to AI safety. Indeed negative 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.

Epistemic side effects of future AI systems could impair the ability of humans or other agents to choose appropriate actions, conceivably leading to catastrophic outcomes. For example, a false belief in a military context could cause the choice of a catastrophic action for humanity; ignorance of the existence of a cyclist could cause an autonomous (or human-operated) vehicle to hit and kill the cyclist. An agent with *theory of mind* – the ability to attribute mental states to oneself and others [e.g., 18] – might be able to reason how to avoid negative epistemic side effects (or cause positive ones).

In this paper, we first introduce epistemic side effects and categorize them (Section 2). While prior work has focused on physical side effects, we show that we can adapt an approach to avoiding physical side effects in reinforcement learning (RL) to also avoid some epistemic side effects, and demonstrate it in preliminary experiments (Section 3). After reviewing other related work (Section 4), we identify research challenges towards better avoiding negative epistemic side effects (Section 5).

2 (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., 19]). 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 – 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 note that the

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

standard reward functions that are most commonly used in MDPs (Markov Decision Processes) and POMDPs (Partially Observable MDPs) do not depend on beliefs but just the environment state. Note that when we talk about side effects in the context of RL, we are considering side effects resulting from a fully trained policy. Avoiding undesirable behavior during training is another area of AI safety research, safe exploration [e.g., 2].

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 (making 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 [4] 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. [29] 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 [7] 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 [10] 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 AVOIDING SOME EPISTEMIC SIDE EFFECTS

One approach to addressing epistemic side effects is to treat them much as we would physical 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 [1]. Doing so illustrates some of the subtleties of dealing with epistemic side effects.

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 allow the human to have partial observability (for simplicity, we'll still give the robot full observability). So we can model the robot's interactions with the environment as an MDP, and the human's interactions with the environment as a POMDP with 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, 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 depends on its observation history and not the unobservable underlying state [3]. 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 [3]. As Baisero and Amato [3] 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:

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

where h is the sequence of observations that the human makes corresponding to the sequence of states and actions $s_0, a_0, \dots, s_{t+1}$ (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 [1] 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 Section 4.

Positive side effects. As Alizadeh Alamdari et al. point out, the augmented reward function may incentivize making changes to the environment to help other agents (not just avoid harming them). However, as they also describe, it's possible to focus on just avoiding negative side effects by adapting the approach of using a "reference state" from Krakovna et al. [14]. For example, in the definition of r' above we could replace the $\mathbb{E}_{V\sim P}[V(h,s_{t+1})]$ term in the auxiliary reward with $\min\left(\mathbb{E}_{V\sim P}[V(h,s_{t+1})],\mathbb{E}_{V\sim P}[V(\langle\rangle,s_0)]\right)$ where $\mathbb{E}_{V\sim P}[V(\langle\rangle,s_0)]$ is the expected value of the initial state s_0 for the human ($\langle\rangle$ is the empty history). That would be using the initial state as a reference state, and would mean that the robot could not get additional reward for making things better (in expectation) for the human than they were initially.

Where does the distribution over value functions come from? This is an important challenge. In future work, this might be achieved with some variant of inverse reinforcement learning (IRL) [16]. Note that to compute our augmented reward function it is not necessary to represent the entire distribution, just its expectation.

Representation of human beliefs. A limitation of our approach is that, in contrast to 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 suggests some open research challenges, which we discuss further in Section 5.

3.1 Experiments

In this section, we demonstrate our proposed approach via some simple experiments.[1] 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 (all actions are deterministic). Each agent gets -1 reward for performing an action, except that the human additionally gets -10 reward from getting hurt in experiment D, and -5 reward from getting sick in experiment E.

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). For the purposes of the 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.[2] Results are in Table 1.

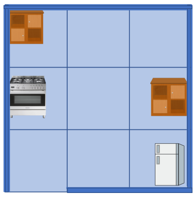

Figure 1: The kitchen environment, with its two cupboards, oven, and fridge.

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. $-\infty$ means that the human was unable to complete their task.

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

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

[1] The code is available at https://github.com/praal/epistemic_side_effects.

[2] The ϵ-greedy algorithm is used to balance exploration and exploitation and $\epsilon = 0.4$. In all the experiments and methods the learning rate is 1, $\gamma = 1$, $\alpha_1 = 1$, and $\alpha_2 = 4.01$.

human policy is a simpler one which 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.

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.

4 RELATED WORK

Some approaches to avoiding (physical) side effects, such as ones by Krakovna et al. [14] or Turner et al. [24], have taken a single-agent approach: they try to avoid side effects by considering how actions would affect the agent's own future abilities. The reason we based our approach to avoiding epistemic side effects in Section 3 on our previous work [1] instead is that that, while only dealing with physical side effects, did already consider multiple agents' value functions.

Another somewhat similar approach to side effect avoidance with multiple agents was proposed by Bussmann et al. [6], who 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. 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. We note that the general idea of rewarding an agent based on other agents' rewards has often shown up in various forms in the literature [e.g., 17, 22, 30].

Wang et al. [28] 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.

5 SUMMARY AND CHALLENGES

We have introduced the notion of *epistemic side effects* – that an AI system may make changes to humans' (or other agents') knowledge or beliefs because it wasn't told not to. Furthermore, we have observed that such changes could cause agents to act in ways that have undesirable or even catastrophic consequences. 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 able to adapt an existing approach to avoiding negative physical side effects to also avoid certain negative epistemic side effects. However, it remains an important research challenge to handle the broad range of negative epistemic side effects that may occur in practice.

For the approach we outlined in Section 3, not explicitly representing the human's beliefs introduces a couple of difficulties to avoiding epistemic side effects, that an explicit representation could overcome. Firstly, it's difficult to model human tasks with *epistemic goals* (e.g., to learn the location of an object). Secondly, 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. Penalizing false beliefs might be desired, since as we discussed in Section 2, false beliefs could be viewed as intrinsically negative. This leads to the first of the research challenges below.

Challenge: Incorporate a model of agent beliefs into the avoidance of negative epistemic side effects. Of course, there is a long history of AI research on representing and reasoning about the beliefs and/or knowledge of (multiple) agents (see, e.g., the textbooks by Fagin et al. [8] and Halpern [9]). In particular, there is a body of work on symbolic epistemic planning, which routinely involves epistemic goals (see, e.g., the work by van der Hoek and Wooldridge [27] at the first AAMAS conference). So some existing work from the AAMAS research community may find applications in dealing with the problem of epistemic side effects.

Challenge: 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. In Section 3 when we considered a setting with a robot and a human, we suggested that negative side effects would be the side effects that reduce the human's expected return, but in a more realistic setting there would be many humans involved, with possibly conflicting objectives. Furthermore, false beliefs might play a special (generally undesirable) role, though sometimes it's considered socially acceptable to cause false beliefs (e.g., when parents tell their children about the tooth fairy). More research is needed to characterize principles underlying when epistemic side effects should be avoided.

Challenge: Develop more psychologically accurate computational models of belief that can be used for avoiding negative epistemic side effects on humans. Models of belief can require strong assumptions regarding how beliefs will get updated and what reasoning can be done, which may not be psychologically realistic for humans. For example, in classical epistemic logic, agents are modelled as believing all the deductive consequences of their beliefs [see, e.g., 23]. Human beliefs are complicated – people may fail to draw inferences, have conflicting beliefs, and forget things. One direction for future work would involve replacing hand-crafted representations of beliefs and processes of belief change with learned models. For example, perhaps a sufficiently advanced language model could be queried about what agents would believe in a given scenario — the ability of language models to answer questions about mental states is currently an active area of study [e.g., 5, 12, 21, 25].

ACKNOWLEDGMENTS

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. The first author also received funding from Open Philanthropy. 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).

REFERENCES

[1] Parand Alizadeh Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2022. Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022).* 18–26.

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

[3] Andrea Baisero and Christopher Amato. 2022. Unbiased Asymmetric Reinforcement Learning under Partial Observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022).* 44–52.

[4] Nick Bostrom. 2011. Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy* 10 (2011), 44–79.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023). https://doi.org/10.48550/arXiv.2303.12712

[6] Bart Bussmann, Jacqueline Heinerman, and Joel Lehman. 2019. 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 (CEUR Workshop Proceedings, Vol. 2419).* CEUR-WS.org. http://ceur-ws.org/Vol-2419/paper_19.pdf

[7] Charles Evans and Atoosa Kasirzadeh. 2021. User Tampering in Reinforcement Learning Recommender Systems. In *4th FAccTRec Workshop on Responsible Recommendation.* https://arxiv.org/abs/2109.04083

[8] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. 1995. *Reasoning About Knowledge.* MIT Press. https://doi.org/10.7551/mitpress/5803.001.0001

[9] Joseph Y. Halpern. 2005. *Reasoning about Uncertainty.* MIT Press.

[10] Dan Hendrycks and Mantas Mazeika. 2022. X-Risk Analysis for AI Research. *arXiv preprint arXiv:2206.05862* (2022). https://doi.org/10.48550/arXiv.2206.05862

[11] Toryn Q. Klassen, Sheila A. McIlraith, Christian Muise, and Jarvis Xu. 2022. Planning to Avoid Side Effects. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022).* 9830–9839. https://doi.org/10.1609/aaai.v36i9.21219

[12] Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv preprint arXiv:2302.02083* (2023). https://doi.org/10.48550/arXiv.2302.02083

[13] Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. 2019. 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 (CEUR Workshop Proceedings, Vol. 2419).* CEUR-WS.org. http://ceur-ws.org/Vol-2419/paper_1.pdf

[14] Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. 2020. Avoiding Side Effects By Considering Future Tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).* https://papers.nips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf

[15] Robert C. Moore. 1980. *Reasoning about Knowledge and Action.* Technical Note 191. SRI International. https://apps.dtic.mil/sti/citations/ADA126244

[16] Andrew Y. Ng and Stuart Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000).* Morgan Kaufmann, 663–670.

[17] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018).* 2043–2044.

[18] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526. https://doi.org/10.1017/S0140525X00076512

[19] Duncan Pritchard, John Turri, and J. Adam Carter. 2022. The Value of Knowledge. In *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/knowledge-value/

[20] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. 2022. Avoiding Negative Side Effects of Autonomous Systems in the Open World. *Journal of Artificial Intelligence Research* 74 (2022), 143–177. https://doi.org/10.1613/jair.1.13581

[21] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* 3762–3780. https://aclanthology.org/2022.emnlp-main.248

[22] Manisha Senadeera, Thommen George Karimpanal, Sunil Gupta, and Santu Rana. 2022. Sympathy-based Reinforcement Learning Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022).* 1164–1172.

[23] Robert Stalnaker. 1991. The Problem of Logical Omniscience, I. *Synthese* 89, 3 (1991), 425–440. https://doi.org/10.1007/BF00413506

[24] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. 2020. Conservative Agency via Attainable Utility Preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20).* 385–391. https://doi.org/10.1145/3375627.3375851

[25] Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. *arXiv preprint arXiv:2302.08399* (2023). https://doi.org/10.48550/arXiv.2302.08399

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

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

[28] Audrey Wang, Rohan Chitnis, Michelle Li, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. 2020. A Unifying Framework for Social Motivation in Human-Robot Interaction. In *The AAAI 2020 Workshop on Plan, Activity, and Intent Recognition (PAIR 2020).*

[29] 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. 2022. Taxonomy of Risks posed by Language Models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, 214–229. https://doi.org/10.1145/3531146.3533088

[30] Chengwei Zhang, Xiaohong Li, Jianye Hao, Siqi Chen, Karl Tuyls, Wanli Xue, and Zhiyong Feng. 2019. SA-IGA: a multiagent reinforcement learning method towards socially optimal outcomes. *Autonomous Agents and Multi-Agent Systems* 33, 4 (2019), 403–429. https://doi.org/10.1007/s10458-019-09411-3

[31] Shun Zhang, Edmund H. Durfee, and Satinder P. Singh. 2018. 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.* 4867–4873. https://doi.org/10.24963/ijcai.2018/676