User-Aligned Assessment of Agent Learning and Operation for Reliable Autonomy

Sandhya Saisubramanian
Assistant Professor, Oregon State University, Oregon, USA

Abstract

Autonomous agents are typically evaluated based on their performance on assigned tasks, using aggregate measures of accuracy. Such assessments often hide misaligned behaviors such as negative side effects. For a learning-enabled system, learning from human feedback is a promising approach to learn to produce user-aligned behaviors in settings where precise specification of user-aligned objectives and their associated reward functions can be challenging. However, in many situations, the agent learns a proxy reward instead of the intended reward, leading to undesirable, unpredictable behaviors. To support long-term reliable autonomy, autonomous systems assessments must *evaluate both agent learning and performance*. To that end, we present techniques to (1) learn to avoid negative side effects that are discovered after deployment; (2) ensure that the agent learns the intended reward and not a proxy from human feedback; and (3) self-monitor its behavior to gradually reduce reliance on humans.

Summary of Recent Research

An autonomous agent's behavior is determined by its reward function. It is challenging to accurately specify the objectives and the associated reward functions for agents operating in complex environments. As a result, agents often operate based on incomplete specifications, which may produce undesirable behaviors.

Concern 1 *Evaluation metrics that focus on aggregate measures of performance accuracy may not uncover the negative side effects of agent actions.*

Consider a household cleaning robot. If the evaluation focuses on the task performance such as the cleanliness of the floor, side effects such as the water sprayed on power sockets may not be uncovered or addressed. However, such undesirable consequences of agent behavior have a significant impact on how users view and interact with AI systems.

Learning from human feedback is a popular approach to train agents in settings where it is challenging to accurately specify the objective and a reward function that produce a desired behavior. A reward function is produced by mapping the information in the feedback to real values. The success of this approach has been documented in many research studies (Ng and Russell 2000; Ramachandran and Amir 2007; Cui et al. 2021; Ibarz et al. 2018).

Concern 2 *The agent may learn a proxy instead of the intended reward function.*

A major barrier to large-scale real-world deployment of agents trained with human feedback is that the agent may have learned a proxy reward function that correlates with the training data. Current learning methods offer no mechanisms to verify if the agent learned a proxy or intended reward function *during* the learning process. This may be uncovered during its performance assessment after deployment. It is also difficult to localize the source of undesirable behavior—whether the concern arises from incorrect learning or incorrect planning (incorrect objective). While there are many other reasons for undesirable behavior, we will focus only on the planning and learning component in this work.

Concern 3 *Assessments based on aggregate measures of accuracy of the agent's policy and its learned reward do not uncover potential unsafe behaviors.*

Below we describe some of the techniques developed to better align agent behavior with user expectations and preferences, by (1) identifying and mitigating negative side effects, without affecting the task performance; (2) refining learned reward function based feedback to agent explanations; and (3) gradually reducing the reliance on humans via agent self-monitoring capabilities. These methods involve human-in-the-loop evaluation and move beyond the traditional notions of accuracy, loss function, or optimal policy for evaluating an autonomous system.

1. Mitigating negative side effects Negative side effects are the unexpected, undesirable consequences of agent actions that arise due to incomplete specification. Negative side effects are often discovered after deployment, using human feedback. In situations where the side effects are not catastrophic, the agent must avoid it, without significantly affecting its task performance. Avoiding negative side effects (NSEs) involves the following steps: (1) gather information about NSEs from different forms of feedback; (2) learn a predictive model of NSEs to generalize the gathered information to unseen situations; and (3) plan to mitigate NSEs, without significantly affecting the task completion. We present learning and planning techniques to avoid *Markovian and non-Markovian* NSE (Srivastava et al. 2023; Saisubramanian, Kamar, and Zilberstein 2020, 2022). Markovian NSEs are learned and represented in a tabular format, and non-Markovian NSEs using a finite state machine. Planning using a learned NSE model is performed

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

using a multi-objective approach with lexicographic reward preferences (Saisubramanian, Kamar, and Zilberstein 2020), a human-agent team approach (Saisubramanian, Kamar, and Zilberstein 2022), and a constraint optimization approach (Srivastava et al. 2023). Our approaches offer a principled way to balance the trade-off between side effects mitigation and task performance.

2. Learning human-aligned reward functions Agents often learn a proxy reward function when presented with expert demonstrations, since multiple reward functions may be consistent with the demonstrated behavior. To support safe deployment, it is necessary to ensure that the agent learns a reward function that is aligned with the demonstrator's intended reward. Existing methods assess reward alignment *after* the learning process is complete, which is ineffective and offers little scope for amending incorrect rewards. In a recent work (Mahmud, Saisubramanian, and Zilberstein 2023), we present an algorithm for verifying reward alignment *during* the learning process in a Bayesian inverse reinforcement learning setting. Our approach generates explanations of the reward model, which are evaluated by a human tester. Based on the tester's feedback, the agent updates its posterior, thereby reducing the ambiguity associated with the learned reward and ensuring that the learned reward is the intended reward. We are currently investigating how the feedback type affects learning the intended reward function and its sample efficiency.

3. Self-monitoring for safe operation Agent actions may sometimes produce novel undesirable effects when the region of operation is expanded or when it is updated to improve performance, based on the data collected. Relying on human feedback to learn a predictive model of unsafe actions from scratch every time the agent operates in a new region is impractical. We want these systems to be able to self-monitor their behavior to detect and mitigate undesirable consequences. In a recent work (Svegliato et al. 2022), we use metareasoning to continuously monitor the underlying task process to detect safety violations and identify the right action to quickly recover from the situation, while minimally interfering with task completion. We are also extending this to cooperative multi-agent settings. Specifically, joint action execution of multiple agents in a shared environment may produce negative side effects if their training does not account for the behavior of other agents or their joint action effects on the environment. Instead of relying on human feedback to discover the side effects associated with all possible agent interactions across all tasks, we present a metareasoning approach to detect and mitigate such side effects. The metareasoner estimates the joint penalty and decomposes it into individual penalties for each agent using credit assignment, thereby facilitating decentralized policy computation.

References

Cui, Y.; Zhang, Q.; Knox, B.; Allievi, A.; Stone, P.; and Niekum, S. 2021. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, 604–626.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Mahmud, S.; Saisubramanian, S.; and Zilberstein, S. 2023. Explanation-guided reward alignment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 473–482.

Ng, A.; and Russell, S. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, 2.

Ramachandran, D.; and Amir, E. 2007. Bayesian Inverse Reinforcement Learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*.

Saisubramanian, S.; Kamar, E.; and Zilberstein, S. 2020. A Multi-Objective Approach to Mitigate Negative Side Effects. In *Proc. of the 29th Intl. Joint Conf. on Artificial Intelligence*.

Saisubramanian, S.; Kamar, E.; and Zilberstein, S. 2022. Avoiding Negative Side Effects of Autonomous Systems in the Open World. *Journal of Artificial Intelligence Research*, 74: 143–177.

Srivastava, A.; Saisubramanian, S.; Paruchuri, P.; Kumar, A.; and Zilberstein, S. 2023. Planning and Learning for Non-Markovian Negative Side Effects Using Finite State Controllers. In *Proc. of the 37th AAAI Conference on Artificial Intelligence*.

Svegliato, J.; Basich, C.; Saisubramanian, S.; and Zilberstein, S. 2022. Metareasoning for Safe Decision Making in Autonomous Systems. In *Proc. of the Intl. Conf. on Robotics and Automation*.