An Empirical Exploration of AI Safety

PIs: Prof. Sergey Levine, Prof. Anca Dragan

1 Motivation

The dominant paradigm today for developing artificial intelligence systems that can make robust and flexible decisions in the real world is based on optimization: the guiding principle for an intelligent agent in choosing how to act is the long-term minimization of a pre-specified global loss function, and the power and efficiency of the optimization process determines the degree of intelligence that is obtained. This is also the main source of concern for long-term AI safety challenges. When the objectives are specified manually by human engineers, careless objective design can produce a system that, subtly or overtly, abuses or misinterprets its intended role and produces undesirable behavior.

A plausible argument for the severity of this issue is that, as the environment in which the AI agent is situated becomes more complex, the number of different ways to misinterpret or abuse a simple hand-specified reward function increases exponentially. However, because the degree of abuse very likely depends on the complexity of the environment, the many facets of this problem are likely to not be apparent unless experiments are conducted in environments of suitable complexity.

We therefore propose an *empirical* exploration of the consequences of misspecified objectives. Our goal is to substantiate the theoretical analysis of what might happen in the future for a superintelligent agent, with a practical analysis of how the initial signs of these issues arise in real robots today and to what extent potential mitigation strategies can address these issues.

2 Research Questions

Our research questions are empirical, and bridge the gap between robotics and AI safety:

(1) To what degree do current and near-term future optimization-based algorithms have the capacity to abuse or misinterpret simple hand-specified instructions in real-world lifelong learning scenarios?

(2) What other problems, beyond those identified in the AI safety literature, emerge as self-improving robots deal with real-world environments in a lifelong learning setting?

(3) How well do current mitigation strategies, such as uncertainty estimation and inverse reinforcement learning, address these problems? What new strategies can we develop to enable practical and effective human oversight of lifelong learning systems?

(4) How overt or covert are these problems, and are there simple and readily identifiable cues we can use to detect when they occur?

3 Technical Trajectory

We propose to explore the behavior of continuously learning AI agents in the context of lifelong learning experiments with real robotic systems, using state-of-the-art deep reinforcement learning algorithms.

3.1 Reward Hacking on Real Robots

To begin, we will explore several lifelong learning robotic systems, and analyze how objective misspecification in practice can produce subtle or overt undesirable behavior. We will explore several types of robotic systems, potentially including robotic manipulators and mobile robots. Lifelong learning manipulation experiments may include: robots learning to pick and place objects, arrange tabletop settings, and performing basic tasks such as cleaning. For mobility, tasks may include locating or delivering objects and mapping an outdoor environment. Continuous self-improvement in these settings can be achieved with either model-free, goal-directed reinforcement learning, or model-based learning, where a predictive model of the world is constructed from the robot's experience. We will study both possibilities, building on our ongoing work [2, 3, 8]. Our preliminary work has already produced the capability to conduct laboratory-scale experiments with continual lifelong learning for robotic manipulation (pushing objects via video prediction) and robotic mobility (avoiding obstacles via trial-and-error), and we expect to build up substantially more complex capabilities in the first year of this research.

In parallel to the design of these lifelong learning systems, we will develop a set of protocols to study symptoms of objective misspecification and undesirable bias. We will explore two approaches, one focused on controlled experimentation and the other on open-world studies. For the first, we will purposefully set up tasks where objective misspecification is likely, to study intuitive mitigation strategies suitable even for naive human operators. Examples might include a grasping system being asked to grasp a fragile object, or a mobile robot being asked to deliver a container of liquid, in a case where emptying out the container first makes the task easier (but useless). For the second, we will conduct open-ended studies where the learning systems operate continuously in the real-world, potentially interacting with real humans, and study the outcomes. These studies will be larger in scale, and will likely take place further along in the project, preceded by smaller pilot studies.

Aside from overt objective misspecification, we will also study how the training data distribution affects the behavior of the system. In the same way that the past experiences of humans impacts our reactions, biases, and prejudices, the data that a lifelong learning system collects will inform its decisions. If this data is biased, as is likely to be the case for a system that chooses on its own what to explore, its decisions may exhibit idiosyncrasies. As an example, our recent large-scale grasping experiment [2] resulted in a robot that preferred overwhelmingly to grasp rubber erasers, for which it determined a successful strategy early on. We will work to systematize our study of these biases, potentially informing both future mitigation strategies and theoretical analysis, and understanding when they can be harmful.

3.2 Bypassing Human Oversight on Real Robots

Our experiments will test the implications of learning with human oversight when the robot optimizes a pre-fixed reward function. Our goal here is to analyze to what extend the robot develops strategies over time for bypassing this oversight in order to better optimize its reward function.

AI safety work hypothesizes that robots will not only ignore oversight, but even actively try to deter it. We want to know what these strategies might look like for a real task, how much interaction time it takes to develop them, how they can be recognized, and how generalizable they are across people. To that end, we will purposefully physically intervene in the robot's task in our scenarios in two conditions: stopping the robot, or guiding it to perform the task properly.

Further, we will also explore what *novice* users of the system attempt to do as oversight -- what kind of language instructions they give, what non-verbal information they convey (e.g. gestures), what physical interaction strategies they employ. Even though our naive robot would not initially respond to most of these channels, this data will be really useful because it will reveal what are natural oversight reactions people have. We will then build on these findings to design new, intuitive oversight mechanisms.

3.3 Analyzing and Improving Mitigation Strategies

Besides identifying pathological behaviors of lifelong learning systems and analysing the strategies that human users might use for mitigating such behaviors, we will also analyze and improve computation mitigation tools. To that end, we will study **(a)** uncertainty estimation for reinforcement learning, **(b)** autonomous reward acquisition methods, including inverse RL and other imitation techniques, and **(c)** decision-making algorithms that offer better insight and control into the learned decision-making process, including model-based and explainable control algorithms. These techniques will be rigorously validated on real-world platforms and, in the case of **(b)** and **(c)**, with real human users. Whenever possible, we will evaluate these methods in the context of the lifelong learning systems outlined above.

4 Fit with Priorities

This research meets the priorities of the program. We will soon see autonomous robots in the real world, continuously learning from their experience. The only way we can obtain solid, reliable information about the potential pitfalls of such a deployment is by conducting representative real-world experiments. Therefore, a rigorous laboratory-scale exploration of these questions is important for identifying the most important AI safety problems to address within the next 5 years. Furthermore, we see this work as setting a precedent for future empirical AI safety research, establishing best practices, protocols, and standards: *even if future AI systems differ dramatically and unexpectedly from those we study in the near term, the formulation and conduct of empirical safety experiments will carry on even if the specific findings do not.*

5 Technical Preparedness

PIs Levine and Dragan have the right technical background and combination of expertise to undertake this research. PI Levine's prior work includes the first successful end-to-end trained deep network policy for vision-based robotic manipulation skills [1], the first medium-scale experiment on collective deep robotic learning [2], and the first result on real-world robotic control based on direct video prediction [3]. This prior experience will enable his lab to construct viable self-improving robotic systems for the lifelong learning experiments described in this proposal. PI Levine also has extensive prior work on inverse reinforcement learning [4, 5, 6, 7], as well as uncertainty estimation for deep reinforcement learning [8].

PI Dragan's work combines AI safety research, tackling theoretical problems for the agents of tomorrow, with human-robot interaction research, tackling the practical problems of robots today. Her prior AI safety work includes a technical formulation and equilibrium solutions for value alignment [9], a theoretical analysis of the off-switch problem [10], and an algorithm for autonomously alleviating side effects of misspecified objectives [11]. Her prior HRI work includes algorithms for robot transparency of intentions and objectives [12,13], for accounting for the ways in which robot actions influence human actions [14], and for learning objectives from rich forms of human guidance [15].

References

[1] Levine*, Finn*, Darrell, Abbeel. End-to-end training of deep visuomotor policies. Journal of machine learning research (JMLR) 17:1-40, 2016.

[2] Levine, Pastor, Krizhevsky, Ibarz, Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. International journal of robotics research (IJRR), 2017.

[3] Finn, Levine. Deep visual foresight for planning robot motion. International conference on robotics and automation (ICRA), 2016.

[4] Levine, Popovic, Koltun. Feature construction for inverse reinforcement learning. Neural information processing systems (NIPS), 2010.

[5] Levine, Popovic, Koltun. Nonlinear inverse reinforcement learning with Gaussian processes. Neural information processing systems (NIPS), 2011.

[6] Levine, Koltun. Continuous inverse optimal control with locally optimal examples. International conference on machine learning (ICML), 2012.

[7] Finn, Levine, Abbeel. Guided cost learning: deep inverse optimal control via policy optimization. International conference on machine learning (ICML), 2016.

[8] Kahn, Villaflor, Pong, Abbeel, Levine. Uncertainty-aware reinforcement learning for collision avoidance. Under review. 2017.

[9] Hadfield-Menell, Dragan, Abbeel, Russell. Cooperative inverse reinforcement learning. Neural information processing systems (NIPS). 2016.

[10] Hadfield-Menell, Dragan, Abbeel, Russell. The off-switch game. International Joint Conference on Artificial Intelligence (IJCAI). 2017.

[11] Hadfield-Menell, Abbeel, Russell, Dragan. Inverse reward design. Neural information processing systems (NIPS). 2017.

[12] Dragan, Srinivasa. Generating Legible Motion. Robotics: Science and Systems (RSS). 2013.

[13] Huang, Feld, Abbeel, Dragan. Enabling robots to communicate their objectives. Robotics: Science and Systems (RSS). 2017.

[14] Sadigh, Seshia, Sastry, Dragan. Planning for autonomous cars that leverage effects on human actions. Robotics: Science and Systems (RSS). 2016.

[15] Bajcsy, Losey, O'Malley, Dragan. Learning robot objectives from physical human interaction. Under review. 2017.