Huawei Technologies Research & Development (UK) Limited
Unit 302, Cambridge Science Park
Milton Road
Cambridge, UK
CB4 0WG
Tel: +44 (0) 1223 887000
Fax: +44 (0) 1223 887288
www.huawei.com

04 August 2021

Title: Safe and Robust Autonomous Decision Making with AI

Sequential decision making (SDM) under uncertainty is a crucial component of many autonomous systems. Applications exhibiting those characteristics are ubiquitous, including but not limited to gaming environments, self-driving, robotics, personalized medicine and financial trading. In recent years, significant progress has been made on developing algorithms that solve large-scale SDM tasks leading to remarkable outcomes. For instance, techniques developed in reinforcement learning and Bayesian optimization -- machine learning methods for SDM -- have been at the heart of alpha-Go and alpha-Go zero, gaming solvers capable of achieving human-level performance.

Inspired by those innovations, many researchers are attempting their application in the real world beyond gaming environments, e.g., in chip design [1], dopamine impulse control [2], self-driving [3], and robotics [4]. Although successful in isolated instances, multiple concerns related to safety and robustness arise when carrying reinforcement learning beyond well-behaved laboratory settings.

Our concerns stem from recent results [5], [6] assessing the robustness and safety of reinforcement learning algorithms:

1. Robustness to changes in transition dynamics is a crucial component for adaptive and safe reinforcement learning in the real world. To illustrate, consider a self-driving car scenario in which we attempt to design an agent capable of driving a vehicle smoothly, safely, and autonomously. A typical reinforcement learning workflow to solving such a problem consists of building a simulator to emulate real-world scenarios, training in simulation, and then transferring resultant policies to physical systems for control. Of course, building exact digital twins of traffic behavior and driving dynamics in large cities is a formidable challenge in and of itself. Henceforth, for the above strategy to succeed, we require controllers learnt in simulation to exhibit robustness to changes in environmental dynamics to cover a broad range of possible scenarios. To assess those properties [5], we evaluate state-of-the-art methods on standard reinforcement learning benchmarks from OpenAI gym and Mujoco that involve the control of simulated robots. We train various algorithms in one environment and test performance when varying dynamic parameters. Results indicate that current methods are highly fragile, losing significant performance upon minor environmental variations. We then propose algorithms that trade-off performance versus robustness, showing broader ranges of control. Our methods define constrained versions of the problem and operate in minimax scenarios. Here, we determine controllers that avoid overfitting on one specific task but rather generalise across scenarios.

2. Reinforcement learning methods assume idealized simulators and require tens of millions of agent-environment interactions gathered by randomly exploring policies. In real-world safety-critical applications, however, such an idealized framework of random exploration with the

Huawei Technologies Research
& Development (UK) Limited
Unit 302, Cambridge Science Park
Milton Road
Cambridge, UK
CB4 0WG
Tel: +44 (0) 1223 887000
Fax: +44 (0) 1223 887288
www.huawei.com

ability to gather samples at ease falls short, partly due to the catastrophic costs of failure and the high operating costs. Hence, if algorithms are to be applied in the real world, safe agents that are sample-efficient and capable of mitigating risk need to be developed. To this end, different works adopt varying safety definitions, where some are interested in safe learning, i.e., safety during the learning process, while others focus on acquiring safe policies eventually. Unfortunately, most of these methods are sample-inefficient and make a large number of visits to unsafe regions. Our interest in algorithms achieving safe final policies while reducing the number of visits to unsafe regions makes us pursue a new framework and demonstrate effective performance in various benchmarks [6]. We formalize a chance constraint constrained problem and combine reinforcement learning with active learning to reduce visits to unsafe regions.

We wish to present those concerns and propose potential solutions to debate the viability of deploying reinforcement learning technologies in the real world. We hope our findings stimulate researchers and practitioners in the field to better assess and test their algorithms before deployment. We will end our paper with a call for open problems that can serve as an initial stepping-stone into a dedicated collaborative research curriculum to enable a new generation of intelligent and adaptive algorithms.

1. (https://analyticsindiamag.com/how-reinforcement-learning-is-advancing-chip-designing/)
2. (https://www.sciencetimes.com/articles/32446/20210724/mice-control-dopamine-impulses-reward-study-shows.htm)
3. (https://arxiv.org/pdf/2002.00444.pdf)
4. (https://techcrunch.com/2021/07/27/cassie-the-bipedal-robot-runs-a-5k/)
5. (https://arxiv.org/abs/1907.13196)
6. (https://arxiv.org/abs/2006.09436)