Interpretable Adversarial Reinforcement Learning

Oliver Chang

University of California, Santa Cruz
elochang@ucsc.edu

Abstract

Autonomous driving has shown significant progress in recent years. The combination of advanced sensors, ample data, and machine learning algorithms has led to the deployment of autonomous vehicles (AVs) in cities like Los Angeles, San Francisco, and Phoenix. While not all humans can drive perfectly, AVs should be able to plan, adapt, and react to environmental disturbances, including irrational human drivers. My research focuses on applying reinforcement learning (RL) techniques to validate AV-related cyber-physical systems (CPS) in realistic environments. I develop a custom RL environment that simulates highway driving scenarios with multiple vehicles. This environment includes a CPS model of adaptive cruise control (ACC), a lane-changing model (MOBIL), and an adversarial agent that learns to drive irrationally. My work extends interpretable RL techniques to continuous control tasks like autonomous driving.

Introduction

Despite the rise in AVs, there are still challenges to overcome (Di Lillo et al. 2024). In particular, human drivers can make irrational decisions, and AVs must be able to handle these situations safely. The California DMV keeps a record of all AV-related accidents, which provides a valuable resource for understanding the challenges faced by AVs in real-world scenarios (Favarò, Eurich, and Nader 2018). Accidents involving AVs and human drivers usually occur when the human driver strikes the AV from behind, implying that humans may misjudge the AV's behavior or fail to notice in time (Goodall 2021).

The goal of an adversarial agent is to cause a CPS to violate a safety specification. This can identify scenarios that lead to failure points in CPSs. If discovering failure points in CPS answers the question of "What causes a system to fail?", then the next question is "Why does the system fail?" To answer this question, I use interpretable RL techniques which allow me to validate learned behaviors, identify model errors, and to ensure continuous improvement in models (Gilpin et al. 2018). Prior work in interpretable RL has been used in video games, operating in few degrees of freedom for simplicity (Campbell et al. 2023; Guo et al. 2023).

Reinforcement Learning for Falsifying Autonomous Driving Systems

While previous research has analyzed historical accident data (Liang et al. 2024), used localized sampling-based falsification (Hernandez et al. 2024), and motion-planning techniques (Orthey, Chamzas, and Kavraki 2023), my approach leverages RL as the falsification tool to discover adversarial driving maneuvers. RL is well-suited for this problem because it can learn emergent behaviors in complex environments. Thus, no prior knowledge, routes, or maneuvers are needed to reduce any bias in the search for failure points.

The environment I develop uses Scenic (Fremont et al. 2019). Scenic, a probabilistic programming language, allows for intuitive scenario orchestration. Scenic's flagship domain is driving. We use MetaDrive as our simulator that Scenic interfaces with. Scenic's role is to generate dynamic scenario compositions such as specifying vehicle placement. Each non-RL vehicle is controlled by ACC and MOBIL. Prior work in ACC falsification investigated single-lane accidents (Hernandez et al. 2024), but my work extends this to multi-lane scenarios, by incorporating MOBIL as the lane-changing model.

The RL agent is trained using Soft Actor-Critic (SAC). SAC is ideal for this task because the joint objective function incentivizes both exploration and exploitation (Haarnoja et al. 2018). The reward function is designed to encourage the adversarial agent to cause a collision while maintaining a safe distance. Therefore, the reward function is sparse, offering a large positive value for causing a collision and a dense signal for driving forward and fast to discover unique trajectories. Experiments reveal that the RL agent can successfully learn an adversarial driving policy that thwarts the CPS's safety constraints. Moreover, the RL agent can generalize to unseen traffic configurations, by perturbing the initial vehicle placements through Scenic.

I am collaborating with my colleague, Kay Vargas, who is developing a novel suite of metrics to evaluate the quality and quantity of counter examples. These metrics will help us understand the effectiveness of our RL-based falsification approach compared to traditional methods. Our current progress demonstrates successfully counter example discovery through RL. We plan to submit our work to AAMAS 2026. Thus, we anticipate our research being peer-reviewed

by the workshop date (January 20-21, 2026). This research is funded in part by the National Center for Transportation Cybersecurity and Resiliency.

Interpretable Reinforcement Learning

Although a RL agent can solve complex tasks, a robust agent requires millions of trial-and-error interactions with the environment in order to learn an optimal policy. For practicality, RL agents can reuse prior knowledge to learn more efficiently. Fine-tuning RL agents in novel tasks is a common practice called transfer learning (Campbell et al. 2023).

An interpretable transfer learning paradigm, action-advising, presents knowledge from a teacher agent to a student agent. Instead of fine-tuning neural network weights, action-advising transfers knowledge through action suggestions, where the teacher provides the student with the optimal action for a given state. However, prior work in action-advising requires an optimal teacher (which is not always available), applies in discrete action spaces, and fails when the student is solving a task that is too different from the teacher's task.

My work addresses these limitations by introducing a self-modulating action-advising algorithm, Dynamically Introspective Action Advising (DIAA). Based on the Introspective Action Advising (IAA), my algorithm dynamically adjusts the threshold for when a teacher should give advice to the student. We adjust the amount of advice given by comparing the relative performance of the student with advice and student without advice by leveraging the off-policy objective difference lemma (Kakade and Langford 2002; Shenfeld et al. 2023). Intuitively, if the student with advice is performing better, then the teacher should give more advice. Conversely, if the student without advice is performing better, then the teacher should give less advice, i.e., follow the student's intuition.

Since we use the off-policy objective difference lemma, DIAA extends IAA to continuous control tasks by using SAC. SAC features a replay buffer which stores previously collected experiences. We store trajectories from the student policies and sample experiences from the buffer uniformly for a fair comparison. By using SAC, we also extend the slate of experiments, accelerating training in tasks like autonomous driving (MetaDrive), robotic control, and locomotion (MuJoCo). Our results indicate a sizeable speedup in learning, reducing the overall number of environment interactions needed to learn an optimal policy. As importantly, DIAA reveals which teacher behaviors are most beneficial for the student. In our driving experiments, we found that a teacher trained on a straight road with no traffic struggled with throttle control when the student was driving on a road with merging traffic and curves. Thus, the student learned to be more cautious and drive slower.

I am working closely with my Ph.D. advisor, Professor Leilani Gilpin, to develop, evaluate, and deploy DIAA. We are planning to submit our work to AAMAS 2026. Hence, we anticipate our research being peer-reviewed by the workshop date (January 20-21, 2026).

References

Campbell, J.; Guo, Y.; Xie, F.; Stepputtis, S.; and Sycara, K. 2023. Introspective action advising for interpretable transfer learning. In *Conference on Lifelong Learning Agents*, 1072–1090. PMLR.

Di Lillo, L.; Gode, T.; Zhou, X.; Atzei, M.; Chen, R.; and Victor, T. 2024. Comparative safety performance of autonomous-and human drivers: A real-world case study of the Waymo Driver. *Heliyon*, 10(14).

Favarò, F.; Eurich, S.; and Nader, N. 2018. Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations. *Accident Analysis & Prevention*, 110: 136–148.

Fremont, D. J.; Dreossi, T.; Ghosh, S.; Yue, X.; Sangiovanni-Vincentelli, A. L.; and Seshia, S. A. 2019. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 63–78. Phoenix AZ USA: ACM. ISBN 978-1-4503-6712-7.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Goodall, N. J. 2021. Comparison of automated vehicle struck-from-behind crash rates with national rates using naturalistic data. *Accident Analysis & Prevention*, 154: 106056.

Guo, Y.; Campbell, J.; Stepputtis, S.; Li, R.; Hughes, D.; Fang, F.; and Sycara, K. 2023. Explainable action advising for multi-agent reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5515–5521. IEEE.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. Pmlr.

Hernandez, C.; Barbosa, D. O.; Lei, Z.; Burbano, L.; Park, Y.; Ukkusuri, S.; and A. Cardenas, A. 2024. D4: Dynamic Data-Driven Discovery of Adversarial Vehicle Maneuvers. In *International Conference on Dynamic Data Driven Applications Systems*, 182–190. Springer.

Kakade, S.; and Langford, J. 2002. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, 267–274. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-873-3.

Liang, C.; Ghazel, M.; Ci, Y.; and Zheng, W. 2024. Analyzing rear-end collision risk relevant to autonomous vehicles by using a humanlike brake model. *Journal of Transportation Engineering, Part A: Systems*, 150(7): 04024031.

Orthey, A.; Chamzas, C.; and Kavraki, L. E. 2023. Sampling-based motion planning: A comparative review. *Annual Review of Control, Robotics, and Autonomous Systems*, 7.

Shenfeld, I.; Hong, Z.-W.; Tamar, A.; and Agrawal, P. 2023. TGRL: An algorithm for teacher guided reinforcement learning. In *International Conference on Machine Learning*, 31077–31093. PMLR.