# Adversarial Robustness of AI Agents Acting in Probabilistic Environments

Lisa Oakley, Alina Oprea, Stavros Tripakis
Khoury College of Computer Sciences, Northeastern University

***Abstract*—As machine learning systems become more pervasive in safety-critical tasks, it is important to carefully analyze their robustness against attack. Our work focuses on developing an extensible framework for verifying adversarial robustness in machine learning systems over time, leveraging existing methods from probabilistic model checking and optimization. We present preliminary progress and consider future directions for verifying several key properties against sophisticated, dynamic attackers.**

***Index Terms*—MDPs, machine learning, adversarial robustness**

Adversarial machine learning is a well-developed area of research analyzing the safety and robustness of machine learning algorithms. A significant amount of research in adversarial machine learning addresses robustness of neural networks trained for image classification tasks. In this research area, one attack is to synthesize small perturbations of the network inputs to craft "adversarial examples" which result in violations to some desired properties [1]–[3]. In this approach, attacks seek to falsify the robustness property of the neural networks. For example, adding adversarially selected noise to an image of a panda bear can cause a convolutional neural network to erroneously label the image "gibbon" with high confidence, while to a human the changes are indistinguishable [4].

In recent years, tools such as Reluplex [5], ReluVal [6], $AI^2$ [7], and Charon [8] attempt to verify robustness in neural networks. In these works, the authors go beyond finding counterexamples to robustness, and attempt to verify the correctness of the algorithms using techniques including abstract interpretation and projected gradient descent.

However, it is often the case that neural networks and other machine learning technologies exist in larger systems, engaging with uncertain environments over long time durations. A natural extension to the rich area of adversarial machine learning is to evaluate the adversarial robustness of AI agents acting over time in probabilistic environments.

Some preliminary work has been done to address this problem. Dreossi et. al [9] use a falsification approach to determine whether adversarial perturbations in a machine learning component of a cyber-physical system can cause a failure in the entire system. Verily [10] takes a verification approach, and leverages existing DNN verifiers to verify safety and liveness properties in a deep reinforcement learning algorithm. Verisig [11] verifies safety properties in closed-loop systems for hybrid systems with sigmoid-based neural network components. Bacci and Parker [12] use model abstraction to provide robustness guarantees for the controller of a deep RL agent. Suilen et. al. [13] use convex optimization to synthesize robust policies which interact with uncertain, partially observable environments.

We propose the use of formal verification methods for probabilistic programs as a framework for ensuring provable adversarial robustness in AI policies acting in probabilistic environments. In this approach, we model the environment as a Markov decision process (MDP), and consider an adversary that can modify the transition probabilities in the environment. The adversary's goal is to find minimal environmental modifications which result in a violation of some previously satisfied property. Some properties of interest include reachability, safety, and expected reward.

In the initial stages of this work, we have reduced selected instances of this problem to quadratic programming. We also developed a tool in python which successfully minimizes the reachability property of any policy over two time steps with bounded perturbations on randomized MDPs. Our tool is easily extensible to more complex MDPs, and we have plans to test our solution on Grid-world [14] environments with large state space. We also plan to explore novel adversarial models for MDPs, as well as additional security properties that can be probabilistically verified in our framework.

Exploring robustness of MDPs has many benefits. Firstly, MDPs and Markov chains are widely used for modeling agent interactions with probabilistic systems with cybersecurity applications such as virus infections in a network [15] and advanced persistent threats (APTs) [16]. Therefore, we have a significant body of real-world applications in which to test our method. Furthermore, using environments described as MDPs allows us to leverage the rich field of probabilistic model checking to implement efficient and provably correct algorithms using state of the art model checking software such as PRISM [17]. Additionally, we can generalize the verifications to many types of agents, including applications to reinforcement learning and other synthesized policies for MDPs and POMDPs [18].

As we continue this research, we plan to make improvements to our tool to address longer time durations and more properties. Additionally, we intend to consider more sophisticated threat models including dynamic attacks which can adapt over time. We believe that leveraging verification methods for probabilistic programs to address adversarial robustness in policies for probabilistic systems is a novel and promising research direction. We intend to continue working to address these challenges and move toward a future of provably robust artificial intelligence.

**Acknowledgements.** This work has been supported by the National Science Foundation under NSF SaTC awards CNS-1717634 and CNS-1801546.

## REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (S & P)*. IEEE, 2017, pp. 39–57.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.

[5] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in *International Conference on Computer Aided Verification*. Springer, 2017, pp. 97–117.

[6] S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana, "Formal security analysis of neural networks using symbolic intervals," in *27th USENIX Security Symposium*, 2018, pp. 1599–1614.

[7] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, "Ai2: Safety and robustness certification of neural networks with abstract interpretation," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 3–18.

[8] G. Anderson, S. Pailoor, I. Dillig, and S. Chaudhuri, "Optimization and abstraction: A synergistic approach for analyzing neural network robustness," in *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 731744. [Online]. Available: https://doi.org/10.1145/3314221.3314614

[9] T. Dreossi, A. Donzé, and S. A. Seshia, "Compositional falsification of cyber-physical systems with machine learning components," *Journal of Automated Reasoning*, vol. 63, no. 4, pp. 1031–1053, 2019.

[10] Y. Kazak, C. Barrett, G. Katz, and M. Schapira, "Verifying deep-RL-driven systems," in *Proceedings of the 2019 Workshop on Network Meets AI & ML*, ser. NetAI19. New York, NY, USA: Association for Computing Machinery, 2019, p. 8389. [Online]. Available: https://doi.org/10.1145/3341216.3342218

[11] R. Ivanov, J. Weimer, R. Alur, G. J. Pappas, and I. Lee, "Verisig: Verifying safety properties of hybrid systems with neural network controllers," in *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, ser. HSCC 19. New York, NY, USA: Association for Computing Machinery, 2019, p. 169178. [Online]. Available: https://doi.org/10.1145/3302504.3311806

[12] E. Bacci and D. Parker, "Probabilistic guarantees for safe deep reinforcement learning," *arXiv preprint arXiv:2005.07073*, 2020.

[13] M. Suilen, N. Jansen, M. Cubuktepe, and U. Topcu, "Robust policy synthesis for uncertain POMDPs via convex optimization," *arXiv preprint arXiv:2001.08174*, 2020.

[14] L. Kaebling, M. Littman, and A. Cassandra, "Learning policies for partially observable environments: Scaling up," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[15] M. Kwiatkowska, G. Norman, D. Parker, and M. Vigliotti, "Probabilistic mobile ambients," *Theoretical Computer Science*, vol. 410, no. 12–13, pp. 1272–1303, 2009.

[16] L. Oakley and A. Oprea, "QFlip: An adaptive reinforcement learning strategy for the FlipIt security game," in *International Conference on Decision and Game Theory for Security*. Springer, 2019, pp. 364–384.

[17] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in *Proc. 23rd International Conference on Computer Aided Verification (CAV'11)*, ser. LNCS, G. Gopalakrishnan and S. Qadeer, Eds., vol. 6806. Springer, 2011, pp. 585–591.

[18] S. Carr, N. Jansen, R. Wimmer, A. Serban, B. Becker, and U. Topcu, "Counterexample-guided strategy improvement for pomdps using recurrent neural networks," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 5532–5539.