Verification of Safety in Artificial Intelligence and Reinforcement Learning Systems

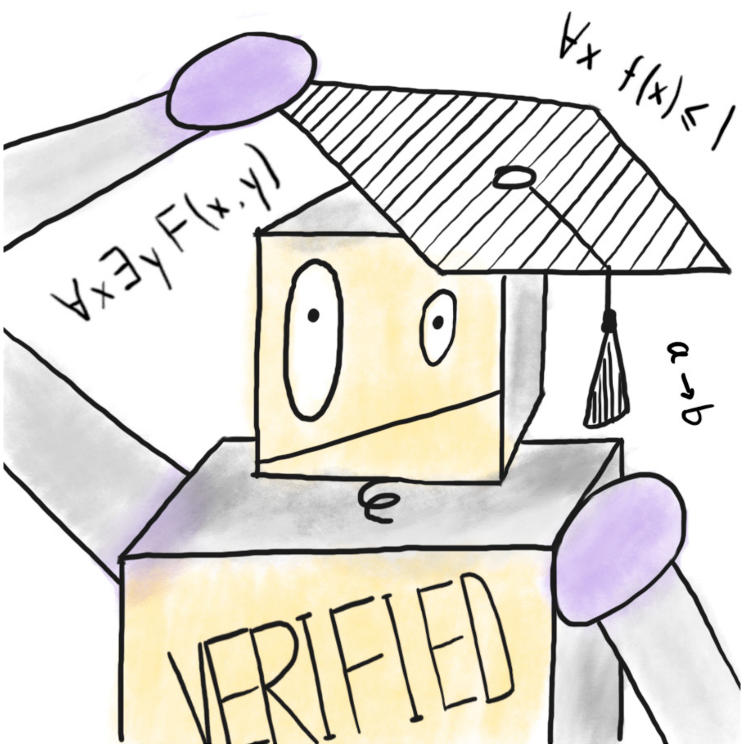
Yanni A. Kouskoulas, Daniel I. Genin, Aurora C. Schmidt, Ivan I. Papusha, Rosa Wu, Galen E. Mullins, Tyler A. Young, and Joshua T. Brulé

ABSTRACT

For complex artificially intelligent systems to be incorporated into applications where safety is critical, the systems must be safe and reliable. This article describes work a Johns Hopkins University Applied Physics Laboratory (APL) team is doing toward verifying safety in artificial intelligence and reinforcement learning systems.

Broad groups of researchers at APL are studying and developing the next generation of autonomous systems. Advances in machine learning and artificial intelligence (AI) enable the autonomous operation of ground vehicles, planes, drones, submarines, and much more. However, to successfully incorporate such complex AI systems into military and safety-critical applications, we must advance our means for ensuring their safe and reliable operation.

The problem is that many of these AI systems learn by optimizing a reward function. The unconstrained maximization of statistical rewards leads to a variety of issues such as reward hacking, unintended consequences, and, for continually-learning systems, catastrophic forgetting. In safety-critical systems, we must be able to guarantee or verify that the system will behave according to the expectations of users as well as others who could be affected by the system. Much attention is being paid to the existence of, and security



vulnerabilities posed by, adversarial examples—one consequence of not being able to verify the performance of a machine learning system. However, there are ample examples of unexpected and tragic consequences where autonomous systems have resulted in loss of life without malicious manipulation.[1–3]

To develop a means for guaranteeing the safe operation of AI-enabled systems, APL researchers have been using formal methods. *Formal methods* describe a wide array of tools and techniques that encompass the formal definition of logical requirements and system descriptions. Tools from formal methods, such as those built on satisfiability modulo theory (SMT), can be used to prove that a given formally described system satisfies a set of desired properties and constraints.[4]

A 2019 independent research and development project, called A ModelPlex Approach to a Verified Robotics Code Kit (MAVeRiCK), extended research for verifying aircraft collision avoidance to create a correct-by-construction fallback controller design that ensures collision-free path planning.[5] The fallback controller ensures safety by taking over from the primary controller whenever a critical state is reached and a particular action must be taken to avoid an imminent collision. This project resulted in a research paper detailing how formally verified safety predicates are used to create a fallback controller with safety guarantees.[6] Furthermore, the work contributed to a library for formal verification of timing computations for turn to bearing maneuvers.[7] This approach to verifying the safety of vehicle navigation is being applied in a larger Air Force Research Laboratory (AFRL)-funded effort for the subtask of creating a verified runtime assurance watchdog controller to ensure the safe testing of autonomous aircraft systems; for example, through guaranteeing the watchdog predictively enforce a vehicles stays within the planned test range geofence.[8]

However, the team recognized that in a number of situations the fallback control architecture may lead to problematic performance of mission objectives. Imagine cases in which the fallback and primary controller interfere with each other so that progress toward the goal is impeded. As a result, APL began an effort called Verified Safe Reinforcement Learning (VSRL) to study alternative approaches for ensuring the safe performance of continually adaptive deep learning systems. The project sought to provide direct guarantees on the performance of a neural network controller trained to avoid collisions with other aircraft while minimizing deviations from the goal. In general, providing guarantees on the outputs of neural networks over the continuous space of potential inputs is too difficult, due to the complex mapping from input to outputs that neural networks embody. The team found promise in a methodology for the encoding of affine networks that may be verified with SMT tools.[9] In 2020, VSRL aimed to demonstrate this approach in the training and verification of small rectified linear unit networks trained to perform path planning or control tasks; creating a library for automatic encoding of pytorch networks into SMT constraints.[10] In addition, the team discovered a method for adjusting the weights of a neural network using an SMT solver to guarantee certain input-output relationships. The approach was published in the 2020 Formal Methods for ML-Enabled Autonomous Systems (FoMLAS) workshop.[11] The effort culminated in a demonstration of the verification of a reinforcement learning network trained to avoid aircraft collisions[12] through the use of safeability concepts to reduce the domain of inputs that must be checked to verify the safety of a neural network controller. This research will enable the design of future systems that can take advantage of machine learning advances as well as formal approaches to guaranteeing safe performance.

REFERENCES

1. R. Gonzales, "Feds say self-driving Uber SUV did not recognize jaywalking pedestrian in fatal crash," *NPR*, Nov. 7, 2019, https://www.npr.org/2019/11/07/777438412/feds-say-self-driving-uber-suv-did-not-recognize-jaywalking-pedestrian-in-fatal-.
2. J. Evans, "Driverless cars: Kangaroos throwing off animal detection software," *ABC News*, Jun. 23, 2017, https://www.abc.net.au/news/2017-06-24/driverless-cars-in-australia-face-challenge-of-roo-problem/8574816.
3. D. Shepardson, "U.S. agency probing two fatal Tesla crashes in Florida since last Sunday," *Reuters*, Mar. 2, 2019, https://www.reuters.com/article/us-tesla-crash/u-s-agency-probing-two-fatal-tesla-crashes-in-florida-since-last-sunday-idUSKCN1QJ0MC
4. C. Barrett, R. Sebastiani, S. Seshia, and C. Tinelli, "Satisfiability modulo theories" in *Handbook of Satisfiability*, vol. 185 of *Frontiers in Artificial Intelligence and Applications*, A. Biere, M. J. H. Heule, H. van Maaren, and T. Walsh, Eds., Amsterdam: IOS Press, Feb. 2009, pp. 825–885.
5. J. Jeannin, K. Ghorbal, Y. Kouskoulas, A. Schmidt, R. Gardner, et al., "A formally verified hybrid system for safe advisories in the next-generation airborne collision avoidance system," *Int. J. Softw. Tools Technol. Transfer.*, vol. 19, pp. 717–741, 2017, https://doi.org/10.1007/s10009-016-0434-1.
6. Y. Kouskoulas, A. Schmidt, J.-B. Jeannin, D. Genin, and J. Lopez, "Provably safe controller synthesis using safety proofs as building blocks," in *Proc. IEEE 7th Int. Conf. in Softw. Eng. Res. and Innov. (CONISOFT '19)*, Oct. 2019, Mexico City, Mexico.
7. Y. Kouskoulas, T. J. Machado, and D. Genin, "Formally verified timing computation for non-deterministic horizontal turns during aircraft collision avoidance maneuvers," in *Formal Methods for Industrial Critical Systems, FMICS 2020, Lecture Notes in Computer Science*, vol. 12327, M. ter Beek and D. Ničković, Eds., Cham: Springer, 2020, https://doi.org/10.1007/978-3-030-58298-2_4.
8. Y. Kouskoulas, R. Wu, J. Brulé, D. Genin, A. Schmidt, and A. Machado, "Good fences make good neighbors: Designing a formally verified predictive geofence," *NASA Formal Methods Conf. (NFM 2021)*, to be published.
9. I. Papusha, U. Topcu, S. Carr, and N. Lauffer, "Affine multiplexing networks: system analysis, learning, and computation," 2018, https://arxiv.org/abs/1805.00164.
10. Lantern-SMT python repository, https://github.com/JHUAPL/lantern-smt.
11. I. Papusha, R. Wu, J. Brulé, Y. Kouskoulas, D. Genin, and A. Schmidt, "Incorrect by construction: Fine tuning neural networks for guaranteed performance on finite sets of examples," in *3rd Workshop on Formal Methods for ML-Enabled Autonomous Syst. (FoMLAS 2020)*, Jul. 2020, https://arxiv.org/abs/2008.01204.
12. D. Genin, I. Papusha, J. Brulé, T. Young, G. Mullins, Y. Kouskoulas, R. Wu, and A. Schmidt, "Formal verification of neural network controllers for collision-free flight," submitted for publication.

Yanni A. Kouskoulas, Affirm Logic, McLean, VA

Yanni A. Kouskoulas is currently working on secure software systems at Affirm Logic Corp. Prior to joining Affirm Logic, he was a member of the Principal Professional Staff and chief scientist for the Enterprise Systems Group in APL's Asymmetric Operations Sector. He has a BS, an MS, and a PhD in electrical engineering from the University of Michigan and an MBA from the University of Maryland, College Park. He currently serves as the technical and thought leader in his group, planning, leading, and executing a variety of research and development efforts. Yanni's research interests include the application of rigorous, mathematical proofs (i.e., formal methods and formal verification) to develop techniques to aid in producing zero-defect software components and control algorithms. His email address is yxkous@gmail.com.

Daniel I. Genin, Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Daniel I. Genin is a mathematician and computer scientist in APL's Asymmetric Operations Sector specializing in application of formal methods in software design and verification. He has a BS in mathematics and a PhD in mathematics/high-performance computing from Pennsylvania State University. Daniel has led and contributed to a number of projects applying a range of formal methods techniques to mission critical systems. His current interests lie in the area safety and correctness verification for cyber-physical systems. His email address is daniel.genin@jhuapl.edu.

Aurora C. Schmidt, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

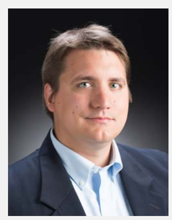
Aurora C. Schmidt is a project manager in APL's Research and Exploratory Development Department. She has a BS and an MS in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT) and a PhD in electrical and computer engineering from Carnegie Mellon University. Aurora's research interests include sensor networks, estimation and coordination problems, signal processing, compressed sensing, optimization, multi-target tracking, control theory, and information and decision-making. Her email address is aurora.schmidt@jhuapl.edu.

Ivan I. Papusha, Space Exploration Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Ivan I. Papusha is a Senior Professional Staff member in APL's Space Exploration Sector. He has a BS and an MS in electrical engineering from Stanford University and a PhD in control and dynamical systems from the California Institute of Technology. His email address is ivan.papusha@jhuapl.edu.

Rosa Wu, Defense Nuclear Facilities Safety Board, Washington, DC

Rosa Wu is an engineer with the Defense Nuclear Facilities Safety Board. She was selected for the agency's Professional Development Program where she got the opportunity for a one-year fellowship with the Johns Hopkins University Applied Physics Laboratory in the Research and Exploratory Development Department. Rosa earned a BS in chemical engineering from the University of Illinois at Urbana-Champaign and a Masters of Engineering in chemical engineering with a specialty in computational informatics from Cornell University. While at APL, Rosa developed skills in machine learning, formal methods, and atmospheric dispersion modeling. As part of the Verified Safe Reinforcement Learning team, she worked on training small linear rectified unit networks and verifying them with satisfiability modulo theories tools. Her email address is rosawu2@gmail.com.

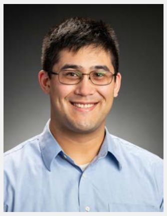


Galen E. Mullins, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Galen E. Mullins is a roboticist in APL's Research and Development Department. He received bachelor's degrees in mechanical engineering and mathematics from Carnegie Mellon University, a master's degree in applied physics from Johns Hopkins University, and a doctorate in mechanical engineering from University of Maryland. His research interests are in robotics, autonomy, machine learning, and numerical optimization. His primary duties include creating thermal analysis tools and generating infrared signature data to support several programs. His background also includes electrical and mechanical design, software development, and system automation. His e-mail address is galen.mullins@jhuapl.edu.

Tyler A. Young, Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Tyler A. Young is a software engineer in APL's Asymmetric Operations Sector. He has a BS in computer engineering and an MS in software engineering, both from Villanova University. As a graduate researcher, he worked in robotics simulation development as the lead developer of the "RAMS" classroom robotics simulator, adapted from NASA's Rover Analysis and Modeling Simulation (ROAMS). He now contributes to the avionics safety sector, developing and testing the ACAS-X family of collision avoidance software. Tyler's current research interests include airspace simulation, AI safety, and cybersecurity. His email address is tyler.young@jhuapl.edu.



Joshua T. Brulé, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Joshua T. Brulé is a member of the Senior Professional Staff in APL's Research and Exploratory Development Department. His background is in causal programming (roughly, the intersection of causal reasoning and inference and programming languages). His research interests include probability theory, statistics, artificial intelligence, functional programming, and explorable explanations. His email address is joshua.brule@jhuapl.edu.