Advancing Learning-based Autonomy with Formal Methods

Hanna Krasowski, UC Berkeley

I. Introduction

Automating real-world systems is challenging due to dynamic environments, perception uncertainties, and model disturbances. Because of this complexity, machine learning (ML) is often considered the most suitable paradigm for real-world system autonomy. However, purely ML-based models lack explainablity, safety guarantees, and depend on large high-quality training datasets. In contrast, model-based methods provide guarantees and explainability for autonomous systems, but often require substantial engineering and expert knowledge to achieve performant models.

In my research, I aim to unlock the potential of ML-based techniques for real-world systems by leveraging formal methods to achieve data-efficient, safe, and interpretable autonomy (see Fig. 1). Advancing ML with formal methods offers two key advantages: 1) Formal methods enhance interpretability and can provide hard safety guarantees. 2) Encoding abstract knowledge with formal methods makes it computationally tractable and can guide even in environments with low-quality or limited data.

Past and Current Work. The first thrust of my research develops algorithms for reliable machine learning. In particular, I focus on *provably* safe reinforcement learning (RL) algorithms that provide hard guarantees for complex safety specifications during training *and* deployment [19, 24, 26]. I demonstrate the capability of my algorithms in realistic dynamic environments such as navigation of unmanned surface vessels [23]. My second thrust uses formal methods, i.e., set-based reachability analysis and temporal logic, to make abstract domain knowledge computationally tractable [33]. Here, I formalize complex specifications for rule-compliant motion planning [21, 25], and, more recently, specifications that are translating qualitative observations in systems biology [27]. My third thrust aims to facilitate realistic motion planning research and increase comparability. To this end, I introduced CommonOcean, a dedicated software and benchmarking platform for maritime vessel navigation [21, 22, 28].

Research Vision. Ultimately, I envision that advancing ML with formal methods will enable the development of a foundational framework for real-world autonomy capable to tap unstructured and multi-modal information sources, including implicit and mathematical representations of system models as well as time-series, text, or traffic data. This foundational framework would feature a modular architecture to account for different abstraction levels and context-adaptablity in order to achieve robust and transparent ML.

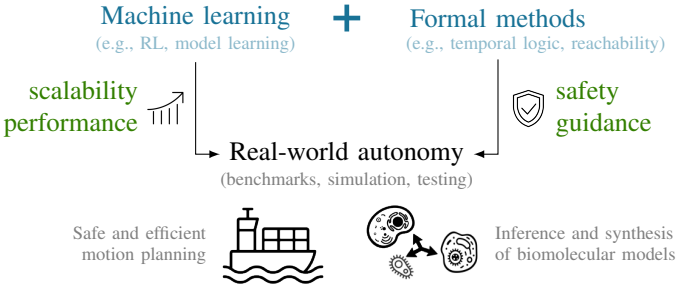

Fig. 1. I develop and integrate formal methods that guide machine learning to obtain effective, data-efficient, safe, and interpretable autonomous systems.

II. Past and Current Work

Provably safe reinforcement learning. For real-world autonomy, regulatory bodies and end-users demand explainable and verifiable systems. RL is capable of solving the complex motion planning tasks required for operation of autonomous systems. Yet, vanilla RL is inherently unsafe due to random exploration and unpredictable behavior in out-of-distribution states. Thus, there is a push for safe RL [17], which typically considers safety specifications softly through the reward function [6] or constraint optimization [1]. Such safe RL approaches achieve safety *only* in expectation. However, for many real-world applications a single safety violation can be disastrous, e.g., a forceful collision of a robot arm with a person could severely injure the person.

To address this gap, I developed *provably* safe RL algorithms [19, 23, 24, 26], which provide hard guarantees for RL agents during training *and* deployment by combining formal methods with RL. Formal methods are rigorous mathematical methods that either synthesize specification-compliant controllers or verify specifications for systems. Specifically, I employ set-based reachability analysis [3] for verifying actions of RL agents. Set-based reachability computes compact sets that enclose all possible system behaviors for a set of inputs and states. Due to the polynomial runtime complexity and continuous-domain computations of set-based reachability analysis, this method scales to high-dimensional state spaces and seamlessly integrates bounded uncertainty and dynamic obstacles [3], which is challenging for alternative verification methods such as model-checking [2], control barrier functions [7, 30], and Hamilton-Jacobi-Isaac reachability analysis [15, 34]. My work addresses two key barriers to the adoption of provably safe RL in the real world: high customization and simplistic safety specifications. In [19], I proposed a

general and real-time-capable algorithm for provable collision avoidance of autonomous systems operating in dynamic environments. In [23], I developed the first provably safe RL algorithm that guarantees compliance with complex traffic rules for multi-agent maritime navigation.

Codifying abstract knowledge with formal methods. For real-world systems, data is multi-modal and sparse or costly to generate or collect. In biology, for example, large-scale genetic and protein data is available, but certain steps in diagnostics and drug discovery still lack adequate models [12], e.g., in situ measurement often remains unfeasible or is prohibitively costly. In such circumstances, it is paramount to tap all available information sources, including abstract domain knowledge such as system dynamics and documented heuristics. In the growing field of neuro-symbolic AI [4, 6, 29], my research combines formal methods with ML to guide learning when available knowledge is abstract.

In particular, I employ temporal logic and reachability analysis to re-shape abstract knowledge into a computationally tractable representation that guides ML for motion planning and system biology [25–27, 33]. For instance, I demonstrated that translating qualitative observations on biomolecular systems to temporal logic enables identifying realistic candidates of biomolecular models, even in the absence of quantitative data [27]. Similarly, integrating abstract knowledge enhances real-world autonomous systems that operate in high-dimensional continuous action spaces. Vanilla RL often struggles in these spaces [31, 32], especially when relevant actions are highly state-dependent and represent only a small subset of the global action space. I developed the first action masking approaches for *continuous* spaces, which use *state-dependent* action sets derived from task knowledge to focus RL exploration on the relevant solution subspace [25, 26, 33]. My experiments on robotics benchmarks in [33] showed this focus leads to faster convergence and can even be necessary to enable learning.

Evaluating real-world readiness. Evaluation of autonomous systems is difficult [8, 11] since realistic open-source simulation environments [10, 20] are challenging to develop and open-access hardware platforms [14, 36] are usually limited to simple tasks. Unmanned surface vessels are a suitable real-world application for autonomous systems research as they feature partially-observable multi-agent and multi-objective tasks with complex environmental disturbances and sparse multi-modal data. Solving this challenging application would reduce ecological, economic, and health risks. In 2023 alone, severe maritime collisions occurred every four days, primarily due to human error [13], resulting in injuries or environmental damage. Intriguingly, autonomous vessels can even have a positive impact on the environment, such as by cultivating seaweed for carbon capture [9, 35].

To tackle this relevant yet under-researched application, I developed CommonOcean [22], the first dedicated open-source framework for maritime motion planning research with easy-to-use benchmarks and tools. In particular, CommonOcean includes six software tools: 1) an interface to connect any motion planner with CommonOcean traffic scenarios, 2) a sailability checker, 3) vessel dynamics models and parameters for multiple vessel types, 4) a formalization of universal traffic rules in temporal logic [21], 5) a scenario converter for maritime traffic data, and 6) a customizable simulation environment that includes a parameterized rule-reactive navigation model for open sea traffic [28].

III. Research Agenda

Algorithms for reliable ML models. Although real-world robotic systems share similar safety and task requirements, their integration typically has to be manually tailored to ensure performance and robustness. Therefore, automatically customizing global requirements to application-specific constraints is needed for real-world autonomy. I envision a modular framework with adaptive safeguards where the models are fine-tuned as more data becomes available and leverage expressive specification blueprints. To this end, I aim to develop a capable context manager, probably based on a foundation model [5, 18], that curates complex safety specifications given the task and provides confidence metrics to decide on a hierarchy of specifications. If specification violations cannot be prevented, the specifications are soundly relaxed while ensuring their maximal satisfaction.

Learning guidance with formal methods. Despite their differences, biological processes and robot motion planning share a common real-world challenge: data for machine learning is costly and often a priori unavailable. My work shows that formal methods can enable learning by making abstract knowledge computationally tractable. However, the burden of syntactically and semantically correct conversion is mostly carried by the engineer. For automated formal-methods-guided ML, I plan to develop universal ML algorithms that leverage various types of abstract knowledge to efficiently search for the optimal solution. To this end, I envision learning semantically-structured embedding spaces that allow for finding solutions in a lower-dimensional space and context-adaptive action spaces, which are automatically derived from diverse multi-modal information sources.

Demonstrating feasibility with CommonOcean. Automating maritime navigation can erase health and environmental hazards and serves as a prime example for demonstrating the impact of my research. Vessel autonomy remains an unsolved robotic challenge due to complex safety requirements, low-frequency traffic data, and control-relevant environmental disturbances. My long-term goal is to develop CommonOcean as the premier platform for research on advanced vessel navigation and decision-making. A next step is expanding the capabilities of the platform to reflect real-world disturbances such as wind, waves, and current. Moreover, I aim to develop an open-science real-world test bed to facilitate realistic research evaluation. This makes CommonOcean attractive for application specialists as well as researchers from robotics and neuro-symbolic AI. **Ultimately, I aim for a foundational ecosystem for reliable and efficient real-world autonomy based on ML and guided by formal methods.**

Acknowledgments

I am grateful to my collaborators for their contributions to the research underpinning this statement.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pages 22–31, 2017.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, pages 2669–2678, 2018.

[3] Matthias Althoff, Goran Frehse, and Antoine Girard. Set propagation techniques for reachability analysis. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1):369–395, 2021.

[4] Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and T. P. Singh. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36(21):12809–12844, 2024.

[5] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

[6] Alberto Camacho, Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. LTL and beyond: Formal languages for reward function specification in reinforcement learning. In *Proc. of the Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 6065–6073, 2019.

[7] Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, pages 3387–3395, 2019.

[8] HeeSun Choi, Cindy Crump, Christian Duriez, Asher Elmquist, Gregory Hager, David Han, Frank Hearl, Jessica Hodgins, Abhinandan Jain, Frederick Leve, et al. On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. *Proc. of the National Academy of Sciences (PNAS)*, 118(1), 2021.

[9] Andreas Doering, Marius Wiggert, Hanna Krasowski, Manan Doshi, Pierre F.J. Lermusiaux, and Claire J. Tomlin. Stranding risk for underactuated vessels in complex ocean currents: Analysis and controllers. In *Proc. of the IEEE Conf. on Decision and Control (CDC)*, pages 7055–7060, 2023.

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proc. of the Annual Conf. on Robot Learning*, pages 1–16, 2017.

[11] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

[12] Mohammed Eslami, Aaron Adler, Rajmonda S Caceres, Joshua G Dunn, Nancy Kelley-Loughnane, Vanessa A Varaljay, and Hector Garcia Martin. Artificial intelligence for synthetic biology. *Communications of the ACM*, 65 (5):88–97, 2022.

[13] European Maritime Safety Agency. Annual overview of marine casualties and incidents 2024. Technical report, EMSA, 2024.

[14] Benjamin David Evans, Raphael Trumpp, Marco Caccamo, Felix Jahncke, Johannes Betz, Hendrik Willem Jordaan, and Herman Arnold Engelbrecht. Unifying F1TENTH autonomous racing: Survey, methods and benchmarks. *arXiv:2402.18558*, 2024.

[15] Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2019.

[16] Thor I. Fossen. *Handbook of Marine Craft Hydrodynamics and Motion Control*. John Wiley & Sons, 2011.

[17] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[18] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv:2405.07987*, 2024.

[19] Niklas Kochdumper, Hanna Krasowski, Xiao Wang, Stanley Bak, and Matthias Althoff. Provably safe reinforcement learning via action projection using reachability analysis and polynomial zonotopes. *IEEE Open Journal of Control Systems*, 2:79–92, 2023.

[20] Nathan Koenig and Andrew Howard. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2149–2154, 2004.

[21] Hanna Krasowski and Matthias Althoff. Temporal logic formalization of marine traffic rules. In *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*, pages 186–192, 2021.

[22] Hanna Krasowski and Matthias Althoff. CommonOcean: Composable benchmarks for motion planning on oceans. In *Proc. of the IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, pages 1676–1682, 2022.

[23] Hanna Krasowski and Matthias Althoff. Provable traffic rule compliance in safe reinforcement learning on the open sea. *IEEE Transactions on Intelligent Vehicles*, pages 1–18, 2024. ISSN 2379-8904.

[24] Hanna Krasowski, Xiao Wang, and Matthias Althoff. Safe reinforcement learning for autonomous lane changing using set-based prediction. In *Proc. of the IEEE Int. Conf. on Intelligent Transportation Systems (ITSC)*, pages 1–7, 2020.

[25] Hanna Krasowski, Prithvi Akella, Aaron D. Ames, and

Matthias Althoff. Safe reinforcement learning with probabilistic guarantees satisfying temporal logic specifications in continuous action spaces. In *Proc. of the IEEE Conf. on Decision and Control (CDC)*, pages 4372–4378, 2023.

[26] Hanna Krasowski, Jakob Thumm, Marlon Müller, Lukas Schäfer, Xiao Wang, and Matthias Althoff. Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking. *Transactions on Machine Learning Research*, 2023.

[27] Hanna Krasowski, Eric Palanques-Tost, Calin Belta, and Murat Arcak. Learning biomolecular models using signal temporal logic. In *Proc. of the Conf. on Learning for Dynamics and Control*, 2025.

[28] Hanna Krasowski, Stefan Schärdinger, Murat Arcak, and Matthias Althoff. Intelligent sailing model for open sea navigation. *arXiv*, 2025.

[29] Karen Leung, Nikos Aréchiga, and Marco Pavone. Backpropagation through signal temporal logic specifications: Infusing logical structure into gradient-based methods. *The International Journal of Robotics Research*, 42(6): 356–370, 2023.

[30] Xiao Li, Zachary Serlin, Guang Yang, and Calin Belta. A formal methods approach to interpretable reinforcement learning for robotic planning. *Science Robotics*, 4(37), 2019.

[31] Evan Prianto, MyeongSeop Kim, Jae-Han Park, Ji-Hun Bae, and Jung-Su Kim. Path planning for multi-arm manipulators using deep reinforcement learning: Soft actor–critic with hindsight experience replay. *Sensors*, 20(20), 2020.

[32] Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2): 945–990, 2022.

[33] Roland Stolz, Hanna Krasowski, Jakob Thumm, Michael Eichelbeck, Philipp Gassert, and Matthias Althoff. Excluding the irrelevant: Focusing reinforcement learning through continuous action masking. In *Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[34] Kim P. Wabersich, Andrew J. Taylor, Jason J. Choi, Koushil Sreenath, Claire J. Tomlin, Aaron D. Ames, and Melanie N. Zeilinger. Data-driven safety filters: Hamilton-Jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, 43(5):137–177, 2023.

[35] Marius Wiggert, Manan Doshi, Pierre F.J. Lermusiaux, and Claire J. Tomlin. Navigating underactuated agents by hitchhiking forecast flows. In *Proc. of the IEEE Conf. on Decision and Control (CDC)*, pages 2417–2424, 2022.

[36] Sean Wilson, Paul Glotfelter, Li Wang, Siddharth Mayya, Gennaro Notomista, Mark Mote, and Magnus Egerstedt. The Robotarium: Globally impactful opportunities, challenges, and lessons learned in remote-access, distributed control of multirobot systems. *IEEE Control Systems Magazine*, 40(1):26–44, 2020. doi: 10.1109/MCS.2019.2949973.