# Combining Runtime Monitoring and Machine Learning with Human Feedback

Lukina, Anna

**DOI**
10.1609/aaai.v37i13.26815

**Publication date**
2023

**Document Version**
Final published version

**Published in**
AAAI-23 Special Programs, IAAI-23, EAAI-23, Student Papers and Demonstrations

**Citation (APA)**
Lukina, A. (2023). Combining Runtime Monitoring and Machine Learning with Human Feedback. In B. Williams, Y. Chen, & J. Neville (Eds.), *AAAI-23 Special Programs, IAAI-23, EAAI-23, Student Papers and Demonstrations* (pp. 15448-15448). (Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023; Vol. 37). American Association for Artificial Intelligence (AAAI). https://doi.org/10.1609/aaai.v37i13.26815

# Combining Runtime Monitoring and Machine Learning with Human Feedback

**Anna Lukina**

Delft University of Technology, The Netherlands
a.lukina@tudelft.nl

### Abstract

State-of-the-art machine-learned controllers for autonomous systems demonstrate unbeatable performance in scenarios known from training. However, in evolving environments—changing weather or unexpected anomalies—, safety and interpretability remain the greatest challenges for autonomous systems to be reliable and are the urgent scientific challenges.

Existing machine-learning approaches focus on recovering lost performance but leave the system open to potential safety violations. Formal methods address this problem by rigorously analysing a smaller representation of the system but they rarely prioritize performance of the controller.

*We propose to combine insights from formal verification and runtime monitoring with interpretable machine-learning design for guaranteeing reliability of autonomous systems.*

## New Faculty Highlights: Extended abstract

A possible way to address the problem of real-time reliability of machine-learned controllers is to introduce a monitor, a piece of software that observes the system and detects dangerous violations automatically given a carefully designed safety specification. This may already help, however, there are several scientific challenges on the way.

To design a monitor, we need to express what a controller knows on an abstract level capturing only the knowledge critical for satisfying a given runtime specification. In our previous work (Henzinger, Lukina, and Schilling 2020), we tackled a similar problem for neural networks, where we abstracted the latent knowledge during training. We then used this abstraction for detecting latent behavior deviating from trained knowledge in prediction time. In general, the abstracted controller may be based on an arbitrary black box.

As we showed in our previous work (Alamdari et al. 2020), decision trees can approximate the knowledge of the black-box reinforcement-learned controllers and reveal safety vulnerabilities. These can further serve as specification for formal synthesis of correct-by-design controllers. Viewing decision-tree learning as a mathematical optimization problem with an objective function and a set of constraints, we can satisfy the given specification precisely. However, most decision tree works, including our own

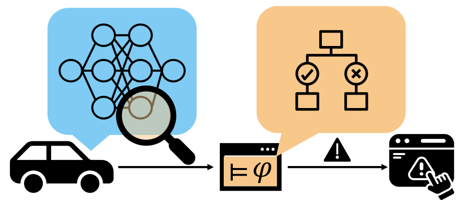


Figure 1: Runtime monitoring machine-learned controllers based on interpretable abstractions and reporting violations of the safety property $\varphi$ to human operators.

(Demirović et al. 2022), primarily focus on classical metrics from machine learning rather than formal safety.

One inherent problem of monitors is that they may not be able to recognize things beyond their specification. If after failure inspection and possible repair the safety specification changes, the monitor should be able to adapt to this feedback. This can be done efficiently for neural networks using active learning (Lukina, Schilling, and Henzinger 2021), and potentially for black-box controllers.

*This research is aimed to fuel the discussion about the open challenges in the intersection of formal methods and machine learning.*



## References

Alamdari, P. A.; Avni, G.; Henzinger, T. A.; and Lukina, A. 2020. Formal Methods with a Touch of Magic. In *2020 Formal Methods in Computer Aided Design (FMCAD)*, 138–147.

Demirović, E.; Lukina, A.; Hebrard, E.; Chan, J.; Bailey, J.; Leckie, C.; Ramamohanarao, K.; and Stuckey, P. J. 2022. MurTree: Optimal Decision Trees via Dynamic Programming and Search. *Journal of Machine Learning Research*, 23(26): 1–47.

Henzinger, T. A.; Lukina, A.; and Schilling, C. 2020. Outside the Box: Abstraction-Based Monitoring of Neural Networks. *ECAI 2020*, 2433–2440.

Lukina, A.; Schilling, C.; and Henzinger, T. A. 2021. Into the Unknown: Active Monitoring of Neural Networks. In Feng, L.; and Fisman, D., eds., *Runtime Verification*, volume 12974, 42–61. Cham: Springer International Publishing. ISBN 978-3-030-88493-2 978-3-030-88494-9.