ISSN 2186-7437

NII Shonan Meeting Report

No.178

Formal Methods for Trustworthy AI-based Autonomous Systems

Ichiro Hasuo
Einar Broch Johnsen
Martin Leucker

October 16–19, 2023

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Shonan 178: Formal Methods for Trustworthy AI-based Autonomous Systems

Ichiro Hasuo, Einar Broch Johnsen, and Martin Leucker

Shonan Village Center, October 16 - 19, 2023

1 Introduction

Formal methods (FM) have a long and successful history in contributing to the reliability and risk minimization of safety-critical software systems [5]. Like other human engineering disciplines, they are based on rigorous mathematical modeling and analysis of system properties. Due to tremendous advances in artificial intelligence (AI), the situation is changing as several of the system functions are increasingly implemented and enabled by machine learning (ML) components; that is, they are not explicitly programmed by humans, but instead automatically generated from large sets of data. Such ML-enabled components are instrumental for the automation of functions in smart, autonomous cyber-physical systems, which already affect our society in many application domains, including self-driving cars, smart grids, smart healthcare, and smart manufacturing. For instance, deep neural networks (DNN) have become an essential tool for tasks such as image perception and object classification in autonomous vehicles. However, data-driven learning capabilities are inherently hard to characterize and their combination with other, more traditional AI (e.g. rule-based) and system engineering tools is not well understood and can introduce additional risks. For example, DNNs are not robust to adversarial perturbations (APs), meaning that even minimal changes applied to the input, often imperceptible to the human senses, can cause the network to completely misclassify the input [4, 20, 12, 9].

In safety-critical systems, such misclassification can lead to dramatic consequences, such as a fatal self-driving accident where the sensor system in the car failed to identify a large truck with a trailer crossing the highway [14]. As the operational behavior of learning-enabled components is a function of the data they are trained on, it can also be distorted by bias or crippled by gaps in the data set. For example, an AI-based recruiting tool at Amazon had to be scrapped because it systematically favored men for technical jobs, just as reflected in the actual data [13].

A fundamental problem here is that it is not possible to learn ethical and legal principles like “all human beings are equal” from existing data only, but such guidelines need to be represented and ensured by other techniques, including logical formalisms. As data-driven learning capabilities may cause unpredictable emergent behavior and can lead to harmful events or violations of social norms and conventions, recent initiatives such as the EU-HLEG Ethics Guidelines for Trustworthy AI [8] call for regulation of AI-based systems, asking to design and operate them in a way that they are trustworthy and comply with safety, security, privacy, robustness, legal, ethical, and explainability requirements up to a level adequate for the application. Similar considerations are also beginning to find their way into pre-normative standards for system design, such as the DIN SPEC 92001-1:2019-04 on AI Life Cycle Processes and Quality Requirements [6].

2 State of the Art and Challenges

The current state-of-the-art on safety and security assurance is based on model-driven system design that consists in rigorously specifying the structure and the behavior using formal models. These models enable static and dynamic analysis and verification techniques that can provide comprehensive guarantees about the correctness of systems [3]. While static techniques such as model checking can perform sophisticated analysis of software without actually executing it, their exhaustive nature makes it difficult to scale them without losing precision. Runtime verification (RV) [18] is a more lightweight verification technique that complements design-time analysis by observing the system at runtime, checking intended properties online and possibly mitigating violations through specified recovery actions. Though model-driven and formal approaches are increasingly adopted in industry, they are generally still applicable only to non-learning systems operating in well-characterized environments, as data-driven learning components and environmental aspects that are not fully known or foreseen at design-time introduce uncertainty in the design process in the form of "black boxes" in the system model.

As the state-of-art of formal verification techniques for ML components is still quite limited, the great difficulty to provide formal guarantees about their behavior is a major obstacle to using ML and AI methods in safety-critical autonomous systems engineering [16]. A recent approach is Reluplex [17], an SMT solver for verifying DNNs that has been successfully evaluated on a DNN implementation of a next-generation airborne collision avoidance system for unmanned aircraft. There also has been a significant effort in recent years to develop tools such as DeepFool [19] and cleverhans [11] that generate perturbations to add to the original training set to further test the ML classifiers. Other recent and more sophisticated testing approaches [7] follow a semantic-based approach, using simulation environments to generate more plausible adversarial examples rather than just small perturbations over the input image. However, because attackers can use inputs never considered before in the testing process, testing is not enough to provide guarantees, and providing formal guarantees about the space of inputs that will be correctly processed remains a major challenge that plays a fundamental role in improving trustworthiness of autonomous systems.

We thus need novel formal techniques that offer high levels of guarantees for large-scale autonomous systems with learning-enabled components. For example, a promising direction is new mechanisms to enforce dynamic assurance of safety and dependability at runtime. A related architectural approach [21] in control theory uses a decision module to safeguard a high-performance controller (e.g., DNN) by switching to a pre-certified baseline controller if certain safe operating conditions are close to being violated. Potential approaches to this challenge may have a tremendous societal and economical impact on our society by reducing the warranty, liability, certification and design costs of autonomous systems and subsystems with learning components.

3 Objectives and Topics of the Workshop

The workshop will be a joint and multi-disciplinary effort to develop novel formal methods for the rigorous design of autonomous systems with learning components. While AI and data-driven learning is essential for achieving autonomy, we believe that it needs to be properly combined and complemented with formal, knowledge/model-based techniques to make the system design and operation as trustworthy as needed. Because in principle, we cannot guarantee that complex engineered systems will be completely free of risks and design flaws, we seek to define different classes of trustworthiness and trade-offs to decide if a system is trustworthy enough for its intended use in terms of functionality, safety, security, privacy, lawfulness and comprehensibility for humans. The workshop aims at fostering novel approaches, capable of capturing and effectively balancing these concerns in autonomous systems, by bringing together key people from the FM and the AI communities.

Technically, we aim for a multi-stage approach where trustworthiness will be established as far as possible at design-time, while ensuring that possible variations during the autonomous system's operation—due to changes in the environmental context, or the self-adaptive behavior of learning capabilities — can be assessed and dealt with at runtime. The envisaged approach will determine which requirements can be guaranteed at the design-time and which ones are left to be monitored and possibly enforced at runtime. This requires tight collaboration between different research communities on FM (prominently from the rigorous system design and the runtime verification area), control theory and robotics, ML and AI, and semantic technologies. A purpose of this Shonan meeting is to establish and consolidate a group of experts, with different perspectives and both from academic and industrial research, to coordinate currently fragmented research activities and address the discussed challenge in diverse application areas such as autonomous driving, robotics, industrial IoT and smart manufacturing, smart medical devices and healthcare, and smart energy grids.

Research topics and questions that were addressed and dealt with during the meeting include:

- Formal modeling languages capable of expressing key properties related to trustworthiness of AI-based autonomous systems with learning components;

- Design-time analysis approaches for static verification of system properties expressed in formal languages, including e.g. techniques for model checking properties of DNNs;
- Runtime verification techniques for AI-based systems including online generation of monitors, detection and diagnosis of critical events, and enforcement of requirements;
- Novel methods to enforce correctness requirements both during design and at runtime to achieve overall autonomic correctness, including properly balancing between them;
- Methods for testing and formal verification of learning and self-adapting capabilities of autonomous systems to provide high-level dependability and robustness guarantees;
- Approaches to couple knowledge generated by runtime analysis with high-level autonomic decision processes (e.g. triggering selective re-learning of components).

These topics were addressed in the form of mini-tutorials given by leading researchers in FM and machine learning/AI to familiarize everyone with the proper terminology, research methodologies and current approaches, a selection of shorter and deeper technical presentations that reported on problems and the state of the art (including tool demonstrations for diverse application domains), and by focused group discussions moderated by the organizers.

4 Meeting Schedule

- 15 October 2023, Sunday evening
  - Arrival
- 16 October 2023, Monday morning
  - Opening by the local Shonan team
  - Opening by the organizers
  - Short self-introduction by the participants
- 16 October 2023, Monday afternoon
  - Short self-introduction by the participants (cont.)
  - Sanjit A. Seshia: *Towards Verified AI: Formal Specification and Environment Modeling* (tutorial)
- 17 October 2023, Tuesday morning

  - Etienne André: *Configuring Timing Parameters to Ensure Opacity* (tutorial)
  - Ichiro Hasuo: *Goal-Aware RSS for Complex Scenarios via Program Logic* (talk)
  - Plenary session: identifying themes for breakout sessions

- 17 October 2023, Tuesday afternoon
  - Erika Ábrahám: *SMT Solving* (tutorial)
  - Breakout sessions B1, B2

- 18 October 2023, Wednesday morning
  - Masaki Waga: *Dynamic Shielding for Reinforcement Learning in Black-Box Environments* (talk)
  - Marija Slavkovik: *Quick(?) and Dirty Intro to AI Ethics* (tutorial)
  - Breakout sessions B3, B4

- 18 October 2023, Wednesday afternoon
  - Excursion

- 19 October 2023, Thursday morning
  - Mahsa Varshosaz: *Formal Specification and Testing for Reinforcement Learning* (talk)
  - Krzysztof Czarnecki: *Automated Driving* (tutorial)
  - Plenary session

- 19 October 2023, Thursday afternoon
  - Departure

5 Overview of Talks and Discussions

5.1 Towards Verified AI: Formal Specification and Environment Modeling

Speaker: Sanjit A. Seshia, UC Berkeley, USA

Abstract: We propose a new probabilistic programming language for the design and analysis of perception systems, especially those based on machine learning. Specifically, we consider the problems of training a perception system to handle rare events, testing its performance under different conditions, and debugging failures. We show how a probabilistic programming language can help address these problems by specifying distributions encoding interesting types of inputs and sampling these to generate specialized training and test sets. More generally, such languages can be used for cyber-physical systems and robotics to write environment models, an essential prerequisite to any formal analysis. In this paper, we focus on systems like autonomous cars and robots, whose environment is a "scene", a configuration of physical objects and agents. We design a domain-specific language, Scenic, for describing "scenarios" that are distributions over scenes. As a probabilistic programming language, Scenic allows assigning distributions to features of the scene, as well as declaratively imposing hard and soft constraints over the scene. We develop specialized techniques for sampling from the resulting distribution, taking advantage of the structure provided by Scenic's domain-specific syntax. Finally, we apply Scenic in a case study on a convolutional neural network designed to detect cars in road images, improving its performance beyond that achieved by state-of-the-art synthetic data generation methods.

Paper: [10]

5.2 SMT Solving

Speaker: Erika Ábrahám, RWTH Aachen, Germany

Abstract: SMT (Satisfiability Modulo Theories) solving is a technology for the fully automated solution of logical formulas. Due to their impressive efficiency, SMT solvers are nowadays frequently used in a wide variety of applications. These tools are general purpose and as off-the-shelf solvers, their usage is truly integrated. A typical application (i) encodes real-world problems as logical formulas, (ii) check these formulas for satisfiability with the help of SMT solvers, and — in case of satisfiability — (iii) decodes their solutions back to solutions of the original real-world problem.

Paper: [1]

5.3 Goal-Aware RSS for Complex Scenarios via Program Logic

Speaker: Ichiro Hasuo, NII, Japan

Abstract: We introduce a goal-aware extension of responsibility-sensitive safety (RSS), a recent methodology for rule-based safety guarantee for automated driving systems (ADS). Making RSS rules guarantee goal achievement – in addition to collision avoidance as in the original RSS – requires complex planning over long sequences of maneuvers. To deal with the complexity, we introduce a compositional reasoning framework based on program logic, in which one can systematically develop RSS rules for smaller subscenarios and combine them to obtain RSS rules for bigger scenarios. As the basis of the framework, we introduce a program logic dFHL that accommodates continuous dynamics and safety conditions. Our framework presents a dFHL-based workflow for deriving goal-aware RSS rules; we discuss its software support, too. We conducted experimental evaluation using RSS rules in a safety architecture. Its results show that goal-aware RSS is indeed effective in realizing both collision avoidance and goal achievement.

Paper: [15]

5.4 Configuring Timing Parameters to Ensure Opacity

Speaker: Etienne André, Université Sorbonne Paris Nord, France

Abstract: Information leakage can have dramatic consequences on systems security. Among harmful information leaks, the timing information leakage occurs whenever an attacker successfully deduces confidential internal information. In this work, we consider that the attacker has access (only) to the system execution time. We address the following timed opacity problem: given a timed system, a private location and a final location, synthesize the execution times from the initial location to the final location for which one cannot deduce whether the system went through the private location. We also consider the full timed opacity problem, asking whether the system is opaque for all execution times. We show that these problems are decidable for timed automata (TAs) but become undecidable when one adds parameters, yielding parametric timed automata (PTAs). We identify a subclass with some decidability results. We then devise an algorithm for synthesizing PTAs parameter valuations guaranteeing that the resulting TA is opaque. We finally show that our method can also apply to program analysis.

Paper: [2]

5.5 Tutorial: Quick(?) and Dirty Intro to AI Ethics

Speaker: Marija Slavkovik, University of Bergen, Norway

Abstract: To have ethical AI two questions need to be answered: i) what is the ethical impact that an AI system can have, and, ii) what does it mean for an AI system to behave ethically. The lack of answers to both of these questions hinder the identification of what are the values or principles that we want upheld by AI and for AI. Identifying these principles is not enough, we also want to define them so that they can be operational, or at least understand what operational means here. There is a gap between moral philosophy and ethically behaving AI. The tutorial contributes towards closing this gap, by motivating researchers to interpret an abstract principle from moral philosophy into an algorithmic property that can be formally specified and measured or computationally implemented. The tutorial uses recent articles in AI ethics that attempt to define and identify pertinent ethical principles, as well as ethically motivated desirable algorithmic properties.

Paper: [22]

5.6 Dynamic Shielding for Reinforcement Learning in Black-Box Environments

Speaker: Masaki Waga, Kyoto University, Japan

Abstract: It is challenging to use reinforcement learning (RL) in cyber-physical systems due to the lack of safety guarantees during learning. Although there have been various proposals to reduce undesired behaviors during learning, most of these techniques require prior system knowledge, and their applicability is limited. This paper aims to reduce undesired behaviors during learning without requiring any prior system knowledge. We propose dynamic shielding: an extension of a model-based safe RL technique called shielding using automata learning. The dynamic shielding technique constructs an approximate system model in parallel with RL using a variant of the RPNI algorithm and suppresses undesired explorations due to the shield constructed from the learned model. Through this combination, potentially unsafe actions can be foreseen before the agent experiences them. Experiments show that our dynamic shield significantly decreases the number of undesired events during training.

Paper: [24]

5.7 Formal Specification and Testing for Reinforcement Learning

Speaker: Mahsa Varshosaz, IT-University of Copenhagen, Denmark

Abstract: The development process for reinforcement learning applications is still exploratory rather than systematic. This exploratory nature reduces reuse of specifications between applications and increases the chances of introducing programming errors. This paper takes a step towards systematizing the development of reinforcement learning applications. We introduce a formal specification of reinforcement learning problems and algorithms, with a particular focus on temporal difference methods and their definitions in backup diagrams. We further develop a test harness for a large class of reinforcement learning applications based on temporal difference learning, including SARSA and Q-learning. The entire development is rooted in functional programming methods; starting with pure specifications and denotational semantics, ending with property-based testing and using compositional interpreters for a domain-specific term language as a test oracle for concrete implementations. We demonstrate the usefulness of this testing method on a number of examples, and evaluate with mutation testing. We show that our test suite is effective in killing mutants (90% mutants killed for 75% of subject agents). More importantly, almost half of all mutants are killed by generic write-once-use-everywhere tests that apply to any reinforcement learning problem modeled using our library, without any additional effort from the programmer.

Paper: [23]

5.8 Automated Driving

Speaker: Krzysztof Czarnecki

Abstract: This tutorial gave a state-of-the-art overview of automated driving, surveying the problems and solutions available today to build automated drivers, as well as the key safety standards involved in automated driving. The talk discussed different levels of automation, and surveyed the levels achieved by different automated pilots, including Mercedes Drive Pilot, Waymo One, Tesla Autopilot, and autonomous emergency braking and autonomous trucking. The talk further discussed the challenges in building a self-driving car and the key safety standards involved.

B1: Specifications & Ethics

The session explored methods to formalize ethical behavior for machines, focusing on deriving formal specifications from examples of ethical and unethical actions. Discussions questioned whether the focus was truly on ethics or related concepts. Three approaches to defining right and wrong for machines were identified: moral theory, authority or social choice, and observation and learning through interaction, such as inverse reinforcement learning (IRL). IRL, which infers values from informal interactions, was highlighted as a method for value alignment, with applications like autonomous driving and extensions to derive formal specifications.

Fairness in machine learning was discussed as a critical ethical consideration. Group fairness metrics, such as demographic parity (equal acceptance rates) and equalized odds (conditioning on attributes), were contrasted with individual fairness, which demands treating similar individuals equally. Tradeoffs between these approaches, particularly in biased real-world settings, were acknowledged. Trustworthiness and explainable AI were also key topics, emphasizing the importance of understanding both the internal states of AI systems and external conditions influencing outputs. Explainability was described as a tunable concept, requiring techniques to balance transparency and system complexity.

Challenges in formalizing ethical specifications were explored, including reconciling vague informal values with precise formal reasoning, designing specification languages that handle heterogeneous and vaguely defined requirements, and addressing aspects that defy formalization through constructs like “oracle” functions. Traffic rules and multi-agent systems were identified as suitable domains for formal ethical specifications due to their well-defined requirements. Some ethical concepts, like fairness, are testable, while others, such as human rights, may remain non-testable.

In conclusion, formalizing ethics in AI requires domain-specific approaches to bridge informal principles with formal reasoning. The session emphasized focusing on areas where ethical requirements are clear, testable, and well-suited for formalization.

B2: Modular AI-Based Systems

The breakout session on Modular AI Systems discussed the design, advantages, challenges, and potential solutions for integrating classical and AI components in modular architectures. These architectures aim to balance the benefits of end-to-end learning with the structure and interpretability of modular designs. Modular systems were contrasted with end-to-end approaches, which often involve monolithic neural networks. Several modular architectures were discussed, such as multi-tasking with shared latent representations and sequential pipelines. In multi-tasking, a feature encoder generates a shared latent representation used by a primary task and auxiliary tasks, enabling efficient multi-task learning. In sequential pipelines, modules—comprising deep neural networks (DNNs) and optionally differentiable classical algorithms—exchange latent representations explicitly. This enables task-specific processing while leveraging the benefits of modular design.

Modular AI systems offer several advantages, including system decomposition and work division, which allow for specialization and distributed development. They also enable reusability, where individual modules can be reused across different systems. Moreover, modular systems offer graceful degradation, meaning that failures in one module may not cascade through the system, improving robustness. Additionally, modular designs can potentially minimize the need for extensive training and testing data, and they allow for easier identification of gaps in the training and testing data.

Despite their benefits, modular AI systems present significant challenges. For example, classical components benefit from well-understood assume and guarantee (A/G) reasoning, but similar methods are not yet established for AI components. In multi-task architectures, compositional A/G rules between modules are difficult to define because modules may operate independently at certain times. In sequential-task architectures, latent space data exchange between modules complicates specifying A/G rules, raising issues like lack of transparency, potential unwanted information leaks, and the correctness of decisions based on latent space representations. Additionally, the incrementality and non-monotonicity of DNN components complicate system-wide guarantees.

The group identified several research directions and strategies to address these challenges. These include causal models and interventions to ensure decisions align with the correct reasons and enhance explainability, as well as specification mining to extract specifications from system behavior and relate global system requirements to individual components. Relating specifications with explanations and developing modular specifications for data-driven end-to-end simulation were also discussed. Furthermore, automated reasoning and reactive synthesis techniques can be applied to ensure system-level correctness, especially in numerical problems.

An example application of modular AI systems discussed during the session was automated driving, where modular architectures can decompose complex tasks like perception, planning, and control into manageable components.

B3: Runtime Assurance

The breakout group on Runtime Assurance discussed the role of runtime mechanisms in ensuring the safety, security, performance, and reliability of systems. These mechanisms aim to compensate for the inability to guarantee safety fully at design time and to improve system learning processes. The group identified three primary objectives: imposing safety properties to ensure adherence to defined safety, security, and performance criteria during operation; enhancing reliability to increase dependability in dynamic and uncertain environments; and improving learning processes by accelerating and refining machine learning through runtime assurances.

Several challenges were highlighted, particularly in modeling and interacting with environments. Simplified assumptions often underlie safety guarantees, leading to discrepancies with real-world scenarios, especially in reinforcement learning (RL) involving deep neural networks (DNNs), where the environment behaves like a black box. It is challenging to adapt when the real environment deviates from expected models. Realistic modeling for architectures such as shields and Simplex was also discussed. Another challenge involves anticipatory semantics, specifically addressing the breaking of assumptions dynamically and deploying end-to-end monitoring in machine learning architectures. Enforcing runtime security policies without disrupting system operations is another significant issue.

The group explored several potential solutions and research questions. Runtime verification (RV) was acknowledged as a partial solution to runtime assurance. Tools and frameworks such as ULGEN and SOTER were noted for enabling formal verification and safe learning. Shields and runtime enforcement mechanisms were proposed to maintain safety properties during operation. Safe reinforcement learning (Safe RL) was discussed as a method to integrate safety into learning processes by using runtime monitors to influence behavior through rewards and penalties. Maintaining consistency between logical system specifications and neural network representations through specification mining was another key area of interest. Simplex architecture, which employs runtime monitoring to detect assumption violations and take mitigation actions such as dynamically adapting shield parameters, was highlighted as a promising approach.

The group proposed fostering collaboration between formal methods (FM) experts and machine learning (ML) researchers, potentially through a COST action or a shared platform. This collaboration would aim to unify efforts to address safety and runtime assurance challenges in AI-driven systems.

B4: Neuro-Symbolic X

The breakout session on Neuro-Symbolic-X explored the integration of subsymbolic approaches, such as neural networks (NNs) and statistical methods, with symbolic methods from formal methods (FM) and knowledge representation (KR). This hybrid approach seeks to unify data-driven inference and deductive reasoning to address diverse challenges in AI.

Sub-symbolic methods, like neural networks, prove beneficial for data-driven tasks that address hard-to-formalize aspects of problems and provide approximations through learning. Symbolic methods, in contrast, bring the advantage of formalizable prior knowledge, explicit reasoning capabilities, and interpretable interfaces, which are critical for domains like medical diagnostics. The session highlighted the complementary strengths of these paradigms, with sub-symbolic methods offering flexible modeling and symbolic methods enabling logical inference and human interpretability.

Several integration strategies were discussed, including:

- Neural networks calling symbolic reasoners and vice versa.
- Differentiable logic representations to bridge neural and symbolic reasoning.
- Neural networks generating reasoning-related code.
- Injecting symbolic knowledge into neural networks using loss functions.
- Employing vector databases for knowledge grounding.

Applications for neuro-symbolic systems include shielding AI models to enhance safety, developing governance and constitutional frameworks for conversational AI and decision-making systems, and advancing robotics and medical technologies.

Key challenges include improving the interpretability of neural and latent representations, which remains a critical obstacle to integrating sub-symbolic and symbolic approaches effectively. Addressing these challenges would enable the combined approach to unlock powerful capabilities across a wide range of applications.

6 List of Participants

- Prof. Ichiro Hasuo (NII, Tokyo, Japan)
- Prof. Martin Leucker (University of Lübeck, Germany)
- Prof. Einar Broch Johnsen (University of Oslo, Norway)
- Prof. Erika Abraham (RWTH Aachen, Germany)
- Prof. Toshiaki Aoki (JAIST, Japan)
- Prof. Marija Slavkovik (University of Bergen, Norway)
- Prof. Andrzej Wasowski (ITU University of Copenhagen, Denmark)
- Prof. Lijun Zhang (Institute of Software, Chinese Academy of Sciences)

- Prof. Etienne André (Université Sorbonne Paris Nord, France)
- Prof. Cyrille Artho (KTH Royal Institute of Technology, Sweden)
- Prof. Reiner Hähnle (TU Darmstadt, Germany)
- Prof. Fuyuki Ishikawa (NII, Tokyo, Japan)
- Prof. Yi Kwangkeun (Seoul National University, Korea)
- Prof. Sanjit Seshia (University of California, Berkeley, USA)
- Dr. Jurriaan Rot (Radboud Uni, Nijmegen, the Netherlands)
- Prof. Gerardo Schneider (University of Gothenburg, Sweden)
- Dr. Masaki Waga (Kyoto University, Japan)
- Prof. Krzysztof Czarnecki (University of Waterloo, Canada)
- Dr. Dejan Nickovic (AIT Austrian Institute of Technology, Austria)
- Prof. Leonardo Mariani (University of Milano-Bicocca, Italy)
- Prof. Ferruccio Damiani (University of Turin, Italy)
- Prof. Ezio Bartocci (TU Wien, Austria)
- Prof. Panagiotis Katsaros (University Of Thessaloniki, Greece)
- Prof. Mahsa Varshosaz (ITU University of Copenhagen, Denmark)
- Mr. Andoni Rodriguez (IMDEA Software Institute, Spain)
- Dr. Ernst Moritz Hahn (University of Twente, the Netherlands)
- Dr. Martin Sachenbacher (University of Lübeck, Germany)
- Prof. Cesar Sanchez (IMDEA Software Institute, Spain)

References

[1] E. Ábrahám, J. Kovács, and A. Remke. SMT: something you must try. In P. Herber and A. Wijs, editors, *Proc. 18th Intl. Conf. on Integrated Formal Methods (iFM 2023)*, volume 14300 of *Lecture Notes in Computer Science*, pages 3–18. Springer, 2023.

[2] É. André, D. Lime, D. Marinho, and J. Sun. Guaranteeing timed opacity using parametric timed model checking. *ACM Trans. Softw. Eng. Methodol.*, 31(4):64:1–64:36, 2022.

[3] E. Bartocci, J. V. Deshmukh, A. Donzé, G. Fainekos, O. Maler, D. Nickovic, and S. Sankaranarayanan. Specification-based monitoring of cyber-physical systems: A survey on theory, tools and applications. In E. Bartocci and Y. Falcone, editors, *Lectures on Runtime Verification - Introductory and Advanced Topics*, volume 10457 of *Lecture Notes in Computer Science*, pages 135–175. Springer, 2018.

[4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezný, editors, *Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML 2013)*, volume 8190 of *Lecture Notes in Computer Science*, pages 387–402. Springer, 2013.

[5] D. Bjørner and K. Havelund. 40 years of formal methods - some obstacles and some possibilities? In C. B. Jones, P. Pihlajasaari, and J. Sun, editors, *Proc. 19th Intl. Symposium on Formal Methods (FM 2014)*, volume 8442 of *Lecture Notes in Computer Science*, pages 42–61. Springer, 2014.

[6] Deutsche Industrie Norm (DIN). SPEC 92001-1:2019-04, 2019. https://www.beuth.de/en/technical-rule/din-spec-92001-1/303650673.

[7] T. Dreossi, S. Jha, and S. A. Seshia. Semantic adversarial deep learning. In H. Chockler and G. Weissenbacher, editors, *Proc. 30th Intl. Conf. on Computer Aided Verification (CAV 2018)*, volume 10981 of *Lecture Notes in Computer Science*, pages 3–26. Springer, 2018.

[8] European Commission. Ethics guidelines for trustworthy AI. Report, 08 April 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[9] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society, 2018.

[10] D. J. Fremont, E. Kim, T. Dreossi, S. Ghosh, X. Yue, A. L. Sangiovanni-Vincentelli, and S. A. Seshia. Scenic: a language for scenario specification and data generation. *Mach. Learn.*, 112(10):3805–3849, 2023.

[11] I. J. Goodfellow, N. Papernot, and P. D. McDaniel. cleverhans v0.1: an adversarial machine learning library. *CoRR*, abs/1610.00768, 2016.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In Y. Bengio and Y. LeCun, editors, *Proc. 3rd Intl. Conf. on Learning Representations (ICLR 2015)*, 2015.

[13] Amazon ditched AI recruiting tool that favored men for technical jobs. The Guardian, 11 October 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

[14] Tesla driver dies in first fatal crash while using autopilot mode. The Guardian, 1 July 2016. https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

[15] I. Hasuo, C. Eberhart, J. Haydon, J. Dubut, R. Bohrer, T. Kobayashi, S. Pruekprasert, X. Zhang, E. A. Pallas, A. Yamada, K. Suenaga, F. Ishikawa, K. Kamijo, Y. Shinya, and T. Suetomi. Goal-aware RSS for complex scenarios via program logic. *IEEE Trans. Intell. Veh.*, 8(4):3040–3072, 2023.

[16] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. In R. Majumdar and V. Kuncak, editors, *Proc. 29th Intl. Conf. on Computer Aided Verification (CAV 2017)*, volume 10426 of *Lecture Notes in Computer Science*, pages 3–29. Springer, 2017.

[17] G. Katz, C. W. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In R. Majumdar and V. Kuncak, editors, *Proc. 29th Intl. Conf. on Computer Aided Verification (CAV 2017)*, volume 10426 of *Lecture Notes in Computer Science*, pages 97–117. Springer, 2017.

[18] M. Leucker and C. Schallhart. A brief account of runtime verification. *J. Log. Algebraic Methods Program.*, 78(5):293–303, 2009.

[19] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR 2016)*, pages 2574–2582. IEEE Computer Society, 2016.

[20] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In R. Karri, O. Sinanoglu, A. Sadeghi, and X. Yi, editors, *Proc. Asia Conf. on Computer and Communications Security (AsiaCCS 2017)*, pages 506–519. ACM, 2017.

[21] L. Sha. Using simplicity to control complexity. *IEEE Softw.*, 18(4):20–28, 2001.

[22] M. Slavkovik. Mythical ethical principles for AI and how to attain them. In M. Chetouani, V. Dignum, P. Lukowicz, and C. Sierra, editors, *Human-Centered Artificial Intelligence: Advanced Lectures*, pages 275–303. Springer, 2023.

[23] M. Varshosaz, M. Ghaffari, E. B. Johnsen, and A. Wąsowski. Formal specification and testing for reinforcement learning. *Proc. ACM Program. Lang.*, 7(ICFP), aug 2023. Proc. 28th Intl. Conference on Functional Programming (ICFP 2023).

[24] M. Waga, E. Castellano, S. Pruekprasert, S. Klikovits, T. Takisaka, and I. Hasuo. Dynamic shielding for reinforcement learning in black-box environments. In A. Bouajjani, L. Holík, and Z. Wu, editors, *Proc. 20th Intl. Symposium on Automated Technology for Verification and Analysis (ATVA 2022)*, volume 13505 of *Lecture Notes in Computer Science*, pages 25–41. Springer, 2022.