

# Safety and Robustness for Deep Learning with Provable Guarantees (Keynote)

Marta Kwiatkowska  
University of Oxford  
UK

marta.kwiatkowska@cs.ox.ac.uk

## ABSTRACT

Computing systems are becoming ever more complex, with decisions increasingly often based on deep learning components. A wide variety of applications are being developed, many of them safety-critical, such as self-driving cars and medical diagnosis. Since deep learning is unstable with respect to adversarial perturbations, there is a need for rigorous software development methodologies that encompass machine learning components. This lecture will describe progress with developing automated verification and testing techniques for deep neural networks to ensure safety and robustness of their decisions with respect to input perturbations. The techniques exploit Lipschitz continuity of the networks and aim to approximate, for a given set of inputs, the reachable set of network outputs in terms of lower and upper bounds, in anytime manner, with provable guarantees. We develop novel algorithms based on feature-guided search, games, global optimisation and Bayesian methods, and evaluate them on state-of-the-art networks. The lecture will conclude with an overview of the challenges in this field.

## CCS CONCEPTS

• **Theory of computation** → **Logic and verification**; • **Computing methodologies** → **Machine learning**; **Neural networks**.

## KEYWORDS

Automated Verification; Deep Neural Networks; Adversarial Examples

### ACM Reference Format:

Marta Kwiatkowska. 2019. Safety and Robustness for Deep Learning with Provable Guarantees (Keynote). In *Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '19)*, August 26–30, 2019, Tallinn, Estonia. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3338906.3342812>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ESEC/FSE '19, August 26–30, 2019, Tallinn, Estonia

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5572-8/19/08.

<https://doi.org/10.1145/3338906.3342812>

## BIOGRAPHY

Marta Kwiatkowska is Professor of Computing Systems and Fellow of Trinity College, University of Oxford. Prior to this she was Professor in the School of Computer Science at the University of Birmingham, Lecturer at the University of Leicester and Assistant Professor at the Jagiellonian University in Cracow, Poland. Kwiatkowska has made fundamental contributions to the theory and practice of model checking for probabilistic systems, focusing on automated techniques for verification and synthesis from quantitative specifications. She led the development of the PRISM model checker ([www.prismmodelchecker.org](http://www.prismmodelchecker.org)), the leading software tool in the area and winner of the HVC Award 2016. Probabilistic model checking has been adopted in diverse fields, including distributed computing, wireless networks, security, robotics, healthcare, systems biology, DNA computing and nanotechnology, with genuine flaws found and corrected in real-world protocols. Kwiatkowska is the first female winner of the Royal Society Milner Award and was awarded an honorary doctorate from KTH Royal Institute of Technology in Stockholm in 2014. She is the winner of two ERC Advanced Grants VERIWARE and FUN2MODEL, and is a coinvestigator of the EPSRC Programme Grant on Mobile Autonomy. Kwiatkowska is a Fellow of ACM and Member of Academia Europea.

## REFERENCES

- [1] Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, Andrea Patane, and Matthew Wicker. 2019. Statistical Guarantees for the Robustness of Bayesian Neural Networks. In *Proceedings, International Joint Conference on Artificial Intelligence (IJCAI)*.
- [2] Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, and Andrea Patane. 2019. Robustness Guarantees for Bayesian Inference with Gaussian Processes. In *Proceedings, Conference on Artificial Intelligence (AAAI)*. 7759–7768.
- [3] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. 2017. Safety Verification of Deep Neural Networks. In *Proceedings, Computer Aided Verification (CAV)*. Springer, 3–29. [https://doi.org/10.1007/978-3-319-63387-9\\_1](https://doi.org/10.1007/978-3-319-63387-9_1)
- [4] Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. 2018. Reachability Analysis of Deep Neural Networks with Provable Guarantees. In *Proceedings, International Joint Conference on Artificial Intelligence (IJCAI)*. 2651–2659. <https://doi.org/10.24963/ijcai.2018/368>
- [5] Matthew Wicker, Xiaowei Huang, and Marta Kwiatkowska. 2018. Feature-guided black-box safety testing of deep neural networks. In *Proceedings, International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*. Springer, 408–426. [https://doi.org/10.1007/978-3-319-89960-2\\_22](https://doi.org/10.1007/978-3-319-89960-2_22)
- [6] Matthew Wicker and Marta Kwiatkowska. 2019. Robustness of 3D Deep Learning in an Adversarial Setting. In *Proceedings, Computer Vision and Pattern Recognition (CVPR) 2019*. To appear.
- [7] Min Wu, Matthew Wicker, Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. 2018. A Game-Based Approximate Verification of Deep Neural Networks with Provable Guarantees. *To appear in Theoretical Computer Science*. CoRR abs/1807.03571 (2018). arXiv:1807.03571 <http://arxiv.org/abs/1807.03571>