Evaluating Architectural Safeguards for Uncertain AI Black-Box Components

Max Scheerer, FZI Forschungszentrum Informatik, Karlsruhe, Germany

Reference
Max Scheerer. *Evaluating Architectural Safeguards for Uncertain AI Black-Box Components.* Dissertation, Karlsruher Institut für Technologie (KIT), 2023, DOI: 10.5445/IR/1000159170

Supervisor: Prof. Dr. Ralf Reussner

Date of Defense: 09.05.2023

Abstract

There have been enormous achievements in the field of *Artificial Intelligence* (AI) which has attracted a lot of attention. Their unverifiable nature, however, makes them inherently unreliable. For example, there are various reports of incidents in which incorrect predictions of AI components led to serious system malfunctions (some even ended fatally). As a result, various architectural approaches (referred to as *Architectural Safeguards*) have been developed to deal with the unreliable and uncertain nature of AI. Software engineers are now facing the challenge to select the architectural safeguard that satisfies the non-functional requirements (e.g. reliability) best. However, it is crucial to resolve such design decisions as early as possible to avoid (*i*) changes after the system has been deployed (and thus potentially high costs) and to meet the rigorous quality requirements of safety-critical systems where AI is more commonly used.

This dissertation presents a model-based approach that supports software engineers in the development of AI-enabled systems by enabling the evaluation of architectural safeguards. More specifically, an approach for reliability prediction of AI-enabled systems (based on established model-based techniques) is presented. Moreover, the approach is generalised to architectural safeguards with self-adaptive capabilities, i.e. self-adaptive systems.

The approach has been validated by considering four case studies. The results show that the approach not only makes it possible to analyse the impact of architectural safeguards on the overall reliability of an AI-enabled system, but also supports software engineers in their decision-making.

Introduction. In the past years, there have been tremendous progress in the field of *Artificial Intelligence* (AI). Especially, *Deep Learning* (a subfield of AI) employs so-called *Deep Neural Networks* (DNNs) that have been successfully applied to various complex learning tasks, e.g. autonomous driving or human-robot-interaction. However, DNNs have reached a level of complexity that their correctness cannot be verified, making them inherently unreliable [2]. For example, Tian et al. [9] enumerates three incidents in autonomous driving in which erroneous predictions of AI models caused accidents (one ended fatally). As a result, various approaches have been developed at the architectural level (referred to as *Architectural Safeguards*) to deal with uncertainty associated with a prediction of an AI component, ranging from n-version programming approaches [1] to simplex architectures [3] and self-adaptive systems [10].

Depending on the domain, some architectural safeguards may be more suitable than others in terms of the non-functional system requirements. For instance, an n-version programming approach positively affects the reliability of the system while causing performance degradation. At design-time, however, it is difficult to anticipate how an architectural safeguard (*i*) reduces the predictive uncertainty of an AI component (and thus improves the overall system reliability) and (*ii*) how it affects other non-functional requirements (e.g. performance), because their effects are only observed at runtime. However, improper selection (with regard to the quality attributes of the system) can lead to economic damage or even to serious malfunctions of the system, which is unacceptable, especially for safety-critical systems.

This dissertation presents a model-based approach to analyse the reliability of AI-enabled systems by systematically considering the predictive uncertainty of AI models (using *Bayesian Modelling*). Moreover, it accounts for architectural safeguards by modelling their effects on the predictive uncertainty of an AI component (and thus on the overall reliability of the system). Because the approach builds upon the *Palladio* approach [4] (providing a repertoire of prediction tools), software engineers can complement the reliability analysis with further analyses (e.g. performance prediction) which greatly supports software engineers in making trade-off decisions.

Contributions. The dissertation encompasses four central contributions:

(C1) Domain-agnostic modelling of AI-specific environments [6, 8]: In the dissertation, a metamodel has been developed which accounts for the modelling of AI-specific uncertainties, their relation to the predictive uncertainty of an AI component and their temporal expansion.

(C2) Reliability prediction of AI-enabled systems [8]: As mentioned in the introductory section, the dissertation builds upon the Palladio approach which provides an existing architectural description language, namely the *Palladio Component Model*, for modelling component-based software architectures and an associated reliability prediction tool (for classical soft-

ware systems). The predictive uncertainty of AI components is represented probabilistically by using Bayesian models and w.r.t. to several uncertainty factors (e.g., brightness variations, sensor noise) directly correlating with the predictive uncertainty. The predictive uncertainty (or its Bayesian representation) is approximated by applying a sensitivity analysis and systematically included in the reliability analysis.

(C3) Evaluation of self-adaptive systems [6, 7]: This contribution generalises the concepts of contribution (C2) to self-adaptive systems acting as dedicated safeguarding mechanisms.

(C4) Classes of architectural dependability assurance [5]: Complementary to contribution (C1)-(C3), the dissertation presents a classification structure for describing the extent to which assurances (w.r.t. a dependability-related system-level property) can be made for a given AI-enabled system.

Validation. The approach was validated using four case study systems. As part of the evaluation, we validated various plausibility properties of the reliability prediction approach using real observations. Moreover, we compared our evaluation framework for self-adaptive systems with two baseline simulators and demonstrated its appropriateness. Ultimately, we discussed and evaluated the applicability of the classification structure (elaborated in contribution (C4)) with a set of representative AI systems.

Outlook. The dissertation presents a reliability prediction approach for AI-enabled systems that takes into account the predictive uncertainty of AI components. Furthermore, the approach enables the evaluation of architectural safeguards considering their effect on the predictive uncertainty and thus on the overall reliability of the system. Although the approach supports software engineers in the decision-making process, it does not provide any formal guarantees. This is, however, a desirable property, especially in safety-critical systems. A promising research direction is therefore to expand the concepts of the dissertation towards formal guarantees (taking into account existing work in the context of verification of AI-enabled systems). Although this is a challenging endeavor, a promising direction is to determine upper and lower bounds on the reliability of an AI system, which is more realistic to accomplish.

References

[1] A. Gujarati, S. Gopalakrishnan, and K. Pattabiraman. New wine in an old bottle: N-version programming for machine learning components. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 283–286. IEEE, 2020.

[2] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[3] P. Musau, N. Hamilton, D. M. Lopez, P. Robinette, and T. T. Johnson. On using real-time reachability for the safety assurance of machine learning controllers. In *2022 IEEE International Conference on Assured Autonomy (ICAA)*, pages 1–10. IEEE, 2022.

[4] R. H. Reussner, S. Becker, J. Happe, R. Heinrich, A. Koziolek, H. Koziolek, M. Kramer, and K. Krogmann. *Modeling and simulating software architectures: The Palladio approach*. MIT Press, 2016.

[5] M. Scheerer, J. Klamroth, R. Reussner, and B. Beckert. Towards classes of architectural dependability assurance for machine-learning-based systems. In *Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, pages 31–37, 2020.

[6] M. Scheerer, M. Rapp, and R. Reussner. Design-time validation of runtime reconfiguration strategies: An environmental-driven approach. In *2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, pages 75–81. IEEE, 2020.

[7] M. Scheerer and R. Reussner. Reliability prediction of self-adaptive systems managing uncertain ai black-box components. In *2021 International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 111–117. IEEE, 2021.

[8] M. Scheerer and R. Reussner. Reliability analysis of architectural safeguards for ai-enabled systems. In *2023 27th International Conference on Engineering of Complex Computer Systems (ICECCS)*, pages 61–70. IEEE, 2023.

[9] Y. Tian, K. Pei, S. Jana, and B. Ray. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering*, pages 303–314, 2018.

[10] Q. Zhu, C. Huang, R. Jiao, S. Lan, H. Liang, X. Liu, Y. Wang, Z. Wang, and S. Xu. Safety-assured design and adaptation of learning-enabled autonomous systems. In *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, pages 753–760, 2021.