Complexity, uncertainty and the Safety of ML

Simon Burton, Benjamin Herd

Complexity, uncertainty and the Safety of ML

Simon Burton
Scientific Director, Safety Assurance
Fraunhofer Institute for Cognitive Systems
Munich, Germany
simon.burton@iks.fraunhofer.de

Benjamin Herd
Safety Assurance of AI
Fraunhofer Institute for Cognitive Systems
Munich, Germany
benjamin.herd@iks.fraunhofer.de

***Abstract*—There is currently much debate regarding whether or not applications based on Machine Learning (ML) can be made demonstrably safe. We assert that our ability to argue the safety of ML-based functions depends on the complexity of the task and environment of the function, the observations (training and test data) used to develop the function and the complexity of the ML models. Our inability to adequately address this complexity inevitably leads to uncertainties in the specification of the safety requirements, the performance of the ML models and our assurance argument itself. By understanding each of these dimensions as a continuum, can we better judge what level of safety can be achieved for a particular ML-based function?**

***Index Terms*—Artificial intelligence, machine learning, safety assurance, uncertainty, complexity**

I. Motivation

Increasing interest in the use of ML to implement perception and planning tasks for safety-relevant cyber-physical systems (e.g. automated driving, industrial robotics) has led to inevitable questions regarding whether or not suitably convincing safety arguments can be made for such systems. As a consequence, the field of trustworthy and safe AI is receiving attention from a regulatory and standards perspectives. Examples of which are the EU proposal for regulations on AI[1] and ongoing standardisation initiatives[2].

We subjectively observe a strong divide between the traditional ML and safety communities. On the one hand, the ML community focuses on solving ever more complex tasks, whose performance is measured in overall accuracy rates. On the other hand, the safety community has relied in the past on the analysis of causal models to determine the underlying causes of individual components faults that may lead to hazardous actions of the system. For any reasonably complex tasks, due to typical properties of ML models such as lack of robustness, bias and prediction uncertainty, a purely statistical testing approach to arguing failure rates required by safety-critical systems is not feasible. Further, the required failure rates (e.g. 10^{-7} failures/hour) are simply not achievable given state-of-the-art perception and planning models for tasks such as automated driving. The causalities of ML errors may be hypothesised in general (e.g. bias in the training data) but are difficult to directly observe or disentangle in order to identify effective safety measures.

We therefore see a need for a *lingua franca* which can be used to exchange concepts between these communities. This position paper summarises and elaborates on the consequences of recent work that proposed a framework for reasoning about uncertainty associated with safety assurance arguments for ML-based functions [1]. Such a framework must be based on a deep understanding of the ML techniques and their underlying theory as well as the ability to formulate convincing safety assurance arguments at the system level. Our aim is to promote discussion and to identify areas of research required to reduce the uncertainty within safety assurance of ML.

II. Semantic gaps, complexity and uncertainty

In [2], the concept of *semantic gaps* was used to express the difficulty of defining an adequately complete set of safe behaviours of an ML-based function. The authors make use of the following definition *"Semantic gap: The gap between intended and specified functionality — when implicit and ambiguous intentions on the system are more diverse than the system's explicit and concrete specification "* [3]. The semantic gap was described as a direct consequence of the following:

- the *complexity and unpredictability of the environment* in which the system operates,
- the *complexity and unpredictability of the system* as well as the system's interactions with other technical systems and human actors (including operators, users, and bystanders), and
- the increasing *transfer of the decision-making responsibility* from a human actor to the system. This can also be considered as an expression of the inherent *complexity and ambiguity of the task* itself.

In [1], we extended these concepts and explored how the resulting uncertainty associated with our understanding of the *environment* and *task*, our *observations* used to develop (train) and evaluate the system and the technical *system* itself (e.g. ML model) can be used to inform the safety assurance task. Furthermore, definitions of types and severity of uncertainty could be used to evaluate the confidence with which arguments and supporting evidence can be proposed for each of these dimensions. The model proposed by the paper is summarised in Figure 1.

This work was performed as part of the ML4Safety project supported by the Fraunhofer Internal Programs under Grant No. PREPARE 40-02702.

[1] artificialintelligenceact.eu

[2] www.iso.org/standard/81283.html, www.iso.org/standard/83303.html

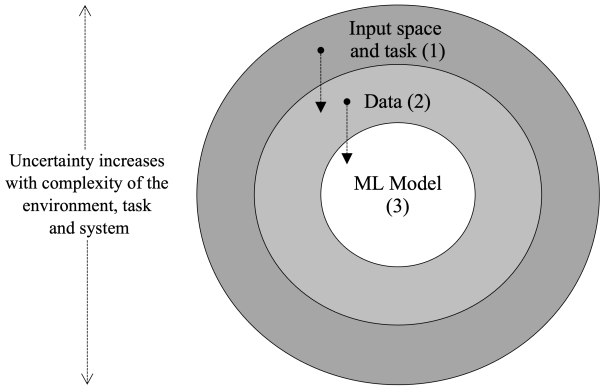

Fig. 1. Dimensions of uncertainty impacting safety assurance of ML

III. Consequences

In [1], we proposed an iterative approach to safety assurance as a means of successively reducing the uncertainty within the assurance argument. In our practical work, e.g. [4], we have shown how a detailed understanding of the environment for a well-defined task, coupled with easy to interpret ML-models acting on a relatively low dimensional input space supported the formulation of a robust argument. Furthermore, when addressing more complex tasks, such as camera-based classification of traffic signs, we discover that the evaluation of errors in the model typically lead to the identification of uncertainties in the data, in turn, possibly exacerbated by a lack of understanding of relevant environmental conditions and the need for more specific evaluation criteria and meaningful metrics [5]. Thus, further promoting an iterative approach to converging on meaningful and convincing arguments.

Ultimately, for safety-critical applications operating within an open-context environment, complexity of the environment, task and system will lead to models with significant residual error rates as well as uncertainty within the assurance arguments. Our ability to make convincing statements about the safety of machine learning functions will depend on how well we can reduce the uncertainty in the dimensions described by Figure 1. Safety assurance must furthermore evaluate the quantity and characteristics of these residual errors in as much detail as possible such that they can be compensated for at the system level, e.g. through restrictions in the operating environment, multi-modal sensor redundancies and run-time evaluation and compensation of known situations which lead to model uncertainty (triggering conditions).

IV. Research perspectives

Given the above observations, we see a number of research questions that, if met, can have a significant impact on our ability to apply ML-technologies to safety-critical applications. These include, amongst others:

- Can complexity in the environment, task and system be defined in a manner that the impact on uncertainty and safety assurance of the resulting ML model be predicted?
- If so, is it possible to decide a priori whether or not a feasible safety argument can be achieved for a given application: given state-of-the-art, availability of suitable data etc. Alternatively, could various applications of ML be ranked according to their relative *assurability*?
- How could concepts from the realm of statistical learning and ML be integrated into a safety assurance framework. For example, the idea of task complexity is closely related to the concept of *learnability* [6] and uncertainty in the observations (data) can be related to *sample complexity*, i.e. the number of samples required for a problem to be efficiently learnable [7].
- Which set of meaningful (from a safety assurance perspective) ML-properties, metrics and target values can be used to directly evaluate the ML model and nevertheless be used to infer system-level safety properties?

The focus on ML-based perception tasks such as camera-based pedestrian recognition for automated driving has motivated much progress in the field of ML safety assurance over the last years. However, from a practical perspective, this may be one of the hardest ML tasks to assure, which has introduced much doubt into the feasibility of applying ML to safety-related tasks in general. We propose gaining experience in less complex ML applications and thus increasing confidence in our approaches to safety assurance to build overall trust in the technology.

The consequences of untrustworthy AI/ML in other fields such as social media, policing and smart cities could have even more widespread consequences due to the scalability of (indirect and therefore unpredictable) harm caused, loss of human agency and discrimination. We believe similar concepts as could also be applied to trustworthy AI in general.

References

[1] S. Burton and B. Herd, "Addressing uncertainty in the safety assurance of machine-learning," *Frontiers in Computer Science*, vol. 5, p. 1132580, 2023.

[2] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artificial Intelligence*, vol. 279, p. 103201, 2020.

[3] C. Bergenhem, R. Johansson, A. Söderberg, J. Nilsson, J. Tryggvesson, M. Törngren, and S. Ursing, "How to reach complete safety requirement refinement for autonomous vehicles," in *CARS 2015-Critical Automotive applications: Robustness & Safety*, 2015.

[4] S. Burton, I. Kurzidem, A. Schwaiger, P. Schleiss, M. Unterreiner, T. Graeber, and P. Becker, "Safety assurance of machine learning for chassis control functions," in *Computer Safety, Reliability, and Security: 40th International Conference, SAFECOMP 2021, York, UK, September 8–10, 2021, Proceedings 40*, pp. 149–162, Springer, 2021.

[5] S. Burton, C. Hellert, F. Hüger, M. Mock, and A. Rohatschek, "Safety assurance of machine learning for perception functions," in *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, pp. 335–358, Springer International Publishing Cham, 2022.

[6] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[7] A. Usvyatsov, "On sample complexity of neural networks," *arXiv preprint arXiv:1910.11080*, 2019.