Uncertainty in Machine Learning Applications

A Practice-Driven Classification of Uncertainty

Michael Kläs[1] and Anna Maria Vollmer[1]

[1] Fraunhofer Institute for Experimental Software Engineering IESE,
Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
{michael.klaes, anna-maria.vollmer}@iese.fraunhofer.de

Abstract. Software-intensive systems that rely on machine learning (ML) and artificial intelligence (AI) are increasingly becoming part of our daily life, e.g., in recommendation systems or semi-autonomous vehicles. However, the use of ML and AI is accompanied by uncertainties regarding their outcomes. Dealing with such uncertainties is particularly important when the actions of these systems can harm humans or the environment, such as in the case of a medical product or self-driving car. To enable a system to make informed decisions when confronted with the uncertainty of embedded AI/ML models and possible safety-related consequences, these models do not only have to provide a defined functionality but must also describe as precisely as possible the likelihood of their outcome being wrong or outside a given range of accuracy. Thus, this paper proposes a classification of major uncertainty sources that is usable and useful in practice: scope compliance, data quality, and model fit. In particular, we highlight the implications of these classes in the development and testing of ML and AI models by establishing links to specific activities during development and testing and means for quantifying and dealing with these different sources of uncertainty.

Keywords: Artificial Intelligence, Dependability, Safety Engineering, Data Quality, Model Validation, Empirical Modelling.

1 Motivation

Systems that make use of models provided by techniques belonging to the domains of machine learning (ML) and artificial intelligence (AI) are becoming increasingly important in our daily life. The terms AI and ML are frequently used interchangeable in this context although differences exist depending on specific definitions. This paper uses the term AI/ML models to refer to computation models trained on empirical data to mimic ‘intelligence’ by transforming inputs to outcomes based on mathematical relationships that are hard to derive by deductive reasoning or simple statistical analysis.

Nowadays, systems that make use of these models do not only recommend movies that we are most likely enjoy [1], but also support the detection of cancer based on images [2] or initiate emergency braking to avoid car crashes [3]. In turn, this means that these models are slowly also becoming a part of safety-relevant systems, where a high risk exists that humans or the environment may be harmed in the case of a failure.

When the relevant existing standards and guidelines for safety-relevant systems (e.g., [4], [5]) were written, however, the usage of AI/ML was not an issue yet and was thus not considered. As a consequence, many techniques proposed in these standards and guidelines appear difficult to apply for systems relying on AI/ML in safety-relevant functions. For example, formal verification techniques cannot be reasonably applied in these kinds of complex models trained on empirical data. Another open question is how to effectively perform mandatory safety reviews for models such as deep convolutional neural networks (CNNs), which have been considered as the state of the art in image recognition since 2012 [6].

Because of the complexity and empirical nature of these models, no guarantee can be provided for their correctness. Thus, a possible consequence could be refusal of the use of these models for safety-critical functions. Traffic sign recognition systems could still provide information to assist human drivers. However, a car would not be allowed to autonomously cross an intersection based on recognized traffic signs and its knowledge of priority rules because it cannot be guaranteed that each stop sign will be recognized correctly in every case.

Alternatives to this strict refusal are being discussed [7][8]; one alternative could be to encapsulate functionality provided by such models and appropriately deal with the inherent uncertainty of their outcomes in the containing system by making use of deterministic and verifiable rules. In this setting, the containing system would be responsible for adequate risk management, taking into account the likelihood that the outcome of the encapsulated model might be wrong, as well as the safety-related consequences of every decision made. In order to allow for informed decisions, encapsulated models would not only have to provide a given service but also describe as precisely as possible the uncertainty remaining in their outcomes. This means that the models would become *dependable* in a figurative sense, according to Avizienis et al.'s definition of systems [9], by delivering their service together with *information about outcome-related uncertainty* that can justifiably be trusted. Based on this information, the containing system could, for example, decide to consider further information sources (as applied in sensor fusion) or adapt its behavior in order to handle the remaining uncertainty adequately. In our scenario of autonomous intersection crossing, the containing system could, for example, use GPS localization as an additional information source or slow down the vehicle, thereby buying time to analyze further images taken of the traffic situation.

At present, model validation and testing commonly focus on determining and optimizing the overall accuracy of the created model (cf. KAGGLE competitions, e.g., the ImageNet Challenges[1]). However, the models' accuracy, which is commonly measured as *error rate*, is only a very generic and therefore weak estimator for the uncertainty remaining in a specific outcome, which is commonly referred to as *prediction uncertainty*. For instance, an error rate of 0.54 % on a test dataset of traffic sign images [10] indicates that the respective model is 99.46 % confident of providing a correct outcome or, conversely, that it is 0.54 % uncertain *on average*. However, for use in safety-relevant functions, this general statement is likely too coarse-grained. Although a reported

[1] https://www.kaggle.com/competitions

accuracy of 99.46 % is excellent, autonomous vehicles simply ignoring one of two hundred stop signs might not be considered sufficiently safe. To be useful, more precise prediction uncertainty estimates are required that consider the situation at hand. For instance, fog or backlight conditions may affect the confidence in the provided outcomes, as may dirt on the camera lens. Moreover, the question needs to be answered of whether the test dataset on which the accuracy of the model was determined matches the situation in which the model is currently being applied.

In order to consider such sources of uncertainty during model development and testing more systematically, an applicable framework and associated terminology would be needed in practice. We especially see the need for a practice-driven classification of the different sources of uncertainty that have to be addressed and quantified. Thus, this position paper proposes a sound and usable schema for classifying uncertainty sources that are relevant in AI/ML models. The main practical benefit is seen in establishing clear links between specific sources of uncertainty and activities performed during model development and testing, and thus the possibility to define concrete means for quantifying and dealing better with the various sources of uncertainty.

The paper is structured as follows: Section 2 provides a short overview of existing classifications of uncertainty. Section 3 introduces the proposed classification, illustrates its application on an example, and discusses its implications. Section 4 closes the paper with an outlook on next steps.

2 Related Work: Existing Classifications of Uncertainty

In general, uncertainty is interpreted as “what is not known precisely”, but it can be characterized differently, e.g., by also considering its impact or causes. Thus, various taxonomies and classifications of uncertainty exist that provide different points of view on uncertainty, such as aleatoric vs. epistemic, irreducible vs. reducible, or the different kinds of inference that introduce them (e.g., predictive, statistical, or proxy) [11].

Mahdavi-Hezavehi et al. present a literature review and overview of different uncertainty studies in the context of system architecture including uncertainty classifications comprising the dimensions location, nature, sources, and level/spectrum [12]. Furthermore, they propose a classification based on the source’s *model* (uncertainty caused by system models due to their abstraction, model drift, incompleteness, complexity, etc.), *goals* (uncertainty caused by a system’s goal-related complications such as outdated goals, goal dependencies, future goal changes, etc.), and *environment* (uncertainty caused by environmental circumstances including execution context, multiple ownership, human involvement). Other uncertainty dimensions reported in the literature consider *resources* (changing or new ones) or *adaptation functions* (automatic learning, sensing, decentralization, etc.). Further uncertainty types are reported by another study considering different publications: *content*, *environment*, *geographical location*, *occurrence*, and *time* [13].

A detailed classification is provided in the context of simulation models by Kennedy and O’Hagan, who distinguish between *parameter*, *parametric*, *structural*, *algorithmic*, *experimental*, and *interpolation* uncertainty [14].

For safety-critical ML applications, Faria distinguishes between sources of output variation on the levels *experience*, *task*, *algorithm*, *implementation*, and *hardware* [15].

All these classifications can help to get a better understanding of the various aspects of uncertainty and may support practitioners in identifying important sources of uncertainty in their context. However, they are hard to apply effectively in practice for rigorous prediction uncertainty quantification because their boundaries are not sharp (e.g., aleatoric vs. epistemic), because they cannot be reasonably quantified and distinguished in a practical AI/ML setting (c.f. the detailed classes of Kennedy and O'Hagan [13]), or because they have no direct links and implications for model building and testing.

3 Sources of Uncertainty in AI/ML-Model Applications

In order to introduce our classification, this section first illustrates common activities in the model building and testing process based on an example. Then the proposed classification is derived by highlighting and grouping major sources of uncertainty that occur in this process. Finally, implications of the proposed classification are discussed.

3.1 Typical Process of Model Learning and Application

Most development and testing of AI/ML models more or less explicitly follows an adaptation of the CRISP-DM [16] approach, which was initially introduced by IBM and comprises the steps business/domain understanding, data understanding, data preparation, modeling, evaluation, and deployment. Next, we summarize the key activities in the process that are relevant for uncertainty in the model application outcomes and illustrate them with an ongoing example.

Based on a specific goal or problem statement, the planned *scope* of the model application is defined. In our example, the scope could be traffic sign recognition in all possible driving conditions of a passenger car on public roads in the target market Germany. Based on the scope definition, *raw data* is gathered in the relevant context to build and test the AI/ML model. In our example, such data could be images taken by cameras in pilot cars driving through Germany for several months, or an existing dataset such as the GTSRB dataset[2] with more than 50,000 traffic sign images. In the next step, the raw data is typically filtered and preprocessed before being used as *cleaned data* to build and test the model because depending on the data source, raw data may suffer from various quality issues that would affect the final accuracy of the model. Moreover, preprocessing of the data makes them more accessible in modeling. In our example, images with specific problems could be filtered, such as very dark images, images with strong backlight conditions or massive lance flares, or blurred images. Preprocessing techniques include, among others, image normalization, Contrast Limiting Adaptive Histogram Equalization (CLAHE) [17], and Single-Image Super-Resolution (SISR) [18]. The clean data is separated into modelling and test data, with the *modelling data* being used to build and cross-validate the AI/ML model and the *test data* to evaluate

[2] http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset

the final model and determine how well the model fits previously unseen data. If the *model* is considered to be sufficiently accurate, it is deployed to productive use (e.g., as part of a driver assistance system) annotated with its *error rate* (e.g., 0.54 %).

3.2 A Practice-Driven Classification of Uncertainty

If we agree on the definition of prediction uncertainty as the likelihood that the provided outcome of a model may be wrong or outside a given range of accuracy, based on the model building and testing process, three major sources of uncertainty can be identified: scope compliance, data quality, and model fit. As we will discuss, the three uncertainty categories are stacked on top of each other; Fig. 1 presents an onion layer model.

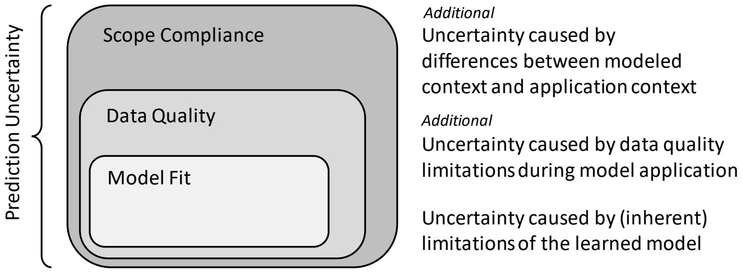
Prediction Uncertainty
Scope Compliance
Data Quality
Model Fit
Additional
Uncertainty caused by differences between modeled context and application context
Additional
Uncertainty caused by data quality limitations during model application
Uncertainty caused by (inherent) limitations of the learned model

Fig. 1. Onion layer model of uncertainty in AI/ML application outcomes.

Model fit. Uncertainty related to model fit is caused by the fact that AI/ML techniques provide empirical models that are only an approximation of the real (functional) relationship between the model input and its outcome. The accuracy of this approximation, which is limited, e.g., due to the limited number of model parameters, input variables considered, and data points available to train the model, represents the model fit. Uncertainty caused by fitting deficits is commonly measured by the error rate, which is calculated when spitting the cleaned data into a training dataset and a test dataset.

There are two important underlying assumptions regarding uncertainty caused by model fit. (1) The model is applied in a setting that is appropriately represented by the test dataset, which is true for the cleaned dataset from which the test data is typically randomly selected. (2) The model is built and applied on data on a homogeneously high quality level (i.e., the data does not suffer from quality issues), which is also more likely in a well-cleaned dataset.

Implication: The average uncertainty caused by model fit can be measured using standard model evaluation approaches applied on high quality data and can be seen as a lower boundary approximation of the remaining uncertainty.

Data quality. In a real setting, all kinds of data collected (e.g., based on sensors but also human input) is limited in its accuracy and potentially affected by various kinds of quality issues. The actual level of uncertainty in the outcome of an AI/ML model is thus affected by the quality (especially the accuracy) of the data on which it is currently applied. Therefore, additional uncertainty that is the result of a delta between the quality of the cleaned data and the data on which the model is currently being applied can be defined as data quality (caused) uncertainty. In our example, confidence in the model outcome may be affected by a camera with lower resolution or a damaged lens as well as by difficult lighting and bad weather conditions such as rain or fog.

Implication: Dealing with data quality uncertainty requires extending the standard model evaluation procedures with specialized setups to investigate the effect of different quality issues on the accuracy of the outcomes in order to provide uncertainty adjustments for cases where the model is applied on data of below-nominal quality. As a consequence, data quality has be quantified and measured not only to annotate raw data with quality information during data preparation, but also to measure the current quality of the data after the model is deployed.

Scope compliance. As we have seen, AI/ML models are built for and tested in a specific context. If these models are applied outside this context, their outcomes can become unreliable (e.g., because the model has to extrapolate). Therefore, the likelihood that a model is currently being applied outside the scope for which it was tested can be defined as scope compliance (caused) uncertainty. In our example, the confidence in the outcome of the model would be heavily affected if the model, which was trained and tested on German traffic signs, were applied in a country that does not follow the Vienna Convention on Road Signs and Signals. Moreover, if the raw data used for model development and testing were collected between May and October, the test dataset would most likely miss images of traffic signs (partially) covered by snow.

Implication: Scope compliance uncertainty can stem from two sources, as illustrated in the example: The model may be applied outside the intended scope or the raw dataset might not be representative of the intended scope. The former can be detected by monitoring relevant context characteristics (in our example, e.g., GPS location, velocity, temperature, date, time of day) and comparing the results with the boundaries of the intended scope. In order to reason about the latter, raw data needs to be annotated with context characteristics (e.g., GPS location, velocity, temperature, date, time of day) in order to compare its actual and assumed distribution in the intended scope.

4 Conclusion

Distinguishing between the three types of uncertainties presented (*model fit*, *data quality*, *scope compliance*) is motivated from a practical point of view because each of these types requires specific means for detecting the related uncertainty and coping with it.

Furthermore, the classification builds on existing ones and enables further focused uncertainty analysis by providing three clearly separated and measurable constructs. In a first approximation, prediction uncertainty can be determined by adjusting *model-fit-caused uncertainty* with a *data quality factor* determined on the basis of the quality of the current input data and the probability of *scope compliance*.

Building upon this classification, we plan to provide a practical framework for capturing uncertainty when building and testing AI/ML models. Additionally, we are planning its evaluation in case studies to demonstrate its applicability and usefulness.

Acknowledgments. Parts of this work is funded by the German Ministry of Education and Research (BMBF) under grant number 01IS16043E (CrESt).

References

1. Carrer-Neto, W., Hernández-Alcaraz, M. L., Valencia-García, R., García-Sánchez, F.: Social knowledge-based recommender system. Application to the movies domain, Expert Systems with Applications, vol. 39, no. 12, pp. 10990-11000 (2012).
2. Sirinukunwattana, K., Raza, S. E. A., Tsang, Y. W., Snead, D. R. J., Cree, I. A., and Rajpoot, N.M.: Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. In: IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1196-1206. (2016).
3. Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C. and Winner, H.: Three Decades of Driver Assistance Systems: Review and Future Perspectives. In: IEEE Intelligent Transportation Systems Magazine, vol. 6, no. 4, pp. 6-22. (2014).
4. ISO 26262: Road vehicles – Functional safety.
5. IEC 61508 ed. 2, Functional safety of electrical/electronic/programmable electronic (E/E/PE) safety related systems.
6. Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Commun. ACM 60(6), 84-90 (2017).
7. Varshney, K. R.: Engineering safety in machine learning. Information Theory and Applications Workshop (ITA), La Jolla, CA, pp. 1-5. (2016).
8. Rasmus, A., Feth, P., Schneider, D.: Safety engineering for autonomous vehicles. IEEE/IFIP Int. Conf. on Dependable Systems and Network Workshops 2016. pp. 200-205. (2016).
9. Avizienis, A., Laprie, J. C., Randell, B. and Landwehr, C.: Basic concepts and taxonomy of dependable and secure computing. In IEEE Transactions on Dependable and Secure Computing, vol. 1, no. 1, pp. 11-33. (2004).
10. Ciregan, D., Meier, U. and Schmidhuber, J.: Multi-column deep neural networks for image classification. IEEE Conference on Computer Vision and Pattern Recognition 2012, pp. 3642-3649. (2012).
11. Booker, J. M., Ross, T. J.: An evolution of uncertainty assessment and quantification. In: Scientia Iranica, vol. 18, no. 3, pp. 669-676 (2011).
12. Mahdavi-Hezavehi, S., Avgeriou, P., Weyns, D.: A Classification Framework of Uncertainty in Architecture-Based Self-Adaptive Systems With Multiple Quality Requirements. In Managing Trade-offs in Adaptable Software Architectures 1, pp. 45–78 (2017).
13. Zhang, M., Selic, B., Ali, S., Yue, T., Okariz, O., Norgren, R.: Understanding Uncertainty in Cyber-Physical Systems: A Conceptual Model. In: A. Wąsowski and H. Lönn (Eds.): ECMFA 2016, Springer LNCS 9764, pp. 247-264. (2016).
14. Kennedy, M. C., O'Hagan, A.: Bayesian calibration of computer models. In: Journal of the Royal Statistical Society, vol. 63 no. 3, pp. 425-464. (2001).
15. Faria, J. M.: Non-determinism and failure modes in machine learning. IEEE International Symposium on Software Reliability Engineering Workshops. 2017, pp. 310-3016- (2017)
16. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. Int. Conference on the Practical Applications of Knowledge Discovery and Data Mining. (2000).
17. Yadav, G., Maheshwari, S. and Agarwal, A.: Contrast limited adaptive histogram equalization based enhancement for real time video system. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2392-2397. (2014).
18. Glasner, D., Bagon, S. and Irani, M.: Super-resolution from a single image. In: IEEE 12th International Conference on Computer Vision, Kyoto, pp. 349-356, (2009).