Complementary Discussion for Challenge Classification to main paper "The Road to Safe Automated Driving Systems: A Review of Methods Providing Safety Evidence"

Magnus Gyllenhammar, Gabriel Rodrigues de Campos, and Martin Törngren, *Senior Member, IEEE*

***Abstract*—This is a complementary document to the main paper *"The Road to Safe Automated Driving Systems: A Review of Methods Providing Safety Evidence"*. In this document some additional details pertaining to each methods' ability to address the eight challenges are provided. The ensuing sections include explicit motivation for the challenge classifications not explicitly given in the main paper. Thus, jointly with the main paper the information herein provides the basis for the classification presented in TABLE VI of the main paper.**

I. Challenges

This is a recap of the challenges identified in the main paper:

Uncertainties:

C-U-env **Uncertainties** associated with the **operational environment** of the ADS,

C-U-inter **Uncertainties** originating from the **interaction** of the ADS with other traffic participants,

Behavioural and structural complexity:

C-B-resp **ADS's responsibility** for all strategic, tactical and operational decisions of the driving task,

C-B-func **Complex set of interwoven functions** and sub-systems required to realise an ADS,

C-B-adapt **Self-adaptation capabilities** of the ADS, in particular, to cope with (temporary) degradations of the system,

Dependability requirements:

C-reqs **High dependability requirements** imposed on the system, e.g. originating from a comparison with human performance, highlighting the contribution of corner and edge cases to the overall safety,

AI and ML components

C-AI **Validation of (black box) components** relying on Artificial Intelligence (AI) and Machine Learning (ML),

Agile development and continuous deployment

C-agile **Frequent releases and continuous learning**, with to a shift to an iterative and agile development process including software upgrades, requiring reduction of safety/assurance case compilation efforts (or re-verification and re-validation of the system).

II. Complementary discussion

The following subsections collect the remaining discussion, complementing that given in the main paper, such that all the cells of the challenge classification of TABLE I of the main paper are explicitly motivated.

A. Operational Design Domain (ODD)

The use of an ODD could help alleviate some difficulties to challenges (C-B-resp), (C-B-func) and (C-reqs) by providing clear boundaries for where the function will be operational. Similarly, also relying on ML-based AI components (related to challenge (C-AI)) might be ameliorated by concretely defining what is inside the operational domain, which could privide a formalised means for out-of-distribution detection.

B. Hazard Analysis and Risk Assessment (HARA)

As the current HARA process relies on mainly manual work, it would be challenging to match with high cadence releases within an agile development process and the impact from incorporation of new evidence through continuous learning is also unclear (i.e. challenge (C-agile)). However, the process itself does not need to be manual, but a solution to overcome that is yet to be defined. As for challenge (C-reqs), regarding the high dependability requirements, it is currently difficult to assess how the HARA will be able to ameliorate this aspect considering the diversity of relevant events to consider. After the first successful deployment of an ADS, and once sufficient operational data becomes available, this might, however, no longer present itself as a challenge, as the completeness of scenarios and events underpinning the HARA could then be deduced from the collected data itself, e.g. following the approach of [1].

M. Gyllenhammar and G. Rodrigues de Campos are with Zenseact

M. Gyllenhammar and M. Törngren are with the Mechatronics division at KTH, Royal Institute of Technology

The research has been supported by the Strategic vehicle research and innovation (FFI) programme in Sweden via the SALIENCE4CAV project (ref 2020-02946), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Swedish Innovation Agency (Vinnova, through the Entice project and the TECoSA centre for Trustworthy edge computing systems and applications).

C. (Qualitative) Process Arguments

Similar to the HARA, process arguments struggle with encapsulating the uncertainties imposed on the ADS through challenges (C-U-env) and (C-U-inter). Further, challenges (C-B-resp), (C-B-func), (C-B-adapt), and (C-reqs) could, in principle, be supported by process arguments, though the quantitative contributions would then need to be better understood. Further, how to merge process arguments and traditional development processes with agile development (to address challenge (C-agile)) remains an open challenge [2].

D. Contract-Based Design (CBD)

As CBD provides a clear interface between system elements, it effectively supports modularity of the elements and consequently ameliorates challenge (C-agile), pertaining to frequent releases and continuous learning. However, the complexity of the ADS (formulated in challenge (C-B-func)) raises the question of the scalability of the approach for highly interwoven functionalities and complex systems. Further, the defined contracts require formalised specification of the assumes and guarantees, which, considering challenges (C-U-env) and (C-U-inter), might be difficult to achieve on the boundary of the ADS towards its environment. When using machine learning-based black-box components (e.g., challenge (C-AI)), the difficulty of applying contracts is even greater, as small perturbations to the inputs might lead to large perturbations in the output [3], which requires highly dependable systems for monitoring and out-of-distribution detection. Finally, encoding formalised contracts for the tactical decisions of the ADS (related to challenge (C-B-resp)), such that they fulfil the safety requirements, represents a considerable challenge.

E. Supervisor Architectures

While the vast number of possible degradations (related to challenge (C-B-adapt)) pose a challenge for the implementation of an appropriate supervisor architecture, supervision is the only solution allowing for an appropriate adaptation to manage the different degradations of the system. Further, supervisor architectures help tackling the first two categories of challenges, namely the uncertainties and the behavioural and structural complexities. Moreover, by deploying anomaly and OoD detection, ML-based components (related to challenge (C-AI)) can be supervised. Lastly, appropriate supervision capabilities might ease the requirements on the assurance efforts done before deployment of the system, and thus ameliorate challenge (C-agile) by reducing time-to-market for each ADS version.

F. Field Operational Tests (FOTs)

The feasibility of FOTs as a V&V method is questionable considering the higher cadence releases resulting from an agile development process, or the sought continuous learning cycle of the system, pertaining to challenge (C-agile).

G. Extreme Value Theory (EVT)

As the results from an EVT analysis are only as good and representative as the data used, EVT faces the same challenges as FOTs with respect to ensuring safety while testing the system in closed-loop (i.e. to assess challenges (C-U-inter) and (C-B-resp)). It is also paramount that the collected data (e.g. through the FOTs) is representative of the actual operating conditions of the ADS. In terms of challenge (C-agile), pertaining to agile development and accommodating continuous learning, the EVT approach helps ameliorating parts of these challenges by leveraging collected data to infer the systems performance level beyond the operational hours used to collect it. However, the reliance on data is detriment to the method's ability to support frequent releases and depending on the collection method this might impose an insurmountable challenge.

H. Scenario-Based Verification and Validation Methods

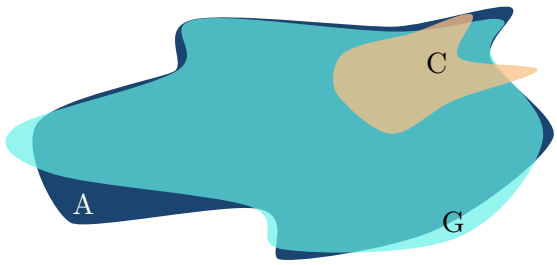

Fig. 1. Illustration of the "scenario space". A corresponding to all possible scenarios in the intended operational design domain of the ADS, G the identified scenarios, and C the safety critical scenarios for the system.

Placing all tactical responsibility on the ADS (i.e. challenge (C-B-resp)) makes the use of scenario-based testing difficult since the ADS might take actions to avoid the initial state of the scenario altogether, rendering the testing results irrelevant (again, corresponding to $G \cap \bar{A}$ of Fig. 1). The complexity of the system and its ability to handle degradations (challenges (C-B-func) and (C-B-adapt) respectively) could, on the other hand, be efficiently validated through scenario-based methods, as they scale according to the testing environments used. Further, the use of simulations helps executing testing and verification quicker and safer than real-time, thus supporting high release cadences (challenge (C-agile)). Additionally, this aspect enables efficient testing of large changes to the system. This said, validation of the models and tools used for simulation itself still impose significant challenges. In particular, model validation includes choices of adequate abstraction levels and level of detail of the various subsystems and the environment for simulation purposes. Further, these choices need to be verified and validated such that the simulation results provides appropriate evidence for the V&V purposes. Moreover, multiple models and simulation compositions will be needed for different purposes.

I. Formal Methods

Despite the merits and advantages of formal methods, one can identify several potential difficulties and limitations with

respect to the safety assurance of an ADS, as detailed below. Indeed, it may be difficult to:

- construct a complete specification with respect to the real operating environment of the ADS, related to challenges in the categories of uncertainty, and behavioural and structural complexity, which may also be exacerbated by challenge (C-reqs),
- scale such methods to cover the entirety of the ADS, corresponding to challenge (C-B-func),
- ensure validity of the verification with respect to the specification when using AI/ML-based components, corresponding to challenge (C-AI), and
- ensure the correctness of the models and parameter values used in conducting the verification, which again stem from the challenge categories of uncertainty, and behavioural and structural complexity.

On the other hand, the successful application of formal verification methods would provide an efficient means to ensure the safety of the ADS and thus support high release cadence as well as continuous learning (i.e. challenge (C-agile)). Formal methods might also help analysing and understanding challenge (C-B-adapt), i.e. the capabilities of degraded modes of the system, and how the system can safely adapt to cope with such changes. Within a well defined setting, and for a limited component, subsystem or specification, it should be noted that formal methods are both suitable as well as useful. However, as noted in [4], any evidence supplied from formal methods should ideally be accompanied by the applicability of that evidence, as well as the formal method/tool used, whenever incorporated into the assurance case.

J. Operational Data Collection

The main purpose of monitoring leading (safety) indicators is to contain the residual risk related to challenges (C-U-env), (C-U-inter), (C-B-resp), (C-B-func), and (C-B-adapt). This type of indicator monitoring could be further enhanced by the use of EVT modelling of the SPIs/KPIs, as suggested in [5], helping to compare sparse data to the high dependability requirements of the system (challenge (C-reqs)).

K. Threat Assessment (TA) Techniques

The intention of the threat metrics is to assess the current threat or risk faced by the system during its operations. However, accurate modelling and assessment incorporating the uncertainties of challenges (C-U-env) and (C-U-inter) into the metrics remain difficult. The metrics themselves provide support for the operational decisions of the ADS (related to challenge (C-B-resp)) and also give a quantitative assessment of the risks irrespective of the complexity of the system, i.e. bypassing challenge (C-B-func).

The focus of most TA metrics is to assess the risk imposed on the vehicle from its external environment. However, integral to these assessments are the capabilities of the ego vehicle. For example, the Brake-Threat-Number (BTN) or Time-To-Collision (TTC) [6] both incorporate the vehicle's braking capability, or at least an estimate thereof. Consequently, the TA could partially support the understanding of the necessary adaptation needed to cope with system degradations related to challenge (C-B-adapt). If coupled with EVT modelling, TA could provide a way to ameliorate the high dependability requirements of challenge (C-reqs), but accurately considering the probabilities of rare events might simultaneously reduce the accuracy of the TA and thus suggesting that these same requirements might pose an obstacle to TA. If the calculation of the metric is reliant on AI/ML-based component, the TA will also be difficult to validate, corresponding to challenge (C-AI). However, if that is not the case, the use of metrics might provide a means to validate the AI/ML-based components in the system based on operational data. The frequency of releases of the ADS (challenge (C-agile)) will not constitute an obstacle for deploying an appropriate TA. Furthermore, such TA techniques would also not be of any particular help when it comes to improving the release cadence of the ADS, except as a means to reduce potential negative impact through monitoring and triggering whenever the system gets into hazardous situations. Continuous data gathering and analysis of the trends of the resulting TAs could however support a learning cycle, also related to challenge (C-agile).

L. Out-of-Distribution (OoD) Detection

OoD detection is believed to be helpful in increasing the reliability of AI/ML-based components and consequently address challenge (C-AI). As the components are ensured to operate within the set of samples known to the algorithm, one can also rely on the validation results provided (based on samples from the same set). The environment uncertainties faced by the ADS, formulated within challenge (C-U-env), might be partly mitigated by the use of OoD detection whereas the uncertainties originating from the interactions, concerning challenge (C-U-inter), are unrelated to the use of an OoD detection method. The challenges of behavioural and structural complexity, challenges (C-B-resp), (C-B-func), and (C-B-adapt), are also not applicable. Even though OoD detection methods might support the validation of AI/ML-based components, ensuring sufficient integrity of the OoD detection itself will be challenging and, as a consequence, the high dependability requirements, formulated within challenge (C-reqs), will present obstacles. As OoD detection methods will have to be trained on the same data as the AI/ML-based components nothing does per se hinder frequent development. In some sense, updating the OoD detection alongside the AI/ML-based components might even be seen as supporting a learning cycle of the system. However, since such updates would be inherent and needed upon any updates to the training data used for the AI/ML-based components, this activity would also require resources upon each update. Notably, if online (in-vehicle) learning is used for the AI/ML-based component this would also require online resources for updating the OoD detection.

M. Dynamic Risk Assessment (DRA)

When approaching a solution to DRA, the more factors and parameters included, the more refined model for situation

awareness can be achievable. However, the problem is that the more factors included, the more data is needed to develop the models that are used to realize the situational awareness. Thus, we face the challenge of state space explosion due to the uncertainties of the operational space of the ADS (related to the challenges (C-U-env) and (C-U-inter)), as discussed in the main paper in terms of the V&V methodologies. However, the ability for a DRA method to handle uncertainties in run-time, formulated within the challenges (C-U-env) and (C-U-inter), is completely dependent on the metrics and models used.

The flexibility of DRA seem to lend itself well to address challenge (C-agile), where, for example, the models underpinning the DRA can be easily updated provided that more operational data becomes available. However, when trying to achieve the desired dependability (challenge (C-reqs)), the question is how to show the reliability of such methods, especially if such needs to be done before the first deployment. Further, given that much of the perception of an ADS and the subsequent construction of its situation awareness are reliant on ML-based algorithms, the question is also how to connect the outputs thereof to the risk estimates of the DRA, an example related to challenge (C-AI). This aspect is especially prominent for the intention prediction task. The complexity of the system (formulated in challenge (C-B-func)) can somewhat be circumvented by using DRA, as the down-stream decision-making can be done in run-time based on the outputs of the relatively less complex DRA system. However, how well these methods are able to accommodate degradations (challenge (C-B-adapt)) is still an open question. Finally, even though an accurate risk assessment at the present time is available through DRA, how to account for the subsequent impact of the tactical decisions of the ADS (i.e. challenge (C-B-resp)) has not yet been discussed in the literature.

N. Degradation Strategies

A Minimal Risk Condition (MRC) effectively ameliorates the impact of foreseeable changes of the uncertainties related to challenges (C-U-env) and (C-U-inter). That is, when it is possible to assess that the operational context suggests uncertainties outside those specified in the ODD, for example, the MRC could be invoked to avoid the associated risks. However, in some cases, such shifts in the uncertainty might not be detected early enough as to let the ADS avoid an accident. Further, finding an appropriate MRC considering the operational uncertainties could be a non-trivial problem.

The use of degradation strategies does neither ameliorate nor support challenge (C-B-resp). The self-adaptation capabilities of the ADS, related to challenge (C-B-adapt), are highly reliant on appropriate MRCs, and the performance and utility of the system can significantly be improved through the use of RODs. The complexity of the ADS, pertaining to challenge (C-B-func), makes the use of a degradation strategy, such as the ROD, more difficult due to the large number of parameters to be considered. The degradation strategies might help mitigate faults or errors in the system and avoid catastrophic outcomes, consequently supporting the achievement of high dependability requirements, related to challenge (C-reqs). If coupled with anomaly detection methods for AI/ML-based components, such as the OoD detection methods, the degradation strategies might support the safe handling of situations problematic to AI/ML-based components. In other words, the use of degradation strategies provides a partial support for solving challenge (C-AI). As for agile development processes/methodologies, related to challenge (C-agile), an update to the ODD or any architectural changes to the ADS would warrant an update of both the MRCs and the RODs. Such updates would require significant efforts and analyses before being deployed, which might inhibit frequent releases of the software, as most of such analysis could likely not be completely automated and would therefore be time-consuming.

O. Run-time Certification

Even though run-time certification mitigates the state space explosion related to challenges (C-U-env), (C-U-inter) and (C-B-resp), the approach faces the same issues as contract-based design. Is it possible to create contracts (e.g. ConSerts) that adequately capture the uncertainties related to challenges (C-U-env) and (C-U-inter), and the flexibility emanating from challenge (C-B-resp)? Similarly, the scalability of run-time certification in the light of challenges (C-B-resp) and (C-B-func), remains to be shown. However, by formalising the interfaces between subsystems and components of the ADS, such methods have the potential to effectively provide a means to support high cadence releases and continuous learning, i.e. challenge (C-agile), similar to contract-based design. Lastly, given appropriate measures and monitors for anomaly and OoD detection, run-time certification approaches might help growing the trust in ML-based components (challenge (C-AI)), as the usage of such components can be adapted given the fulfilment of their respective demands.

P. Dynamic Safety Management (DSM)

Deferring the assurance of the tactical decisions to run-time is proposed in order to ameliorate the effects of the operational uncertainties related to challenges (C-U-env) and (C-U-inter). The considerable complexity of the required risk assessment techniques to support DSM seem to be exacerbated by the ADS's responsibility for tactical decisions, i.e. challenge (C-B-resp). However, the impact from the complexity of the ADS itself (challenge (C-B-func)) could be ameliorated, as argued for in the DRA section of the main paper, by relying on a relatively less complex system for DRA/DSM. Degradation capabilities (related to challenge (C-B-adapt)) would not only be solved through DSM but could further support a more elaborate handling of any type of degradation, effectively providing understanding for the RODs of each degradation of the system. However, this presumes appropriate self-awareness capabilities of the system.

By being highly dependent on the DRA methods, the lack of proof of reliability on the part of the DRA approaches is also inherited by the DSM, making it difficult to assure the reliability of the resulting actions of the ADS. Consequently, it might be difficult to quantify and assure the DSM method before deployment, at least when comparing to the high

dependability requirements related to challenge (C-reqs). This would be exacerbated if one is reliant on AI/ML-algorithms for the implementation of either the DRA or the DSM, in which case, the validation of such components would impose an obstacle, i.e. related to challenge (C-AI). However, the DSM might also provide a means to rely on AI/ML-components for path planning, as the risk of each generated path could effectively be assessed through DRA, see e.g. [7].

Assuming that the models underpinning the DRA and DSM are updated based on collected operational data, they would promote a learning cycle of the system (corresponding to challenge (C-agile)). Further, having assured a method for DSM would also support the frequent changes of other components in the system (the first aspect of challenge (C-agile)), as suggested in [8].

Q. Precautionary Safety (PcS)

While the PcS methodology alleviates some of the restrictions of a worst-case design-time assumption with respect to the operational uncertainties of the ADS (challenges (C-U-env) and (C-U-inter)), the scalability aspects has not been exhaustively addressed. Thus, the ability for this method to overcome challenge (C-B-func) is still an open question. The approach of [9] suggests that challenges (C-reqs) and (C-AI) could be overcome, but at the (initial) expense of a reduced performance of the system. Furthermore, placing the tactical responsibility with the ADS (challenge (C-B-resp)) might in turn impact the event exposure rates, which are the central tenet of the PcS method. Consequently, it remains unclear how well PcS could work for releases of a specific ADS (version) without considerable closed-loop data from that specific version of the system. How a design methodology based on precautionary principles can help support agile development and frequent releases of challenge (C-agile), also remains to be seen. However, the incorporation of operational data, as suggested in [9], suggests that this methodology could support the continuous learning aspect related to challenge (C-agile).

REFERENCES

[1] B. Kramer, C. Neurohr, M. Büker, E. Böde, M. Fränzle, and W. Damm, "Identification and quantification of hazardous scenarios for automated driving," in *Int. Symposium on Model-Based Safety and Assessment*. Springer, 2020.

[2] J.-P. Steghöfer, E. Knauss, J. Horkoff, and R. Wohlrab, "Challenges of scaled agile for safety-critical systems," in *Int. Conf. on Product-Focused Software Process Improvement*. Springer, 2019.

[3] 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 *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2018.

[4] E. Denney and G. Pai, "Evidence arguments for using formal methods in software certification," in *Int. Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE, 2013.

[5] D. Åsljung, C. Zandén, and J. Fredriksson, "A Risk Reducing Fleet Monitor for Automated Vehicles Based on Extreme Value Theory," *techrxiv*, 5 2022.

[6] D. Åsljung, J. Nilsson, and J. Fredriksson, "Using extreme value theory for vehicle level safety validation and implications for autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 4, pp. 288–297, 2017.

[7] M. Gyllenhammar and H. Sivencrona, "Path planning in autonomous driving environments," Mar. 24 2022, US Patent App. 17/477,906.

[8] ——, "Risk estimation in autonomous driving environments," Mar. 24 2022, US Patent App. 17/477,920.

[9] M. Gyllenhammar, G. Rodrigues de Campos, F. Sandblom, M. Törngren, and H. Sivencrona, "Uncertainty aware data driven precautionary safety for automated driving systems considering perception failures and event exposure," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2022.