Preprint version of:
Gyllenhammar et al. "A Safety Argument Fragment Towards Safe Deployment of Performant Automated Driving Systems" in
B. Gallina et al. (Eds.): SAFECOMP 2025, LNCS 15954, 2025.
https://doi.org/10.1007/978-3-032-02018-5_15

A Safety Argument Fragment Towards Safe Deployment of Performant Automated Driving Systems

Magnus Gyllenhammar[1,2][0000−0001−9020−6501], Gabriel Rodrigues de Campos[1][0000−0002−8218−6915], and Martin Törngren[2][0000−0002−4300−885X]

[1] Zenseact, Lindholmspiren 2, 417 56 Gothenburg, Sweden
magnus.gyllenhammar@zenseact.com
[2] KTH Royal Institute of Technology

Abstract. In this paper we present a safety argument fragment to contribute towards solutions to several key factors of relevance towards deployment of safe Automated Driving Systems (ADSs). Firstly, we address the need for exhaustive safety requirements by considering vehicle level, quantitative safety requirements. Secondly, situation awareness is employed to dynamically adapt the ADS' decision-making. Thirdly, the ADS' situation awareness is extended with constraints following Precautionary Safety (PcS) principles to ensure the fulfilment of the quantitative safety requirements. Fourthly, the models and assumptions supporting steps two and three are ascertained through the use of an operational design domain, which the ADS is designed to operate within. Furthermore, the paper contrasts the proposed argument with the state of the art in safety assurance to identify the key challenges still remaining.

Keywords: Safety Argument · Automated Driving Systems · Safety Assurance · Research Gaps · Situation Awareness · Precautionary Safety.

1 Introduction

In the transition to Automated Driving Systems (ADSs), approaches and methods for safety assurance that have once been effective for previous generations of automotive systems, and other safety-critical applications, are no longer practical, efficient (in terms of time and resources) nor performant [11]. This not only halters the deployment of highly automated systems, such as ADSs, but also inhibits frequent software releases, which is not only a business opportunity but also a safety imperative [20]. For example, while formal methods provide a crucial piece of the puzzle for safety assurance, they cannot offer the panacea that fully solves the safety concerns for an ADS [11,26]. Indeed, worst-case assumptions of formal methods can result in suboptimal performance, and assuming rule-following on the part of other traffic participants might, instead, render the system unsafe [27]. What is needed is therefore a method that accounts for all cases that an ADS might face while in operation, including those rare, high-risk events. Such a method also needs the capabilities of dynamically adapting to the operational situation, as well as the available capabilities of the ADS. Furthermore, the method needs to ascertain the vehicle level safety requirements, and such requirements, in turn, need to be exhaustive, i.e. capturing all aspects of the ADS's operations.

In this paper, we present a safety argument fragment aimed at addressing all the above aspects and capable of enabling timely deployment of a performant ADS. The argument fragment provides a complement to already existing design approaches and targets the development of safe operations of the ADS.

The contributions of the paper can be summarised as follows:

- A **safety argument fragment** in Goal Structured Notation (GSN), paving the way towards safety assurance of performant ADSs; and
- **Research gaps** identifying the missing pieces to realise the key parts of the presented safety argument fragment.

The paper is organised as follows. Background of relevant concepts are given in Sec. 2, and Sec. 3 provides a discussion on related work. Sec. 4, presents the proposed safety argument fragment using GSN, including decomposed goals, strategies and contexts. The research gaps related to the decomposed goals are also highlighted and all identified research gaps are collected in Table 1. Moreover, in Sec. 5, the goals are mapped onto an overview of a corresponding ADS. The work is discussed in Sec. 6 and conclusions are given in Sec. 7.

2 Background

In [37], a safety assurance strategy for autonomous vehicles (AVs, or as discussed in this paper: ADSs) is presented. Wardziński [37] argues that through *Situation Awareness (SAW)* the ADS should be able to adapt its tactical decisions in order to operate safely. We note that the essence of Wardziński's approach later has come to be referred to as tactical [21,32] and further extended to Precautionary Safety (PcS) [12,27].

The *SAW model* provides an accurate model of the operational situation of the ADS based on an understanding of the operational context, as well as of the internal state of the ADS. The SAW model is subsequently used for the ADS to derive (safe) tactical and operational behaviour. With this framing, the task of maintaining a safe ADS can then be broken down into a few sub tasks (largely following [37, Fig. 5]):

1. the reliability and accuracy of the SAW model needs to be assured;
2. appropriate safe behaviour needs to be derived, given the SAW model;
3. the ADS needs to produce this behaviour; and
4. the assumptions of the SAW model need to hold during operations.

The SAW needs to include an understanding of the external conditions, including uncertainties, as well as the internal state of the ADS. Consequently, SAW provides an enabler for adaptivity and fault-tolerance of the ADS. To achieve this, there is a need for the perception system and the vehicle platform, to provide estimates of the current states and capabilities of the respective subsystems, not just what the external world looks like. An example of what this might look like, for a high-level overview of an ADS, is shown in Fig. 1.

Precautionary Safety (PcS), takes SAW one step further by providing constraints on the driving policy of the ADS that fulfil quantitative safety requirements [12,27], e.g. in the form of a QRN [38]. In essence, the PcS constraints are found by predicting the

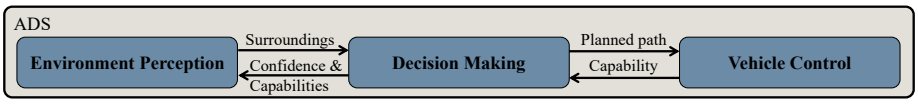

Fig. 1. The environment perception (EP) and vehicle control blocks provide estimates, predictions, capabilities as well as uncertainties to the decision-making block to enable SAW.

probability of all applicable loss events, i.e. accidents or near-accidents, and comparing these to the acceptable frequencies of the QRN. The PcS constraints derived from the loss event probabilities could be seen as an *a priori* statistical knowledge that complements the runtime observations of the SAW. Note that, in this process, it is central to account for not only statistically likely events but also rare events with high severity outcomes, such as, e.g., an animal suddenly running out in front of the vehicle [12].

For the purpose of the argument fragment of this paper, we make use of such PcS principles within the SAW, and the Operational Design Domain (ODD) provides the scope for the design intent of the system and delimits the design-time activities for the ADS [28]. Specifically, the ODD can be used to confine the ADS's operations to where the models and assumptions, used in the design and development, are valid [13].

3 Related Work

The safety case is a central safety activity, entailing to develop: *"[...] a structured argument, supported by a body of evidence that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given environment."* [35, §13.2.1]. For functional safety of automotive systems, ISO 26262 [16] provides many elements for the construction of a relevant safety case and, further, ISO 5083 [19] provides methods targetting ADSs. However, while ISO 5083 [19] provides guidelines towards useful methods, such guidelines remain on a high-level as to avoid being too restrictive. Furthermore, ISO 8800 [18] provides guidance towards developing safe and secure Machine Learning (ML)-based systems and components, and could therefore be valuable to assure the (possibly) ML-based components used to construct the SAW model of the safety argument fragment proposed in this paper.

Safety arguments/cases are well practised, but as discussed in [22] there are many common pitfalls, especially when considering autonomous vehicles. Koopman et al. [22] also indicate that a full safety case for autonomous vehicles would likely need a heterogeneous approach, drawing upon several different methods, something that is also highlighted in [11]. The argument fragment presented in this paper suggests such a heterogeneous approach, drawing upon several different methods and contexts for ADS development and further clarifies and concretises the safety argument suggested in [37]. The presented argument fragment encapsulates not only functional safety aspects, as captured by ISO 26262 [16], but also safety of the intended functionality [17] by imposing vehicle level quantitative safety requirements, or quantitative risk acceptance criteria, as per ISO 5083 [19].

From an industrial perspective, we have seen (partial) safety cases being provided by, e.g. Waymo [8] and Aurora [2]. However, these industry safety cases remain at a

high-level, making them too vague for direct application. Furthermore, they include too many aspects (out of necessity of course) which make them difficult to consume and discuss in an academic context. The argument fragment presented herein outlines key methods and goals, while also clearly pointing out the remaining research gaps and open questions within this scope. Consequently, we believe that the argument fragment we present here provides an easily accessible and clearly defined foundation for further academic and industrial development towards safe deployment of ADSs.

4 The Safety Argument Fragment

Drawing upon the argument in [37], we now present our safety argument fragment for performant ADSs. The argument fragment holds the same core principles as Wardziński [37], and listed in Sec. 2 above. However, we add three central elements:

1. A connection to exhaustive quantitative safety requirements in the form of a QRN;
2. Integration of precautionary tactical safety; and
3. A connection, via the ODD, of models and assumptions from design- to runtime.

The proposed argument fragment is formulated in GSN [1] and presented in Fig. 2. Note that the provided fragment provides a complement to existing design activities and does not include all aspects of developing and designing the system, the platform nor the architecture for realising the ADS.

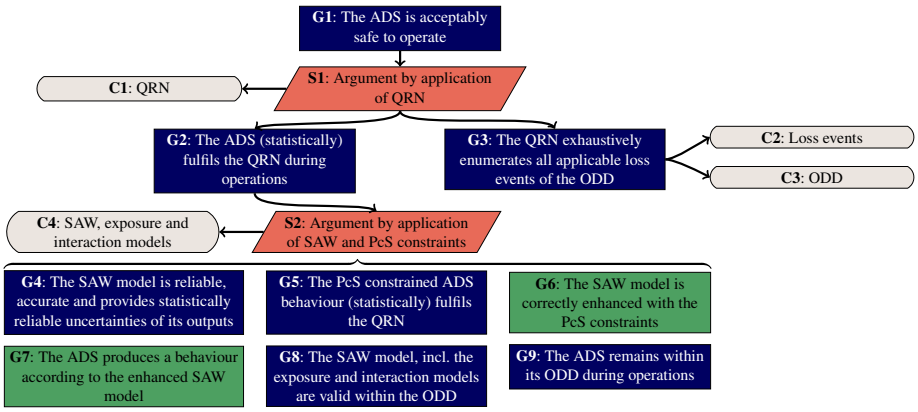

Fig. 2. Illustration of the proposed safety argument fragment for ADSs. Goals (G) are given in navy, strategies (S) in brick, and contexts (C) in sand. Goals in green are not broken down further but nevertheless hold open research challenges. Acronyms: Precautionary Safety (PcS), Situation awareness (SAW), Operational Design Domain (ODD), Quantitative Risk Norm (QRN).

In the following subsections, considerations for each of the goals formulated in Fig. 2 are explored. Note that goal **G2** is not given a separate section as this goal is already broken down through strategy **S2**, as presented in Fig. 2. Furthermore, note that

implementation goals and activities related to V&V are considered here to be part of the lower level goals and are therefore not included explicitly. Likewise, the concrete solutions for how to implement the argument fragment are considered to be future work. Consequently, solution nodes are omitted from the provided argument fragment. For some of the sub-goals, we do however highlight the need for supportive evidence from "evaluation and V&V". Such evidence would include, again following Wardziński [37, Fig. 6], evidence based on simulations, the analysis of simulated and recorded real scenarios, and operational system performance statistics. In addition to breaking down the different sub-goals we also highlight which goals require additional research activities, indicated with green colour in Fig. 3 through Fig. 5. The goals without dedicated figures are highlighted in green in Fig. 2 and the remaining research activities are given in the associated sub-sections.

4.1 G1: Acceptable Safety

The aim of this paper is not to explore all different considerations for what should be included in the safety term related to **G1** of Fig. 2. However, key to the argument presented in this paper, is the presence of *quantitative* vehicle-level safety requirements. The underlying idea is that PcS methods, used for deriving tactical decision constraints, require quantitative levels to work with. This is also in line with ISO 5083 where quantitative risk acceptance criteria are suggested [19]. As discussed in, e.g., [29], different aspects of safety could be considered to include, e.g., limits to the: risk of harm; transfer of harm; and risk of harm in specific situations. Note, however, that not all safety considerations need be captured by quantitative requirements. Qualitative considerations need to be developed alongside, and as a complement, to the presented safety argument fragment of this paper. More in-depth discussions on qualitative elements and the interplay between quantitative and qualitative aspects are, however, not discussed further here.

4.2 G3: Exhaustive Safety Requirements

One of the key contributions, and reasons for, introducing the QRN in [38] was to provide an exhaustive set of safety requirements. This has proven difficult when enumerating hazards, see e.g. [33]. The possible categories of loss events, as opposed to hazards, are independent of the specific system realisation of the ADS making it possible to introduce a generic category capturing all remaining "other" loss events.

Note that the exhaustiveness of the loss events needs to be valid within the ADS's defined ODD. To support this the ODD needs to be well enough specified and understood. The considerations for avoiding ODD exits are addressed in goal **G9**.

4.3 G4: SAW Reliability

Just as Wardziński [37], we acknowledge the need for a reliable and accurate SAW model to support safe decision-making of the ADS. In Fig. 2, we have chosen to highlight the importance of the statistical reliability of the SAW model and the uncertainties

of its associated outputs. Such statistical reliability is crucial to be able to accurately discern consequence probabilities and support the PcS methods. While a central aspect, ensuring the reliability and accuracy of the SAW model is a challenging task for a complex ADS operating in open and uncertain environments. Furthermore, how to design and develop a perception system that is able to produce such reliable estimates, while potentially, fully or partly, relying on ML-based approaches, remains an open research question. Burton et al., [4], present a safety argument towards reducing the impact of functional insufficiencies of ML-based systems. This, however, remains challenging and is reflected in the green colour of goals **G10**, **G11**, **G13** and **G14** of Fig. 3. The green colour of **G12** reflects the need for best practises and evidence of full implementations.

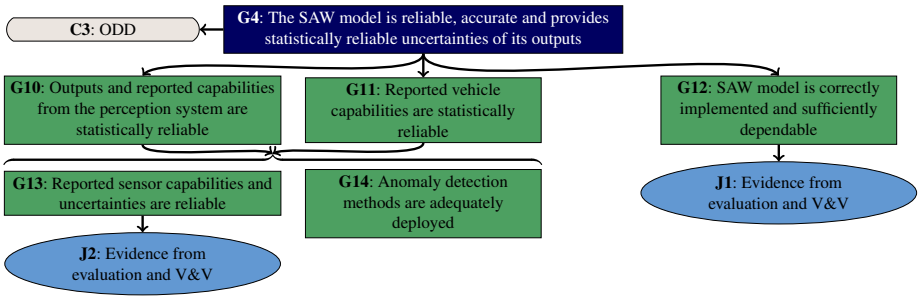

Fig. 3. Presents the goals (G), context (C) and justifications (J) for addressing goal **G4** relating to the reliability and accuracy of the SAW model. Green goals correspond to goals with open research questions.

Promising approaches for constructing a useful and reliable SAW includes dynamic risk assessment [9,25] as well as different threat metrics to judge the probability of collisions or failures, e.g., [7,31]. Note, however, that these approaches focus on the external situation and do not address the ability to assess available capabilities of the ADS and the vehicle platform, as discussed in, e.g., [23]. Nor do they consider the evasive abilities of the ADS as the PcS approaches do [12,27].

To be able to assess consequence probabilities for all applicable adverse events, there is a need to include models for exposure of different adverse events as well as interaction models with other traffic participants. While perhaps not commonly included in the SAW model, these models remain central for the purpose of the argument fragment presented and for the use of the PcS principles.

4.4 G5: PcS Constraints

Given the SAW model, including the exposure and interaction models, the next phase entails finding appropriate constraints on the tactical and operational decisions of the ADS. Following the principles of PcS [12,27] this can be broken down into a two-part process as depicted through goals **G15** and **G16**, see Fig. 4. The consequence probabilities of the applicable loss events are estimated, and these consequence probabilities

are compared to the acceptable frequencies of the QRN. While initial work exists towards supplying both of these parts, see e.g. [12,27], work remains to generalise the PcS approach to cover all applicable loss events, as well as to assess the scalability of the approach when applied to the full complexity of an ADS. Also the tool qualification, related to goal **G17**, requires additional efforts.

The PcS constraints might either be provided in the form of an occupancy grid or as an ability to check the fulfilment of a particular trajectory or set of (planned) decisions. The former approach opens up for optimisation-based trajectory planning whereas the latter supports a doer-checker-type architecture. These aspects are elaborated in Sec. 4.6 relating to goal **G7**.

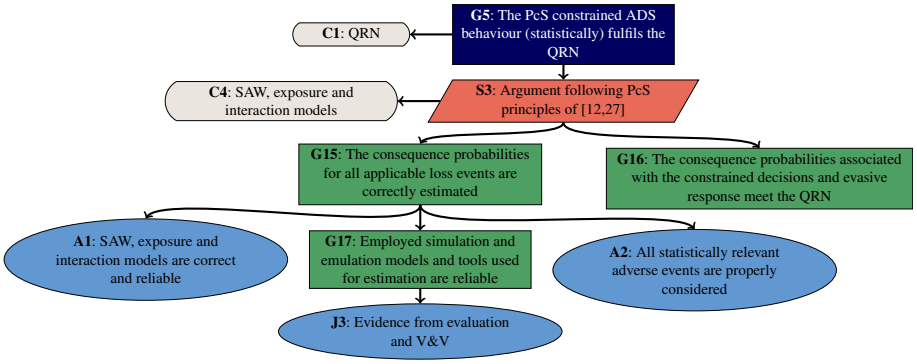

Fig. 4. Depicts the argument for addressing goal **G5**, including the contexts (C), strategy (S), subgoals (G), justifications (J) and assumptions (A). Goals in green hold open research challenges.

Here it is pertinent to also circle back to the different safety considerations safety discussed related to goal **G1** above. While the "statistical fulfilment of the QRN" primarily refers to a fleet-level conformance to the QRN, nothing inhibits this approach from incorporating requirements on the limit to risk of harm in specific situations. Thus, the resulting PcS constraints can be both situation- and vehicle-specific, while also ensuring fleet-level fulfilment of the safety requirements.

4.5 G6: Enhanced SAW

To enable effective trajectory planning while considering the determined PcS constraints we propose to enhance the SAW model with these constraints. The motivation for this is to facilitate appropriate tactical decision-making and the planning thereof. The details for such incorporation remains an open research question.

4.6 G7: Constrained Trajectory Planning

In the trajectory planning phase, there are two main directions when using the SAW model, the selection of which partly impacts the needed outputs from the enhanced

SAW model following goal **G6**. The first direction relates to the use of a supervisory [34] or a doer-checker-type architecture of the ADS [24]. In this context, the enhanced SAW model would be used to accept or reject provided trajectory candidates.

The second direction encompasses an optimisation-based approach, such as, e.g., Model Predictive Control (MPC). Here, the fidelity and required computations of the enhanced SAW need to be much higher to provide concrete constraints to the formulation of the optimisation problem. One could also consider formal methods in the form of logic- or set-based approaches, see e.g. [5, Sec. II.C] or combinations thereof.

To provide the trajectory planning with appropriate information, there is a continued need to investigate methods for how to construct the (enhanced) SAW model and how to convey the information of the PcS constraints in a reliable, yet efficient, way. The first direction, discussed above, would require less computational effort for the construction of the PcS enhanced SAW model. However, an optimisation-based approach would instead be able to come closer to optimal performance. In both cases, there would be a need to employ some kind of surrogate model for estimating the PcS constraints during operations, as for example described in [14]. The details of such a surrogate model, however, remain an open research question.

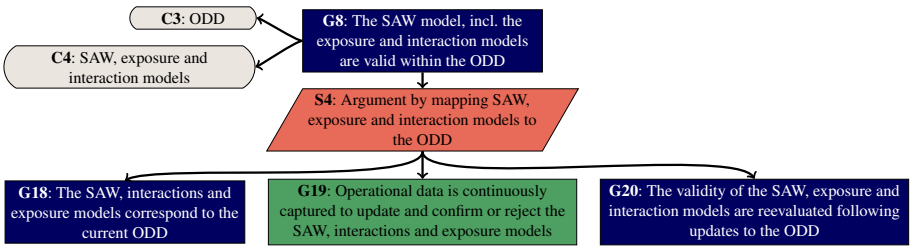

Fig. 5. The argument for goal **G8**, to ensure mapping between models and the ODD.

4.7 G8: Validity of the SAW, Exposure and Interaction Models

Considering the central role played by the SAW model in the presented argument fragment, it is clearly paramount to ensure its reliability, accuracy and validity. The strategy for fulfilling this, i.e. meeting goal **G8**, is provided in Fig. 5. Here, we suggest a mapping between the models supporting SAW and the ODD, similar to the mapping between a use case and the ODD, as suggested in [13]. Any of the assumptions or data to construct the models are therefore selected from within, and validated towards, the intended ODD. To ensure that this mapping remains valid, there is a need to continuously capture operational data. Note that this goal, **G19**, does not refer to real-time checks but rather relates to a continuous feedback loop on fleet level, corresponding to the development cycle of the ADS. The safety implications and details for such continuous data capture from operations remain an open question. This monitoring capability is similar to that needed for monitoring safety performance indicators [36], with the difference being the focus on data that relate to the SAW, exposure and interaction models. One

could, of course, also consider real-time checks during operations, such as described in e.g. [15], but this is not considered necessary for the purpose of the argument presented here.

4.8 G9: Avoiding ODD Exits

For the mapping between the SAW and the ODD to be relevant, the ADS needs to remain within the ODD during operations. For that purpose, we can make use of the ODD exit strategies suggested in [13]. Appropriate triggering conditions need to be in place for detecting ODD exits, and the ADS needs to respond suitably. This can be achieved through handing (back) control to a fallback-ready user (i.e. a human driver if available) or by transitioning into a minimal risk condition [10,28].

Despite best efforts in employing the ODD exit strategies, there might be cases where the predictive power of the trigger conditions are not enough and the ADS finds itself outside its ODD. In such cases, there is a need to quickly identify this and swiftly respond to resolve it. This constitutes a remaining residual risk that will always be present but one that should be decreasing with more and evaluation, V&V and operational evidence of the contrary.

5 The Argument in Relation to an ADS

To concretise the proposed safety argument fragment, the high-level overview of the ADS, presented in Fig. 1, is here expanded and the goals of the argument fragment are mapped onto this expanded view. This mapping is shown in Fig. 6.

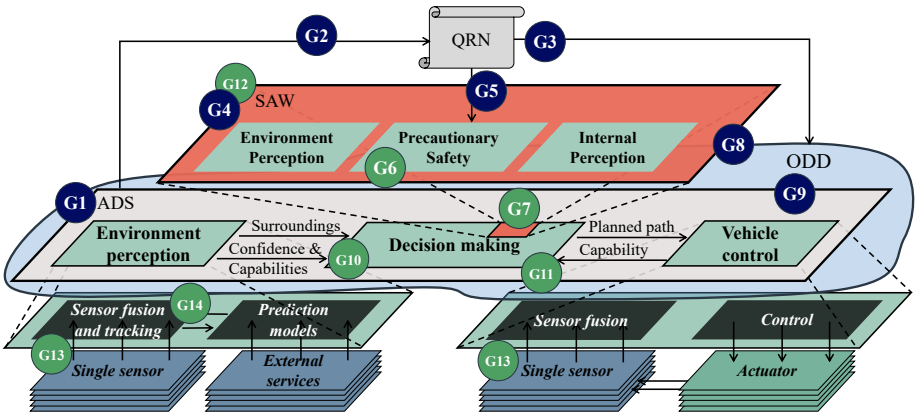

Fig. 6. The goals of the argument fragment and the associated components of the ADS. The colours of the goals correspond to those used throughout the argument fragment.

The EP block perceives the external environment, possibly with support from external services such as, e.g., vehicle-to-infrastructure communication. Similarly, the vehicle control block gauges the ADS's internal capabilities, e.g. to steer and brake. These

form the environment and internal perception that, jointly with PcS principles, make up the (enhanced) SAW. The SAW model itself subsequently make up part of the decision making block of the ADS.

The overarching goals **G1** – **G3** and **G9**, relate to the ADS's safety requirements, the QRN and the ODD. The construction and reliability of the SAW is captured by **G4**; its relationship to the QRN by **G5**; its extension through PcS principles by **G6**; and the relationship between the SAW and the ODD by **G8**. Moreover, **G7** captures the ability for the ADS to produce appropriate behaviour given the SAW. Furthermore, the sub-goals to achieve a reliable estimate, from the sensors all the way up to the SAW, is captured by goals **G10** – **G14**.

Note that the sub-goals related to goal **G5** and Fig. 4, and goal **G8** and Fig. 5 are not depicted in Fig. 6. The positioning of these sub-goals are already captured by the position of the corresponding main goal, i.e. **G5** and **G8** respectively.

6 Discussion

The presented safety argument fragment provides an initial step towards timely safety assurance of performant ADSs. The argument includes several open research gaps to be fully practicable, as indicated throughout Sec. 4. These research gaps are collected in Table 1.

Goal	Research gap
G6	Incorporation of PcS constraints within the SAW model
G7	Surrogate model of the enhanced SAW for runtime usage
G10, **G11**, **G13**	Reliable estimation of perception, vehicle, and sensor capabilities
G12	Reliability of the SAW model and its implementation
G14	Deployment of anomaly detection methods
G15, **G16**, **G17**	Generalisation and scalability of the PcS approach
G19	Safety implications from continuous capture of operational data

Table 1. Articulation of the identified research gaps and the goals from the argument fragment.

6.1 Pitfalls for Safety Assurance

Some of the pitfalls for safety arguments [22], are relevant to discuss further here. Especially, considering how the construction of the SAW relies on data, models, assumptions and simulations. The risk of missing rare events [22, Sec. 2.4.1], as part of the simulation effort, could pose significant problems. However, this could be ameliorated by relying on exposure models (of events, behaviours, environmental elements, etc), for which data can be collected independently of the ADS realisation. Furthermore, by analysing crash statistics, the risk of missing rare events can be minimised. Of course, there will always be a residual risk related to this aspect, as some rare events are yet to happen, but this approach should give a good bound on the probability of occurrence of such unseen rare events.

The pitfall of simulation data validity [22, Sec. 2.4.4], relates to tooling qualification, as e.g. mentioned in relation to goal **G17** in Fig. 4. This remains an open challenge, and one which warrants specific care when the subsequent system or model (i.e. SAW) is used for highly safety-critical decisions. When constructing the SAW, there might also be a risk of violated assumptions, when using formal approaches [22, Sec. 2.5.1]. However, these types of violations could be minimised by the use of an ODD and the matching process suggested related to goal **G8**, see Fig. 5.

6.2 Continuous Assurance

By relying on SAW, the proposed safety argument fragment enables the decoupling of the safety-related activities for different parts of the system from the overall system's safety. In particular, as long as the interfaces between the components providing input to and consuming the outputs from the SAW remain valid, the models that support the SAW can be updated, as related to goal **G8**, without necessarily having to redo the safety activities for other modules. For example, the safety argument for the perception module will remain valid even though an exposure model used for the SAW is updated. Similarly, the confidence in the ADS executing (safe) decisions given the SAW model would also remain, despite updates to the SAW model itself. Consequently, the proposed argument also provides a stepping stone towards dynamic safety cases [6]. For this to be applicable, however, the modularity of the safety argument needs to be maintained throughout the development and V&V activities. Furthermore, the decoupling also provides a means to circumvent some of the need for detailed safety cases for ML-based components, since the impact from these are constrained to the SAW. This does not, however, mean that we can do without appropriate safety cases, such as suggested in, e.g. [4]. Despite ISO 8800 [18], safety cases for complex systems based on ML-based components still remain an open challenge.

6.3 Future Work

In terms of future work, the research gaps identified throughout this paper, and listed in Table 1, deserve further attention. Furthermore, details of the proposed argument fragment should be investigated, including the concrete solutions to each of the sub goals. In a related vein, it would be useful to extend the GSN argument to include confidence arguments using assurance claim points [1, Sec. 1:5]. This would help quantify the uncertainties related to using ML-based methods for perceptions and might further support the analysis of residual risks pertaining to the argument. Additional future work entails, evaluation of the practical applicability of the proposed argument fragment within a relevant industrial context or use case. It would also be interesting to investigate how some of the goals could be framed as contracts. This would entail both considerations from contract-based design [3], as well as how such contracts could be evaluated at runtime [30].

7 Conclusions

This paper develops a safety argument fragment to supports timely deployment of safe and performant Automated Driving Systems (ADSs). It builds upon, and advances, the safety argument initially proposed by Wardziński [37], and brings forward three main contributions, by: (i) incorporating quantitative safety requirements via a QRN; (ii) using SAW and PcS approaches to provide a link between ADS tactical decisions and QRN fulfilment; and (iii) utilising the ODD as an information carrier to ensure validity of the models across design- and run-time. The proposed safety argument fragment illustrates the usefulness of a SAW model as an intermediate abstraction model for achieving safety of ADSs and it also highlights the need for future research into the reliability of the SAW model, especially when relying on ML-based components.

The argument fragment is presented in GSN and some of the initial nine goals are further decomposed through sub-arguments, resulting in a total of 20 goals. Considerations for each of the initial nine goals are discussed and research gaps are identified. The presented safety argument fragment provides a stepping stone towards safety assurance of performant ADSs by modularising the safety considerations for perception, SAW, and trajectory planning while enabling the fulfilment of strict quantitative safety requirements.

Acknowledgments. The authors want to thank the anonymous reviewers for their valuable feedback that has significantly improved the content and quality of the final paper. The research has been partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and partially supported by the Swedish Innovation Agency (Vinnova, through the TADDO2 FFI project and the TECoSA centre for Trustworthy edge computing systems and applications).

References

1. Assurance Case Working Group et al.: Goal structuring notation community standard. Tech. rep., SCSC-141C, Safety Critical Systems Club. (2021)
2. Aurora team: Safety Case Framework development and tailoring (Feb 2025), https://blog.aurora.tech/safety/safety-case-framework-development-and-tailoring
3. Benveniste, A., Caillaud, B., Nickovic, D., Passerone, R., Raclet, J.B., Reinkemeier, P., Sangiovanni-Vincentelli, A., Damm, W., Henzinger, T., Larsen, K.G.: Contracts for system design. Ph.D. thesis, Inria, Rapport de recherche RR-8147 (2012)
4. Burton, S., Gauerhof, L., Heinzemann, C.: Making the case for safety of machine learning in highly automated driving. In: Computer Safety, Reliability, and Security: SAFECOMP Workshops, Trento, Italy, September 12, 2017, Proceedings 36. pp. 5–16. Springer (2017)
5. Dahl, J., de Campos, G.R., Olsson, C., Fredriksson, J.: Collision avoidance: A literature review on threat-assessment techniques. IEEE Transactions on Intelligent Vehicles **4**(1), 101–113 (2018)
6. Denney, E., Pai, G., Habli, I.: Dynamic safety cases for through-life safety assurance. In: Int. Conf. on Software Engineering. vol. 2. IEEE/ACM (2015)
7. Eggert, J.: Risk estimation for driving support and behavior planning in intelligent vehicles. at-Automatisierungstechnik **66**(2), 119–131 (2018)

8. Favaro, F., Fraade-Blanar, L., Schnelle, S., Victor, T., Peña, M., Engstrom, J., Scanlon, J., Kusano, K., Smith, D.: Building a credible case for safety: Waymo's approach for the determination of absence of unreasonable risk. arXiv preprint arXiv:2306.01917 (2023)
9. Feth, P.: Dynamic behavior risk assessment for autonomous systems. Ph.D. thesis, Fraunhofer Verlag (2020)
10. Gyllenhammar, M., Brännström, M., Johansson, R., Sandblom, F., Ursing, S., Warg, F.: Minimal risk condition for safety assurance of automated driving systems. In: Int. Workshop on Critical Automotive Applications: Robustness & Safety (CARS) (2021)
11. Gyllenhammar, M., de Campos, G.R., Törngren, M.: The road to safe automated driving systems: A review of methods providing safety evidence. IEEE Transactions on Intelligent Transportation Systems (2025). https://doi.org/10.1109/TITS.2025.3532684
12. Gyllenhammar, M., de Campos, G.R., Sandblom, F., Törngren, M., Sivencrona, H.: Uncertainty aware data driven precautionary safety for automated driving systems considering perception failures and event exposure. In: Intelligent Vehicles Symposium (IV). IEEE (2022)
13. Gyllenhammar, M., Johansson, R., Warg, F., Chen, D., Heyn, H.M., Sanfridson, M., Söderberg, J., Thorsén, A., Ursing, S.: Towards an operational design domain that supports the safety argumentation of an automated driving system. In: ERTS (2020)
14. Gyllenhammar, M., Zandén, C., Vakilzadeh, M.K.: In-vehicle system for estimation of risk exposure for an autonomous vehicle (Aug 02 2023), EU Pat. EP4219262A1
15. Gyllenhammar, M., Zandén, C., Vakilzadeh, M.K., Falkovén, A.: Methods and systems for automated driving system monitoring and management (Oct 21 2021), EU Pat. EP3895950A1
16. ISO: 26262:2018 Road vehicles – Functional safety (2018)
17. ISO/PAS: 21448:2019 Road vehicles - Safety of the intended functionality (2019)
18. ISO/PAS: 8800:2024 Road Vehicles – Safety and artificial intelligence (2024)
19. ISO/TS: 5083:2025 Road vehicles – Safety for automated driving systems, Design, verification and validation (2025)
20. Johansson, R., Koopman, P.: Continuous learning approach to safety engineering. In: CARS-Critical Automotive applications: Robustness & Safety (2022)
21. Johansson, R., Nilsson, J.: Disarming the trolley problem–why self-driving cars do not need to choose whom to kill. In: CARS (2016)
22. Koopman, P., Kane, A., Black, J.: Credible autonomy safety argumentation. In: 27th Safety-Critical Systems Symposium. pp. 34–50 (2019)
23. Nolte, M., Jatzkowski, I., Ernst, S., Maurer, M.: Supporting safe decision making through holistic system-level representations & monitoring–a summary and taxonomy of self-representation concepts for automated vehicles. arXiv preprint arXiv:2007.13807 (2020)
24. Phil Koopman: Safety Requirements. In: Carnegie Mellon University – 18-642: Embedded Software Engineering (2020), https://users.ece.cmu.edu/~koopman/lectures/ece642/31_SafetyRequirements.pdf
25. Reich, J., Wellstein, M., Sorokos, I., Oboril, F., Scholl, K.U.: Towards a software component to perform situation-aware dynamic risk assessment for autonomous vehicles. In: European Dependable Computing Conf. (EDCC). Springer (2021)
26. Riedmaier, S., Ponn, T., Ludwig, D., Schick, B., Diermeyer, F.: Survey on scenario-based safety assessment of automated vehicles. IEEE access **8**, 87456–87477 (2020)
27. Rodrigues de Campos, G., Kianfar, R., Brännström, M.: Precautionary safety for autonomous driving systems: Adapting driving policies to satisfy quantitative risk norms. In: Intelligent Transportation Systems Conf. (ITSC). IEEE (2021)
28. SAE: SAE J3016:202104 - SURFACE VEHICLE RECOMMENDED PRACTICE - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (2021)

29. Sandblom, F., Rodrigues de Campos, G., Hardå, P., Warg, F., Beckman, F.: Choosing risk acceptance criteria for safe automated driving. In: Int. Workshop on Critical Automotive Applications: Robustness & Safety (CARS) (2024)
30. Schneider, D., Trapp, M.: Conditional safety certification of open adaptive systems. ACM Transactions on Autonomous and Adaptive Systems (TAAS) **8**(2), 1–20 (2013)
31. Schreier, M., Willert, V., Adamy, J.: An integrated approach to maneuver-based trajectory prediction and criticality assessment in arbitrary road environments. IEEE Transactions on Intelligent Transportation Systems **17**(10), 2751–2766 (2016)
32. Schöner, H.P., Antona-Makoshi, J.: Testing for tactical safety of autonomous vehicles. In: 30th Aachen Colloquium Sustainable Mobility (2021)
33. Sulaman, S.M., Beer, A., Felderer, M., Höst, M.: Comparison of the FMEA and STPA safety analysis methods–a case study. Software quality journal **27**(1), 349–387 (2019)
34. Törngren, M., Zhang, X., Mohan, N., Becker, M., Svensson, L., Tao, X., Chen, D.J., Westman, J.: Architecting safety supervisors for high levels of automated driving. In: Intelligent Transportation Systems Conf. (ITSC). IEEE (2018)
35. U.K. Ministry of Defence, Defence Standard 00-56 Part 1: Safety Management Requirements for Defence Systems – Part 1: Requirements (London, UK Issue 7 20170228)
36. Underwriters Laboratories: 4600: Standard for Evaluation of Autonomous Products (2020-04-01)
37. Wardziński, A.: Safety assurance strategies for autonomous vehicles. In: Harrison, M.D., Sujan, M.A. (eds.) Computer Safety, Reliability, and Security. pp. 277–290. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
38. Warg, F., Skoglund, M., Thorsén, A., Johansson, R., Brännström, M., Gyllenhammar, M., Sanfridson, M.: The quantitative risk norm – a proposed tailoring of HARA for ADS. In: Int. Conf. on Dependable Systems and Networks Workshops (DSN-W). IEEE/IFIP (2020)