Presented in – Critical Automotive applications: Robustness & Safety (CARS) Workshop 2024
(Held in conjunction with the 19th European Dependable Computing Conference (EDCC) 2024)

Choosing Risk Acceptance Criteria for Safe Automated Driving

Fredrik Sandblom, Gabriel R. de Campos, Peter Hardå
Zenseact AB
Göteborg, Sweden
{fredrik.sandblom, gabriel.campos, peter.harda}
@zenseact.com

Fredrik Warg
RISE Research Institutes of Sweden
Borås, Sweden
fredrik.warg@ri.se

Fredrik Beckman
Magna Electronics
Linköping, Sweden
fredrik.beckman@magna.com

***Abstract*—It is easy to agree that an automated driving system shall be safe, but it is an on-going discussion what safe means. Several Risk Acceptance Criteria (RAC) candidates have been suggested, but a closer analysis indicates that not all of them are related to risk in a traffic safety sense and that perhaps they are better described as properties that an ADS should be designed to exhibit for other reasons. This paper discusses safety aspects of Automated Driving System (ADS) features and the different incentives and arguments that drive the design of an ADS. More precisely, this paper explores different design goals for safe automated driving and puts forward a combination of Risk Acceptance Criteria (RAC) for limiting the risk of harm. These criteria are motivated and contextualized using a simple real-world traffic example. Furthermore, it is also shown why run-time risk transfer is unavoidable in any system that makes tactical decisions under uncertainty and why this motivates avoiding thought-examples such as the trolley problem as basis for ADS design.**

***Index Terms*—Risk acceptance criteria, safety, automated driving, automated vehicles**

I. Introduction

During the last decades the automotive industry has gone, and continues to go, through structural transformations powered by new breakthroughs on electrification, connectivity, and automation. In particular, recent advances in perception and compute technologies, as well as on situational awareness and threat assessment techniques, have led to high expectations on a rapid development of Automated Driving Systems (ADSs). However, automated driving in highly complex traffic scenarios is a difficult task, in particular considering the quality and performance metrics that society expects.

Safety assurance and safe driving are perhaps the two most challenging tasks in the development of an ADS, as such elements are not only transversal to the whole system, but also define vehicle level behaviours that determine user perception and acceptability. Indeed, tactical decisions made by an ADS should be safe by design and the ADS should have the ability to drive in any situation, including unexpected events and mistakes by human road users, in a suitable manner. In the automotive domain, safety has traditionally been defined as the absence of unreasonable risk of harm in terms of physical injuries or damage to the health of persons [1]. Furthermore, tactical decisions made by an ADS should arguably also be justifiable from an ethical perspective [2], [3]. Recent standards such as ISO 21448 [4] and the upcoming ISO TS 5083 for ADS safety [5] requires that risk acceptance criteria (RAC) are defined as part of the safety assurance process. A top level quantitative RAC (QRAC) represents a utilitarian approach to safety, and criteria based on deontological principles can complement the QRAC to further justify the design.

This paper focus on the safety of ADS features, and the different motivations and arguments that drive the design of an ADS. More specifically, we will approach different elements of safety, and discuss what makes a system safe from the perspective of limiting the risk of harm in relation to perceived safety and following traffic rules. First, we discuss the different design goals for ADS features and the underlying defining elements (Section II). We later articulate some of the risks and challenges associated with them, as well as pertinent Risk Acceptance Criteria (RAC) (Section III). Finally, we provide a simple real-world example to put in context all of the aforementioned elements, and exemplify how system requirements can be formulated based on the aforementioned RACs (Section IV). Two main contributions of this paper are, first, the identification of RAC that affect the magnitude of the risk of harm such that one can apply the highest rigour to those and, second, a formulation and reasoning around what such acceptance criteria can look like.

A key aspect to this paper is the necessity of the safety properties of an ADS to be not only quantifiable but also that the ADS performs risk balancing at run-time with predictable results. Furthermore, this paper also unveils the necessity of an underlying prioritisation of the multiple RACs. More precisely, we will argue that the different RACs should not be treated or prioritised in the same way, as by treating them all as equally important, one may focus on the wrong aspects.

II. ADS design goals

Different manufacturers of ADS features may have different goals for their products. Most would however agree that design goals include providing customer value and that the

This research has been supported by the Strategic Vehicle Research and Innovation (FFI) programme in Sweden, via the project SALIENCE4CAV (ref. 2020-02946).

ADS is safe to use. The following criteria were recently suggested [6], showing the multiple dimensions across which ADS acceptance criteria span:

- Achieving a Positive Risk Balance (PRB)[1] in comparable conditions
- Mitigating risk transfer onto vulnerable populations
- Avoiding negligent computer driver behaviour
- Conforming to industry consensual safety standards
- Meeting regulatory risk requirements
- Addressing ethical and equity concerns

While these are good examples of criteria contributing to ADS features being accepted as safe, they tackle different aspects of acceptance that can be approached with different methods and rigour. Our focus in this work is the formulation of criteria related to risk of harm, which can be expected to be most challenging in terms of acceptable rates, such that these criteria can be treated with sufficient rigour. For instance, injuring a person should arguably be tolerated less often than failing to stop at a stop sign in the absence of other road users. Although users are well-equipped to determine whether a feature is useful to them, it is not possible for a user to determine that a safe feature is safe. The user would be attempting to solve the billion miles problem [7] on their own.

To ensure that users trust the system, it is likely that design goals include exhibiting behaviour that is perceived as being safe; conforming to ethical principles in line with those of the user, following traffic rules, et cetera. In this paper, however, we propose to explicitly sort out the risk acceptance criteria concerned with traffic safety, i.e., risk of harm, such that they can be treated with the highest rigour. Fig. 1 shows an example set of acceptance criteria contributors, where those associated with traffic safety and risk of harm are decoupled from those associated with perceived safety and compliance to traffic rules. Note that the categories may be different to an ADS than to a human driver since the respective risk assessments are not identical due to differences in the underlying capabilities. This entails that behaviour perceived as safe when exhibited by a human may not have the same effect when executed by an ADS, and what constitutes reckless driving can also differ. Furthermore, the analysis provides justification that additional requirements needed for the ADS to be accepted as safe may not need to be implemented to the same rigour. Adequate treatment of such acceptance criteria, such as perceived safety, legal requirements and their inter-play with traffic safety, is left to future research.

III. Risk Acceptance Criteria

A key, underlying idea behind this work is that the ADS design and risk acceptance criteria are the elements to be scrutinised, rather than the run-time decisions of an ADS which occurs as a result of the design. As further discussed in Section IV, run-time uncertainties will pose a very real challenge in determining, in advance, how to treat each possible situation, simply because the situation perceived by the vehicle will differ from the actual situation for which there may exist a precedent. Therefore, we join the crowd of publications that conclude that situation based dilemmas, such as the trolley problem, add little value to a discussion on ADS safety, e.g., [6], [8]. Due to the uncertainties mentioned above, and in contrast to [6], we therefore do not believe that avoiding "intentional risk transfer" is a defensible strategy. In fact, doing so would introduce a "point of no return" where a potential accident is made certain by excluding other potential options to avoid it, without being able to determine, with certainty, that it is unavoidable.

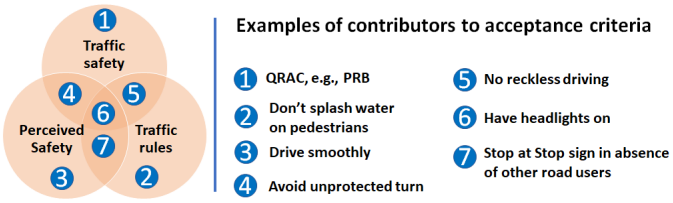

Fig. 1. Acceptance criteria contributors can be grouped with respect to whether they contribute to i) traffic safety in a statistical sense, ii) perceived safety based on the perception of the rider and surrounding traffic users, and iii) legal requirements on how to drive a vehicle.

Hence, it is argued here that a discussion on ADS safety is facilitated by using only RACs that express a requirement on the risk of harm. This would be consistent with the domain definition of safe [1] as well as how safe is popularly defined: as the condition of being protected from harm or other danger [9]. Additional criteria can still be used when justified, but criteria that do not reduce risk of harm can be identified and addressed outside the safety argumentation.

We suggest that an ADS that fulfils the following three groups of acceptance criteria is safe in terms of risk of harm, if the chosen thresholds are justified:

- **Limit to risk of harm** — an argument for a sufficient level of safety on a complete aggregate level, and an assessment of the total impact on society.
- **Limit to transfer of harm (Fairness)** — an argument ensuring that no unfair risk transfer on a complete aggregate level is introduced.
- **Limit to risk of harm in specific situations (Situational excellence)** — the ability to fulfil expectations on behaviour and performance in selected situations.

The implications of the aforementioned risks and the potential associated acceptance criteria are detailed in the following.

A. Limit to risk of harm

There is consensus that risk acceptance criteria to limit the risk of harm is needed, see industry standards such as ISO 26262 [1] and the Safety of the Intended Functionality [4]. Multiple methodologies have been proposed [10], [11] for automated driving and a new ADS specific ISO specification is being developed [5].

[1]PRB is typically understood as the property that deployment of an ADS feature should improve road safety by reducing the aggregate amount of physical harm incurred to road users compared to conventional driving [2].

Real-world safety is often described in quantitative numbers, e.g., accident or injury rates [12], [13] and a reasonable goal for an ADS is to be able to predict how it will perform compared to the actors it will replace, or relative other road users. Therefore we propose to use a QRAC, described in some works as adopting a Quantitative Risk Norm (QRN) [12]. A QRAC has been used to derive precautionary methods for making tactical decisions in ADS features [14], can be compared to reference data, and has the additional benefit that it is straightforward to modify it to comprise criteria in the next section (III-B).

$$\left\{ \begin{pmatrix} c_1 \\ f_1 \end{pmatrix}, \begin{pmatrix} c_2 \\ f_2 \end{pmatrix} \dots \begin{pmatrix} c_m \\ f_m \end{pmatrix} \right\}$$

Example QRAC: m ordered pairs of consequences $\{c_1, \dots, c_m\}$ and frequencies $\{f_1, \dots, f_m\}$.

B. Limit to transfer of harm

The potential transfer of harm from one population group to another is relevant to safety [2]. That is, on an aggregate level, the distribution of harm between different population groups shall be consistent with what is intended or expected; this can be described as being fair.

We suggested in section II to separate elements of perceived safety from those related to risk of harm. Therefore, the risk for each individual population group in a traffic environment before introducing an ADS feature should be considered and refined into risk acceptance criteria, but how the ADS should prioritise in any specified situation should not be considered here. Arguably, if no single population group is exposed to a risk higher than before, no transfer of harm has been introduced. This can be achieved by refining the QRAC in the previous section, as long as the groups are large enough to be statistically meaningful.

$$\left\{ \begin{pmatrix} c_1 \\ f_1 \\ G \end{pmatrix}, \begin{pmatrix} c_2 \\ f_2 \\ G \end{pmatrix} \dots \begin{pmatrix} c_n \\ f_n \\ G \end{pmatrix} \right\}$$

Example: n ordered triples of consequences $\{c_1, \dots, c_n\}$, frequencies $\{f_1, \dots, f_n\}$, and a group identifier, G, e.g., a road user category.

Transfer of harm is avoided at design time by limiting the risk of harm for each individual population group. Therefore, there is no need to detail run-time rules to resolve situations where the accident is assumed inevitable; the point of decision in situations such as the trolley problem occurs rarely enough not to be relevant in the context of risk of harm or transfer of harm.

C. Limit to risk of harm in foreseeable situations

Contrary to groups III-A and III-B, the argument for a quantitative limit to the risk of harm does not explicitly rely on criteria in this group. But if the rate of harm is not significantly affected by these criteria, i.e., these criteria are not needed to avoid harm, how can they be relevant? One answer is that an ADS may be required to perform at least as well as, e.g., humans or state-of-the-art ADAS features, also in situations so rare that they affect the total risk of harm only to a minor extent. Such criteria can be justified as long as the ADS can reach the situation. However, it is not straightforward to predict how the risk of harm is affected by criteria in this group. The risk could even increase in a relative sense while still meeting a QRAC. The explanation to this paradox is that the effect of these requirements on any driving outside the specified situation or comparison can be hard to analyse.

Example 1: Avoid collision when a vulnerable road user (VRU) steps out in the path $^n/_m$ times.

Example 2: Reacting faster than a reference model or perform well in simulations of reconstructed crashes [15].

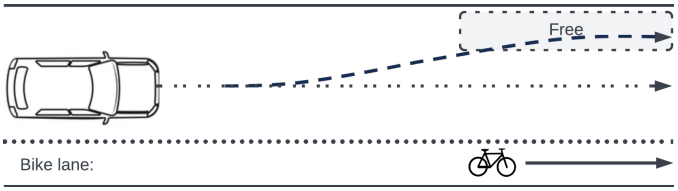

Fig. 2. A car approaches a bicycle with an area that is classified as free. Strategies for passing the bicycle in this situation include driving straight ahead (dotted arrow) or swerving to the left in the lane (dashed arrow). The figure also illustrates that selecting either path over the other constitutes a risk transfer between road users present in the situation, including undetected road users due to the inherent uncertainties in perception, but no statistical risk transfer.

IV. An illustrative example — overtaking a bike

This section provides a simplified illustrative example of a pertinent traffic situation, around which the above mentioned concepts, risks and RACs will be articulated and discussed.

Meeting the suggested risk acceptance criteria does not predict exactly how a given driving task will be executed. However, it is illustrative to see what safety-prediction that fulfilment of the acceptance criteria justifies. To make the example more tangible we select an ordinary traffic situation, shown in Fig. 2, and analyse what fulfilment of the acceptance criteria described in Sections III-A to III-C can mean.

Fig. 2 shows a car that is about to overtake a bicycle on a one-way street and an area on the road, denoted as free. The figure can illustrate either an actual traffic situation, or what an ADS perceives at run time. In either case, it is relevant to keep in mind that an actual situation can differ from the perceived situation and vice versa. This distinction becomes relevant if one formulates acceptance criteria that stipulates situation specific behaviour (see III-C), which we will come back to in the sections below.

A. Effect of limiting risk of harm

Regardless of whether the situation describes the actual state or the perceived state, meeting QRAC predicts that the anticipated overtake together with all future driving will not result in harm more often than the design goal.

Whether this particular overtake will result in a collision or not cannot be deduced from the QRAC.

B. Effect of limiting transfer of harm

Assuming that Fig. 2 describes the actual state, i.e., the only VRU present is the bicycle, limiting the transfer of harm implies that the bicyclist is not less safe due to the overtaking vehicle being controlled by an ADS. Assuming instead that the figure describes the situation as perceived by the ADS, undetected road users can exist and there is a non-zero probability that there is no bicycle present. In any case, if the criteria for limiting the transfer of harm are met, any road user is protected by design.

The distinction between actual or perceived situation makes no difference here — the limit to transfer of harm is clearly independent of how we draw a picture. The example is intended to illustrate that a satisfied safety claim is independent of how the vehicle assesses any particular situation. This is relevant in the next paragraph.

Whether this particular overtake will result in a collision or not cannot be deduced.

C. Effect of limiting harm in foreseeable situations

Assuming that Fig. 2 describes the actual state and the situation is correctly perceived by the ADS, any behaviour resulting from RAC in this family will have the expected effect. Assuming that the figure describes the perceived state but the situation is incorrectly perceived by the ADS will, however, result in the ADS applying a strategy outside the intended situation, which can lead to unexpected results. Consider the unlikely situation where the space marked as "free" is, in-fact, misclassified and contains a road user. If the vehicle were to switch plan from the dotted arrow to the dashed arrow, risk is transferred from the bicyclist to the undetected road user.

Should the ADS therefore not steer to avoid risks? Of course it should. But not because it is certain that it has correctly assessed the situation and "knows" the rule at hand, but because systematically choosing low-risk alternatives is a good strategy for limiting the risk of harm (see Sec. IV-B). Perceived situations will be uncertain and predicting them adds further uncertainty, and the ADS can therefore not be sure that a certain situation is taking place nor that a given action will have the expected outcome. Consequently, criteria originating from very unlikely or very specific situations, e.g., the trolley problem, can lead to unexpected results since the exposure to the situation is so rare that any situation-specific behaviour may be executed due to a misconception, with unpredictable results.

Whether the particular overtake in Fig. 2 will result in a collision or not cannot be deduced, but contrary to before, a prediction valid for this situation is available. The prediction is valid if the situation is correctly assessed by the ADS. The effects of situation-specific decision making when the situation is not perceived correctly is unknown.

V. Conclusions

This paper discusses different aspects and considerations on safety of ADS features, in terms of traffic safety, legal/rule compliance as well as perceived safety from the eyes of users and society in general.

In particular, we argue for the separation of risk of harm from other acceptance criteria, such as perceived safety and traffic rules, such that they can be treated with the highest engineering rigour. We argue that QRAC shall be established for ADS features, that transfer of harm shall, and can, be justifiable in a statistical sense, and that RAC complementary to QRAC can be used to ensure sufficient performance in foreseeable situations.

Finally, we also show that transfer of risk of harm is always present since perception data and predictions are affected by uncertainties. Therefore, situational requirements for rare situations offer little value but may add risk, since they will likely be executed mainly due to a misconception.

References

[1] ISO, "ISO 26262:2018: Road vehicles - functional safety, 2nd ed." 2018.
[2] J.-F. Bonnefon *et al.*, "Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility," Publication Office of the European Union, 2020.
[3] Federal Ministry of Transport and Digital Infrastructure, "Ethics commission on automated and connected driving," Germany, 2017.
[4] ISO, "ISO 21448:2022 road vehicles – safety of the intended functionality," 2022.
[5] ——, "ISO/CD TS 5083 – road vehicles – safety for automated driving systems," 2023.
[6] P. Koopman and W. H. Widen, "Breaking the tyranny of net risk metrics for automated vehicle safety," *SSRN Research Paper Series*, nov 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4634179
[7] N. Kalra and S. M. Paddock, *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* Santa Monica, CA: RAND Corporation, 2016.
[8] R. Johansson and J. Nilsson, "Disarming the trolley problem–why self-driving cars do not need to choose whom to kill," in *Workshop CARS 2016-Critical Automotive applications: Robustness & Safety*, 2016.
[9] "Wikipedia," https://en.wikipedia.org/wiki/Safety, accessed: 2024-01-18.
[10] N. Webb, D. Smith, C. Ludwick, T. Victor, Q. Hommes, F. Favaro, G. Ivanov, and T. Daniel, "Waymo's safety methodologies and safety readiness determinations," 2020.
[11] M. Wood *et al.*, "Safety first for automated driving," https://www.aptiv.com/docs/default-source/white-papers/safety-first-for-automated-driving-aptiv-white-paper.pdf, accessed: 2024-01-18.
[12] F. Warg, M. Skoglund, A. Thorsén, R. Johansson, M. Brännström, M. Gyllenhammar, and M. Sanfridson, "The quantitative risk norm - a proposed tailoring of hara for ads," in *IEEE/IFIP International Conference on Dependable Systems and Networks Workshops*, 2020.
[13] M. Lindman, I. Isaksson-Hellman, and J. Strandroth, "Basic numbers needed to understand the traffic safety effect of automated cars," in *IRCOBI Conference*, 2017.
[14] G. Rodrigues de Campos, R. Kianfar, and M. Brännström, "Precautionary safety for autonomous driving systems: Adapting driving policies to satisfy quantitative risk norms," in *International Intelligent Transportation Systems Conference*, 2021.
[15] "Collision avoidance effectiveness of an automated driving system using a human driver behavior reference model in reconstructed fatal collisions," 2022. [Online]. Available: https://waymo.com/safety/collision-avoidance-benchmarking