Derivation of quantitative risk acceptance criteria for automated driving systems

Jan Erik Stellet, Bernd Müller, Susanne Ebel*

Abstract: Defining risk acceptance criteria for automated driving systems is an essential step for a successful release and avoidance of field incidents. Despite several regulatory provisions and normative frameworks, there is not yet a common understanding and approach, particularly concerning quantitative criteria. This work firstly gives a structured analysis of requirements and approaches. The second contribution is the proposal of a new approach targeting effectively no fleet incidents (ENFLI).

Keywords: Automated Driving, Safety, SOTIF, risk acceptance

1 Motivation

The safety of a product, i.e., not causing harm, is one (although not the only) crucial property for its persistent success on the market and avoidance of legal risks for the manufacturer. However, since perfect safety is typically not achievable, the question what defines acceptable safety risks, i.e., risk acceptance criteria (RAC), arises.

For the safety assurance of automated driving systems (ADS), the definition of defensible (quantitative) RAC is a highly relevant topic. Technical regulations, e.g., UNECE R157 [1] or EU 2022/1426 [2] include high-level provisions yet do not prescribe specific criteria. Frameworks and guidance are provided in industry standards such as ISO 21448 [5] or the upcoming ISO TS 5083. Several publications, e.g., [6–8] address gaps and questions regarding the practical application of these frameworks. However, despite these efforts, a common understanding and accepted approach has not been reached yet.

To advance the discussion, this paper provides context on the applicability of quantitative RAC (Sec. 2). It then summarises regulatory and normative requirements on RAC (Sec. 3) as well as commonly known approaches for the derivation of quantitative RAC (Sec. 4). To address an observed inconsistency between these approaches and the safety level achieved by mature automotive systems, a new approach is proposed, exemplified, and discussed (Sec. 5).

2 Context on quantitative risk acceptance criteria

While safety and risk acceptance are a very broad topic, this paper will focus on quantitative RAC. They typically are to be interpreted within some context and refer to some but not all safety relevant properties of a system. It is crucial to keep the limitations on the

*The authors are with Robert Bosch GmbH, Cross-Domain Computing Solutions, Stuttgart, Germany.

scope and potential purpose of quantitative RAC in mind when evaluating approaches for their applicability in the ADS domain.

In the understanding adopted in this work, quantitative RAC address only hazards at the vehicle level.[1] From these RAC, further quantities can be derived that can also relate to lower levels of an ADS system architecture. According to ISO 21448 [5], verification and validation (V&V) targets "provide evidence that the (risk) acceptance criteria are met". For an extensive analysis of the use of RAC and other quantities we refer to [7].

There are different purposes of quantitative RAC that can be considered but not all purposes can be addressed equally well as is shown in Fig. 1. This underlines that quantitative RAC are only a subset of a more holistic set of acceptance criteria.

A quantitative risk acceptance criterion refers to the number of some (countable) critical event. While there are multiple options, some, like the number of accidents with fatalities, are lagging measures that can be only applied after a product has been introduced to the market. They are not directly useful for product engineering. Leading measures, like the number of safety goal violations (ISO 26262) or ocurences of hazardous behaviour (ISO 21448), are a more appropriate choice.

Furthermore, a quantitative criterion addresses the acceptability of consequences of faults (including functional insufficiencies and their consequences, cf. ISO 21448 [5]), failures or malfunctions. However, for some faults, quantitative RAC are not applicable at all, e.g., classical software bugs, misuse scenarios or security related attacks.

Finally, every quantitative criterion is only meaningful with reference to the measurement principle with which it is evaluated. Generalised conclusions must be handled with care. Usefulness for safety considerations essentially depends on, e.g., whether all circumstances that allow an observation of the failure phenomenon are sufficiently addressed by the measurement principle.

3 Requirements from regulations and standards

There are ADS type approval regulations which contain provisions on (quantitative) RAC, namely UNECE R157 [1], EU 2022/1426 [2] and the German AFGBV[2] [3]. Furthermore, the standard ISO 21448 [5] addresses the safety of the intended functionality (SOTIF) for E/E systems where proper situational awareness is essential to safety derived from complex sensors and processing algorithms.

Tab. 1 highlights similarities and differences in these texts with respect to RAC. Note that this is based on the authors' interpretation of partly differing terminologies.

[1] Note that, while this understanding is in line with ISO 21448 [5, clause 6.1], a closer look at Annex C 2.1 of the standard shows that there can be more than one RAC on the vehicle level. In fact, the example mentions RAC with three different scopes namely (1) an "original acceptance criterion" which is used as an aggregate figure over all accident/incidents, (2) an acceptance criterion for individual accidents/incidents, and (3) following a decomposition of the former to hazardous behaviours, an "acceptance criterion of this behaviour".

[2] *Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung*, engl.: German implementing ordinance for automated and autonomous vehicles

Purpose	Addressable with quantitative RAC
Compliance to type approval regulations & safety standards	▪ *Required* by SAE L3 and L4 type approval regulations. ▪ Quantitative RAC *can be used* according to ISO 21448. ▪ Other safety standards rely on mostly qualitative measures. → Depending on applicable regulation or standard ■■□
Social, market or legal acceptance	Public and legal reaction to statistically similar field incidents varies strongly. → Limited usefulness of of quantitative RAC □□□
Demonstration of safety / safety case	Typical safety case consists of many qualitative and quantitative arguments & evidences. → A contribution but never the whole demonstration of safety ■□□
Derivation of architecture and design targets	Safety is not a directly measurable property, but quantitative targets can help in evaluating architectures & designs. → Very useful application of quantitative (vehicle-level) RAC ■■■
Derivation of V&V targets	V&V targets can be better argued if derived from RAC. → Very useful application of quantitative RAC ■■■

□□□ Not addressable ■■■ Well addressable

Figure 1: Potential purposes and how well they can be addressed by quantitative RAC.

Table 1: Summary of provisions on quantitative RAC.

	Type approval regulations			**Standard**
	UNECE R157	**EU 2022/1426**	**AFGBV**	**ISO 21448**
Types of risks in scope:	Functional safety and SOTIF-related			SOTIF-related
Use of qualitative or quantitative RAC:	Qualitative and quantitative			Qualitative or quantitative or both
Principle for the derivation of quantitative RAC	Comparison to human driver: "Unreasonable risk means the overall level of risk ... compared to a competently and carefully driven manual vehicle" [1, Annex 4, 2.16]	"Unreasonable risk means the overall level of risk ... compared to a manually driven vehicle in comparable transportation services and situations" [2, Article 2, 28.]	"Maß an Sicherheit ... höher als das Maß an Sicherheit bei Fahrzeugen, die von Personen geführt werden" [3, Anlage 1, 10.]	Examples given include selection or combination of the principles outlined in Sec. 4. A rationale is required. [5, clause 6.5]

Table 2: Comparison of commonly referred to principles for the derivation of RAC.

Principle	Summary	Reference of risk	Acceptance criteria
MEM	Acceptable risk is calculated based on the lowest rate of mortality for human individuals in the general population.	The lowest rate of mortality for human individuals	The individual risk (fatalities per person and time) caused by the system is lower than the tolerable risk derived from MEM.
ALARP	Risks are separated in three bands: 1) intolerable, 2) ALARP, 3) broadly acceptable. Individual risks in band 2) are reduced to a level considered "reasonably practicable" by weighing the risk against the effort needed to further reduce it.	The change in collective risk associated with each option/ safety measure.	If the costs of a measure are judged to be disproportionate to the safety benefits, then the measure is judged not to be necessary to further reduce the risk. Several factors need to be considered in this judgement, e.g., state of the art.
GAMAB /GAME	Comparison of two systems: the new system must be globally as safe as or safer than the existing one.	Reference system – could be the human driver if no comparable technical system exists.	The new system is less or equally risky compared to the existing system.
PRB	Allows counterbalancing of residual risks against safety benefit.		

4 Survey of existing approaches

Several approaches for the derivation of RAC have been developed in different application areas. In the railway domain the definition of risk acceptance criteria is described in the CENELEC safety standard EN 50126. There, the principles MEM (Minimum endogenous mortality), ALARP (As low as reasonably practicable) and GAMAB/GAME (Globalment au moins bon / équivalent) are described as methods to define risk acceptance criteria. In addition, the Positive risk balance (PRB) argumentation is mentioned in the code of ethics by the German national ethics committee on automated and connected driving. Tab. 2 provides a basic comparison of these principles.

In the MEM principle, the reference of risk is independent of the technology to be developed. However, it stands to reason to use a reference of risk that is more specific for the automated driving domain (e.g., human traffic fatal accident rates) and therefore, the acceptability of the generic MEM value is questionable. The same applies to the ALARP principle, where quantitative values known from the literature for separation between the three bands of risks are unspecific for ADS.

In order to use GAME/GAMAB for the introduction of new ADS it seems obvious to

refer to human driver accident statistics as easily explainable reference. However, there are several pitfalls in obtaining comparable and defensible reference values, see e.g., [8] and the application example in Sec. 5.3. In this respect, applying the PRB principle faces the same challenges as with GAMAB. In addition, PRB could be misused to offset risks from systems with little or no safety benefit.

In summary, all commonly known approaches feature limitations in the applicability to ADS. Hence, no single approach has yet been considered as gold standard.

5 Proposed approach

In the following, first, in Sec. 5.1, a new risk reference is introduced – the ENFLI (effectively no fleet incidents) approach – which can be used alternatively or in addition. Second, an overall framework is described in Sec. 5.2 which combines multiple approaches for the derivation of quantitative RAC. The framework is illustrated with an example in Sec. 5.3 and the findings are discussed in Sec. 5.4.

5.1 New risk reference: Effectively no fleet incidents (ENFLI)

The previously outlined approaches from the literature have in common that some *rate* of a critical event (e.g., accidents with fatalities per operating time) is used as a risk reference. Exemplary values, see e.g., [6], such as less than one fatal accident per ten million hours of ADS operation, are very small from the perspective of an individual vehicle user or affected non-user.

However, there is a serious implication if an ADS that just achieves such a target rate will be released in a vehicle fleet with a typical size for privately-owned vehicles. Simple calculations with a typical fleet size (e.g., $\geq 100\,000$ vehicles) and an expected average operating time of the ADS per vehicle over its lifetime (e.g., $\geq 1000\,\mathrm{h}$) yield that the *expected value of the number* of critical events over the entire fleet is clearly larger than one. In practice, this can mean that it is expected that accidents will be caused by the fleet of ADS equipped vehicles over their lifetime. Given that serious consequences can result from even a single critical event, this raises doubts whether defensible RAC for an ADS could be derived with such rate-based criteria.

Therefore, the aim in the following is to make the probability of a critical event so small that in a realistic series application such an event is *effectively never expected to occur* over the lifetime of the product. The method is conceptually derived from industry practices for quantitative risk assessments that are used to evaluate potential field issues.

The criterion can be mathematically expressed with a probability or an expected value for the number of critical events over the product lifetime:

$$\mathrm{P}\big(\text{Number of critical events} > 0\big) \ll 1 \text{ or } \mathbb{E}\left[\text{Number of critical events}\right] \ll 1\,. \qquad (1)$$

Numerically there is no significant difference between the two formulae.[3] We assume that

[3] Consider that on the one hand, the expected value for the number of critical events X reads $\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot \mathrm{P}\big(X = k\big)$. On the other hand, the probability of $X > 0$ events can be written as $\mathrm{P}\big(X > 0\big) = \sum_{k=1}^{\infty} \mathrm{P}\big(X = k\big)$. Thus, the only difference between the probability and expected value approach is that the probabilities $\mathrm{P}\big(X = k\big)$ for $k > 1$ in the sum are weighted by k instead of 1. However, those probabilities should be very small given that the entire sum shall be $\ll 1$.

the expected value argumentation is easier to communicate, hence we use this in the sequel.

To calculate an expected value of a number of critical events, the event type needs to be defined. Preference is given to the number of **c**riticality **n**ormalised **s**afety **g**oal **v**iolations (CNSGV) which is a leading measure, cf. Sec. 2. The term *criticality normalised* refers to a distinction according to the criticality of the safety goal as expressed by the ASIL rating in ISO 26262 [4]. In the proposal presented here, a difference in the ASIL rating is allowed to have an influence on the numerical values.

Much smaller than one is not a directly usable formula, so a more precise definition than $\ll 1$ must be introduced. Consistent with risk assessment experiences a range with a safety margin of about two orders of magnitude is proposed. Then, the main condition can be formulated as

$$\mathbb{E}\left[\text{CNSGV}\right] < E_{\text{limit}} \text{ where } E_{\text{limit}} \in [0.01, 0.05] \ . \tag{2}$$

While in each application an exact value will be assigned to E_{limit}, we consider it important to keep the range in mind. This also reflects the inherent uncertainty of the problem.

Typically, the quantification will be applied on a per safety goal basis. If multiple safety goals are relevant this might be critiqued as not conservative enough, but this is considered consistent with ISO 26262 [4]. As long as the number of safety goals is not too large, the margin implied in $E_{\text{limit}} \in [0.01, 0.05]$ is judged as being sufficient.

To compute $\mathbb{E}\left[\text{CNSGV}\right]$, an adequate fault model is required. Two standard cases are:

1. **Rate:** The RAC can be reasonably modelled as a per hour (or per another unit of time) failure rate r_{RAC}. This is useful when the failure phenomenon can continuously cause hazardous behaviour. Then $\mathbb{E}\left[\text{CNSGV}\right] = r_{\text{RAC}} \cdot RIF$ with some risk influencing factors combined in the variable RIF. This yields

$$r_{\text{RAC}} < E_{\text{limit}}/RIF \ . \tag{3}$$

2. **Probability on demand:** The RAC can be reasonably modelled as a per situation failure probability p_{RAC}. This is particularly useful when the failure phenomenon can only cause hazardous behaviour in a particular, "discrete" situation, e.g., a parking situation. Similarly, one obtains

$$p_{\text{RAC}} < E_{\text{limit}}/RIF \ . \tag{4}$$

Deriving risk influencing factors (combined in RIF) for the corresponding system and safety goal is an important part in the application of the ENFLI approach. Typical examples for risk influencing factors are listed below:

- **Number of vehicles:** Almost every risk scales with the number of vehicles in the fleet. Therefore, this parameter should contribute to the risk modelling which is a major feature of ENFLI compared to other approaches.
- **(Average) operating time of the function under consideration over the vehicle lifetime** either as a time value (failure modelling as a rate) or as a unitless expected number of situations (failure modelling as a probability on demand).

- **Normalisation factor derived from the ASIL classification parameters of risk:** E.g., equal to one order of magnitude per ASIL from 1.0 for ASIL D to 0.001 for ASIL A. Note that one should not inadvertently take the operating time of the function into account twice, explicitly and as part of an exposure parameter in the ASIL classification.

There might be qualitative arguments to relax the RAC number. For this, an expert judgement with a rationale is needed and the overall reduction should be restricted to two orders of magnitude. If such a rationale is available and with a choice of $E_{\text{limit}} = 0.01$ one obtains $\mathbb{E}\,[\text{CNSGV}] < 1$ from (2) which still implies that essentially no incident is expected, although no buffer is left. The following arguments may e.g., be considered to argue that no such buffer is needed:

- **Introduction of new technology:** The assumption here is that with the introduction of a new technology it is to some extent more easily accepted by society that "safety perfection" is less realistic than for mature technologies. This may cause some level of leniency towards the technology. A prerequisite is that if some incidents occur the causes are investigated, analysed, and removed.

- **Positive risk balance:** The assumption is that society is more willing to accept a risk if the introduction of the technology significantly reduces another risk. This is not the same as making the PRB a quantitative criterion of its own.

5.2 A generally applicable framework

The introduced ENFLI approach is particularly beneficial for innovative high-risk systems where little field experience of comparable systems is available and in case of an initial market introduction with a limited fleet size. However, there are cases in which other approaches can be used to derive RAC.

Thus, it is proposed to combine several approaches depending on certain criteria. The criteria and resulting approaches are visualized in a decision graph in Fig. 2. Three expert decisions are required as shown in the upper part:

1. Is one or more **quantitative RAC needed** at all or are qualitative criteria sufficient? The answer and rationale are part of the overall safety case.

2. Is an **unlimited market introduction possible** from a safety point of view, considering the available evidence?

3. Is there **sufficient data from comparable reference systems** available to derive risk acceptance criteria according to the GAMAB principle? Is a **demonstration of lower risk compared to human drivers** required by relevant regulations?

Depending on which path in the decision matrix is followed, some or all the activities shown in the lower part of Fig. 2 become relevant. Note that the risk reference derived from human accident statistics in the GAMAB (comparison to human driver) approach needs to be decomposed to reference values for the relevant hazardous behaviours.[4]

[4]See ISO 21448 [5, Annex C2.1] where an acceptance criterion is decomposed by dividing by the conditional probability of exposure to a scenario where the hazardous behaviour can lead to harm, $\mathrm{P}(E|\,HB)$, the probability that the hazardous behaviour is not controllable, $\mathrm{P}(C|\,E)$, and that the severity of the harm created matches the severity addressed with the acceptance criterion, $\mathrm{P}(S|\,C)$.

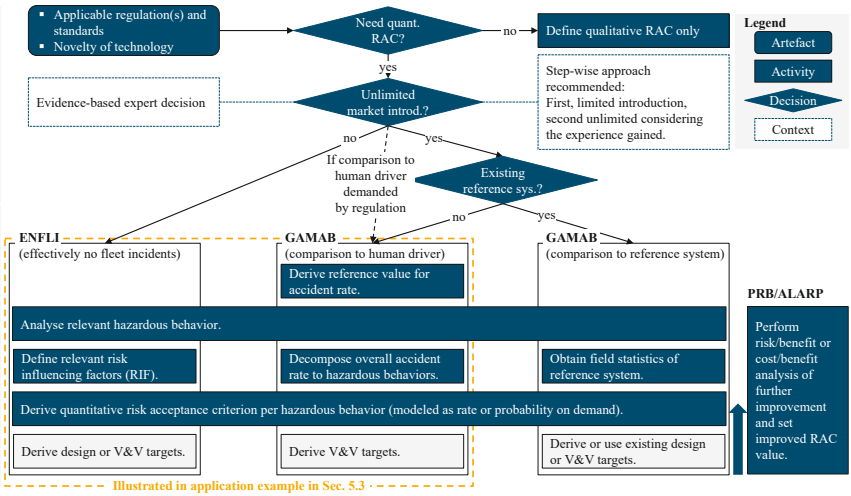

Figure 2: General framework to consider for the derivation of quantitative risk acceptance criteria for AD systems.

5.3 Application example

To illustrate the previously introduced risk references and associated activities, the ENFLI and the GAMAB (comparison to human driver) approach will be used to derive acceptance criteria for one hazardous behaviour of a hypothetical ADS.

The hypothetical system under consideration is an automated lane-keeping system (ALKS) for passenger cars which controls the longitudinal and lateral movement including lane change manoeuvres without further driver intervention, cf. [1]. The operational design domain (ODD) is limited to motorways with structurally separated driving directions and a travelling speed of up to 130 km/h.

Out of several safety goals relevant for the ALKS the example considers “collisions due to erroneous lateral guidance shall be avoided”. Parameters and assumed values relevant for the derivation of the RAC are given in Tab. 3.

For the ENFLI approach, the rate-based criterion (3) with $E_{\text{limit}} = 0.05$ yields

$$r_{\text{RAC,ENFLI}} < \frac{E_{\text{limit}}}{RIF} = \frac{0.05}{10\,000 \cdot 1000\,\text{h} \cdot 0.01} = 5 \times 10^{-7}\,/\text{h}\,. \tag{5}$$

The GAMAB (comparison to human driver) criterion is derived in three consecutive steps: First, a global reference value for the rate of fatal accidents by human drivers is required. This requires statistics on the number of accidents and the distance travelled for a comparable ODD, e.g., limited to passenger cars on motorways. We refer to the data sources and methodology applied in [8] for this example. Furthermore, a value for the distance between two accidents can be converted to a rate over time assuming an average travelling speed. This way, one obtains an average rate of $\approx 1.5 \times 10^{-7}\,/\text{h}$.

Although purely statistical fluctuation of this average can be additionally accounted for, the average is nonetheless dominated by non-rule conform drivers and includes older vehicles with lower technical standard. Unfortunately, there is no established approach or data basis on which a defensible margin for correcting these effects could be derived. Thus, to be on the safe side, a pessimistic safety margin of approximately two orders of magnitude is assumed which leads to a global reference value of $r_{\text{ref, global}} = 1 \times 10^{-9}\,/\text{h}$ for fatal accidents. This margin is to be regarded only as a proposal and other values might be defensible as well.

Second, the global reference value for fatal accidents from the first step is decomposed into individual rates for accidents related to specific hazardous behaviours. In this case, the share of accidents $\eta_{\text{specific,HB}}$ due to "erroneous lateral guidance" is estimated. This is done by estimating the share of relevant accidents where crossing a lane marking was the first event from the German in-depth accident study (GIDAS) [9].

Furthermore, the fault model which links the hazardous behaviour "erroneous lateral guidance" to the occurrence of a fatal accident is established. This is reflected in three conditional probabilities $\mathrm{P}\big(E|\,HB\big)$, $\mathrm{P}\big(C|\,E\big)$ and $\mathrm{P}\big(S|\,C\big)$, see Tab. 3 for explanations and assumed values for the example.

Third, the decomposition of the global reference value $r_{\text{ref, global}}$ yields the following acceptance criterion for the hazardous behaviour "erroneous lateral guidance":

$$r_{\text{RAC,GAMAB}} < \frac{r_{\text{ref, global}} \cdot \eta_{\text{specific,HB}}}{\mathrm{P}\big(E|\,HB\big) \cdot \mathrm{P}\big(C|\,E\big) \cdot \mathrm{P}\big(S|\,C\big)} = \frac{1 \times 10^{-9}\,/\text{h} \cdot 0.5}{0.01} = 5 \times 10^{-7}\,/\text{h}\,. \quad (6)$$

Therefore, for a limited market introduction with only 10 000 vehicles and 1000 h of operating time, the RAC for the considered hazardous behaviour derived with the ENFLI approach is the same as when applying the GAMAB principle.

However, while the GAMAB approach is independent of the fleet size, the results from the ENFLI approach scale linearly with the fleet size. In practice, the latter cannot grow unboundedly. This motivates to estimate the order of magnitude of a limit value $\overline{r}_{\text{RAC,ENFLI}}$. To this end, we consider a hypothetical ADS with the following properties:

- The safety goal can be violated continuously over time, thus the RAC can be reasonably modelled as a per hour rate.
- A deployment to a large-scale platform is intended and thus a fleet size of $\leq 1 \times 10^7$ vehicles is assumed. Fleets beyond that size are very rare and the size increase is not by an order of magnitude. These rare cases are hence covered by the margin implied in $E_{\text{limit}} \in [0.01, 0.05]$.
- Concerning the contributions to the risk influencing factor, an average vehicle lifetime of 10 000 h, 100 % operating time of the function and a normalization factor of 1.0 based on the maximum ASIL classification are assumed.

Then, inserting into (5) yields a limit value of

$$\overline{r}_{\text{RAC,ENFLI}} < \frac{E_{\text{limit}}}{RIF} = \frac{0.01}{1 \times 10^7 \cdot 10\,000\,\text{h} \cdot 1.0} = 1 \times 10^{-13}\,/\text{h}\,. \quad (7)$$

This value might look very conservative but comparing it with the level achieved by mature safety related automotive systems, at least with respect to fatalities, this is not far from reality.

Table 3: Relevant parameters and assumed values for application example in Sec. 5.3.

Parameter	**Value**	**Rationale**
Contributions to risk influencing factor (RIF) in ENFLI approach		
Number of vehicles	10 000	Initial market introduction of a limited amount
(Average) operating time of the ALKS over vehicle lifetime	1000 h	Average vehicle lifetime of 10 000 h and thereof 10 % ALKS operation given ODD limitation
Normalization factor derived from the ASIL classification	0.01	Based on classification of "erroneous lateral guidance" as ASIL B due to low severity in motorway situations as determined from accident statistics.
Risk reference derived from accident statistics		
Global reference value for the rate of fatal accidents on motorways involving passenger cars	$r_{\text{ref, global}} = 1 \times 10^{-9}\,/\text{h}$	• Number of fatal accidents with passenger cars on German motorways in 2015-2019 from DESTATIS [10] • Annual distances travelled by passenger cars on motorways as in [8] • Conversion to rate per hour assuming 110 km/h average travelling speed • Additional safety margin of two orders of magnitude
Share of accidents due to specific hazardous behaviour	$\eta_{\text{specific,HB}} = 0.5$	Based on analysis of accidents described in GIDAS [9] triggered by passenger cars with electronic stability control on motorways.
Conditional probability of exposure to a scenario where the hazardous behaviour can lead to harm	$\mathrm{P}(E \mid HB) = 1.0$	Practically in all scenarios for "erroneous lateral guidance"
Probability that the hazardous behaviour is not controllable	$\mathrm{P}(C \mid E) = 1.0$	Pessimistic choice
Probability that the injury severity of harm due to an uncontrolled hazardous behaviour is fatal	$\mathrm{P}(S3 \mid C) = 0.01$	Based on analysis of accidents described in GIDAS [9] where crossing a lane marking was the first event.

5.4 Discussion

While this outcome is not surprising given the introductory remarks in Sec. 5.1, it raises again the question for which purposes quantitative RAC can be useful. Concerning the purpose of designing a system (architecture), it makes sense to consider targets derived according to the ENFLI approach to develop a sufficiently strong safety architecture that allows for a scaled-up market introduction. However, concerning the purpose of deriving V&V targets, it will become extremely challenging to demonstrate that an RAC nearing $\overline{r}_{\mathrm{RAC,ENFLI}}$ is achieved before market introduction.

Thus, it is important to differentiate between a quantified risk value as an acceptance criterion and quantified confidence obtained from V&V results that the former is achieved. Even if there is a gap between the statistical confidence obtained from V&V results before market introduction and the RAC, qualitative measures can prevent this gap from materialising as excess risk.

One qualitative argument with a quantifiable risk-limiting effect is scrutiny in the observation of field incidents. If the field observation and control of operation are effective, the excess risk is limited to singular critical events. Note that the choice of (normalised) safety goal violations as the type of critical event used in the RAC definition has an additional risk-limiting influence.

Clearly, the gap between RAC values derived using the ENFLI approach and quantified confidence obtained from achievable V&V targets can be made small by initially introducing only a limited quantity of the system in the market. The experience gained with this fleet can inform a subsequent larger introduction.

6 Summary

Quantitative RAC are a subset of a more holistic set of acceptance criteria and while they are clearly useful for some purposes, e.g., evaluating a system architecture or deriving V&V targets, other purposes are better addressable by qualitative criteria and measures. Moreover, quantitative criteria are well suited to address some risks, e.g. stemming from functional insufficiencies, but are not applicable to others, e.g., security attacks.

Known approaches for defining quantitative RAC are either too generic to yield defensible criteria for an ADS (e.g., MEM), or their application raises additional questions for which however sufficiently detailed data and models is lacking (e.g., GAMAB , PRB).

We identify that there is a conceptual issue that such approaches consider the rate of critical events normalised per vehicle whereas, in practice, a manufacturer releases an entire fleet of vehicles. Motivated by the fact that serious consequences can result from even a single critical event in the field, this paper proposes a new risk reference targeting effectively no fleet incidents (ENFLI).

We present an illustrative example, which shows that the RAC values obtained with the ENFLI approach quickly outgrow those derived from a comparison against the accident rate of human drivers (GAMAB approach). On the one hand, this is a desirable property of an RAC as it helps designing a sufficiently strong system safety architecture that allows for a scaled-up market introduction. On the other hand, a gap can occur before release between the RAC value and quantifiable confidence obtained from V&V results. However, as we discuss, the size of such a gap does not imply a proportionally

5.4 Discussion

While this outcome is not surprising given the introductory remarks in Sec. 5.1, it raises again the question for which purposes quantitative RAC can be useful. Concerning the purpose of designing a system (architecture), it makes sense to consider targets derived according to the ENFLI approach to develop a sufficiently strong safety architecture that allows for a scaled-up market introduction. However, concerning the purpose of deriving V&V targets, it will become extremely challenging to demonstrate that an RAC nearing $\overline{r}_{\mathrm{RAC,ENFLI}}$ is achieved before market introduction.

Thus, it is important to differentiate between a quantified risk value as an acceptance criterion and quantified confidence obtained from V&V results that the former is achieved. Even if there is a gap between the statistical confidence obtained from V&V results before market introduction and the RAC, qualitative measures can prevent this gap from materialising as excess risk.

One qualitative argument with a quantifiable risk-limiting effect is scrutiny in the observation of field incidents. If the field observation and control of operation are effective, the excess risk is limited to singular critical events. Note that the choice of (normalised) safety goal violations as the type of critical event used in the RAC definition has an additional risk-limiting influence.

Clearly, the gap between RAC values derived using the ENFLI approach and quantified confidence obtained from achievable V&V targets can be made small by initially introducing only a limited quantity of the system in the market. The experience gained with this fleet can inform a subsequent larger introduction.

6 Summary

Quantitative RAC are a subset of a more holistic set of acceptance criteria and while they are clearly useful for some purposes, e.g., evaluating a system architecture or deriving V&V targets, other purposes are better addressable by qualitative criteria and measures. Moreover, quantitative criteria are well suited to address some risks, e.g. stemming from functional insufficiencies, but are not applicable to others, e.g., security attacks.

Known approaches for defining quantitative RAC are either too generic to yield defensible criteria for an ADS (e.g., MEM), or their application raises additional questions for which however sufficiently detailed data and models is lacking (e.g., GAMAB , PRB).

We identify that there is a conceptual issue that such approaches consider the rate of critical events normalised per vehicle whereas, in practice, a manufacturer releases an entire fleet of vehicles. Motivated by the fact that serious consequences can result from even a single critical event in the field, this paper proposes a new risk reference targeting effectively no fleet incidents (ENFLI).

We present an illustrative example, which shows that the RAC values obtained with the ENFLI approach quickly outgrow those derived from a comparison against the accident rate of human drivers (GAMAB approach). On the one hand, this is a desirable property of an RAC as it helps designing a sufficiently strong system safety architecture that allows for a scaled-up market introduction. On the other hand, a gap can occur before release between the RAC value and quantifiable confidence obtained from V&V results. However, as we discuss, the size of such a gap does not imply a proportionally

growing excess risk in the field when field observation is effective.

References

[1] United Nations Economic Commission for Europe, "01series of amendments to UN Regulation No. 157: Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping," 2022.

[2] European Commission, "Commission Implementing Regulation (EU) 2022/1426: Uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles," in *Official Journal of the European Union (Vol. 65, L 221)*, pp. 1–64, 2022.

[3] Bundesministerium für Digitales und Verkehr, "Verordnung zur Regelung des Betriebs von Kraftfahrzeugen mit automatisierter und autonomer Fahrfunktion und zur Änderung straßenverkehrsrechtlicher Vorschriften," in *Bundesgesetzblatt (Jg. 2022, Teil I, Nr. 22)*, pp. 986–1010, 2022.

[4] International Organization for Standardization, "ISO 26262:2018 Road vehicles – functional safety" Geneva, Switzerland, 2018-12.

[5] International Organization for Standardization, "ISO 21448:2022 Road vehicles – safety of the intended functionality," Geneva, Switzerland, 2022-06.

[6] C. Amersbach, T. Ruppert, N. Hebgen and H. Winner, "Macroscopic Safety Requirements for Highly Automated Driving in Urban Environments," in *13. Graz Symposium Virtual Vehicle (GSVF)*, Graz, 2020.

[7] L. Putze, L. Westhofen, T. Koopmann, E. Böde, and C. Neurohr, "On Quantification for SOTIF Validation of Automated Driving Systems," *2023 IEEE Intelligent Vehicles Symposium (IV)*, Anchorage, 2023.

[8] F. Fahrenkrog, L. Drees, and F. Raisch, "Implications of the positive risk balance on the development of automated driving," in *Traffic Injury Prevention (24sup1)*, pp. 124–130, 2023.

[9] D. Otte, J. Michael, and H. Carl, "Scientific approach and methodology of a new in-depth investigation study in Germany called GIDAS," in *Proceedings of the International Technical Conference on the Enhanced Safety of Vehicles*, 2003. Parameters used in the example were derived from the data available in GIDAS until 12/2019.

[10] DESTATIS (Federal Statistical Office of Germany), "Verkehrsunfälle in Deutschland," `https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/_inhalt.html#sprg478574`.