*Authors' version (postprint)*

# The Quantitative Risk Norm - A Proposed Tailoring of HARA for ADS

Fredrik Warg, Rolf Johansson, Martin Skoglund, Anders Thorsén, Mattias Brännström, Magnus Gyllenhammar, and Martin Sanfridson

**Published in:**

Proceedings of 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)

**Presented at:**
6th International Workshop on Safety and Security of Intelligent Vehicles (SSIV 2020), 2020-06-29

**DOI:** 10.1109/DSN-W50199.2020.00026



*2020-08-17*

# The Quantitative Risk Norm - A Proposed Tailoring of HARA for ADS

Fredrik Warg, Martin Skoglund, Anders Thorsén
*RISE Research Institutes of Sweden*
Borås, Sweden
{fredrik.warg, martin.skoglund, anders.thorsen}@ri.se

Rolf Johansson
*Autonomous Intelligent Driving*
Göteborg, Sweden
rolf.johansson@aid-driving.eu

Mattias Brännström, Magnus Gyllenhammar
*Zenuity AB*
Göteborg, Sweden
{mattias.brannstrom, magnus.gyllenhammar}@zenuity.com

Martin Sanfridson
*Volvo Technology AB*
Göteborg, Sweden
martin.sanfridson@volvo.com

***Abstract*—One of the major challenges of automated driving systems (ADS) is showing that they drive safely. Key to ensuring safety is eliciting a complete set of top-level safety requirements (safety goals). This is typically done with an activity called hazard analysis and risk assessment (HARA). In this paper we argue that the HARA of ISO 26262:2018 is not directly suitable for an ADS, both because the number of relevant operational situations may be vast, and because the ability of the ADS to make decisions in order to reduce risks will affect the analysis of exposure and hazards. Instead we propose a tailoring using a quantitative risk norm (QRN) with consequence classes, where each class has a limit for the frequency within which the consequences may occur. Incident types are then defined and assigned to the consequence classes; the requirements prescribing the limits of these incident types are used as safety goals to fulfil in the implementation. The main benefits of the QRN approach are the ability to show completeness of safety goals, and make sure that the safety strategy is not limited by safety goals which are not formulated in a way suitable for an ADS.**

***Keywords*-ADS, automated driving, hazard analysis, HARA, functional safety, ISO 26262, risk norm.**

## I. Introduction

The development of automated driving systems (ADS) [1] has seen major investments in recent years. An ADS can perform all of the dynamic driving task of a vehicle for an extended period of time. The hopes are that such systems will provide more efficient, accessible, and safer transport solutions. But showing that they are in fact safe has been identified as one of the major challenges [2], [3]. The much-used ISO 26262:2018 standard [4] covers functional safety for electrical/electronic (E/E) functionality in road vehicles. The standard prescribes a dedicated qualitative *hazard analysis and risk assessment* (HARA) method where the output is a set of top-level safety requirements, or *safety goals* (SGs), that must be shown to be complete and consistent. This is a crucial part of the safety argument and body of evidence, or *safety case*, that is needed to show that a function is safe. In this paper we argue that this method has crucial shortcomings when it comes to producing useful SGs for an ADS, and therefore propose an alternative method for such systems.

As discussed further in Sec. II, the defining difference is the absence of a human driver and as a consequence transfer of tactical decisions to the ADS. This means that the system has to show a safe behaviour as a result from the combined tactical and operational decisions; it can use different strategies for handling variations in the traffic and environment, as well as performance variations, and plan the driving in order to reduce exposure to situations where the risk of accidents is deemed too high. This can be done by for example increasing longitudinal and lateral distance margins and set a speed that is adjusted to safely taking care of predicted possible incidents. The work of guaranteeing safety is thereby enabled by adjusting the proactive decision making, rather than only addressing errors in automated subsystems due to faults independent on the traffic situation. This also means that enumeration of operational situations, or relevant scenarios that can occur when using the feature, in the HARA is both intractable and unnecessary. Intractable since the number of potentially relevant operational situations may be vast, making an argument for completeness a very difficult task; and unnecessary since much of this complexity can be confined in the solution domain by means of using tactical decisions and an appropriately defined operational design domain (ODD) [5] to reduce risks. Another consequence of these differences is that the practice of formulating hazards based on simple HAZOP [6] style failure modes is less suitable for an ADS.

For these reasons, we firstly propose to replace the fixed risk assessment criteria of ISO 26262 with the establishment of a *quantitative risk norm* (QRN). This norm is essentially a budget of acceptable frequencies of incidents (including accidents) assigned to a number of consequence classes with different severity, where the frequency budget for each consequence class has a strict limit. Then secondly, as an output of the

This research has been supported by Vinnova via the project ESPLANADE (ref 2016-04268).

activity, the safety goals will be to avoid all listed incidents to below their allotted frequencies. This renders the task of listing hazards and operational situations unnecessary for the purpose of establishing SGs. In the paper, we refer to this as the *QRN approach*. As explained in Sec. III we include both safety, or the absence of accidents causing injuries, and quality, or the absence of other undesirable traffic incidents, in this norm. Part of the work in establishing a risk norm is mapping incidents into a set of incident types, where the incidents should be partitioned in a way that will support development and verification, while making sure the risk budget for all consequence classes are met.

Benefits of the QRN approach are that: completeness of SGs can be ensured by defining the incident types according to the MECE principle (mutually exclusive and collectively exhaustive), as suggested in Sec. III-B, so that any possible conceivable incident falls into one of the classes; and that the SGs become independent of the strategies used to achieve them, i.e. they are not affected by the ability of the ADS to avoid risks by tactical decisions. As elaborated on in Sec. IV, it does not mean that analysis of scenarios or failure modes in the system will not be needed; this will most likely be efficient in order to establish some parts of a functional safety concept (FSC) to fulfil the SGs. So while much of the complexity remains for the FSC, there will also be a significant degree of freedom to define a safety strategy within the confines of the risk norm and the ODD.

While current functional safety standards favor qualitative frameworks with discrete safety integrity levels, we discuss in Sec. V how using our proposed risk norm hints at the usefulness of a quantitative framework for safety assurance, and how this could be used to avoid some problems with the ASIL decomposition and inheritance rules. Finally, after relating our method to other work in Sec. VI, we draw some conclusions in Sec. VII.

In summary, the main contributions in this paper are:

- A novel method to derive safety goals using a quantitative risk norm with consequence classes and incident types for relevant failures, which we also propose can be used as a tailoring of ISO 26262:2018 HARA.
- Including safety and the absence of other undesired traffic incidents in the same risk framework.
- Revisiting the discussion on social acceptance and how to relate to this topic when developing an ADS feature.

## II. Hazard Analysis and Risk Assessment

### A. HARA in ISO 26262:2018

The goal of a hazard analysis is to find all potential hazards, and the *operational situations* where, if the hazards occur, they may lead to an accident. In ISO 26262:2018, hazard is defined as "potential source of harm [injuries to humans] caused by malfunctioning behaviour of the item [function on vehicle level]". A risk assessment is made for each combination of hazard and operational situation, called hazardous event (HE), taking into account the potential *severity* of the hazard

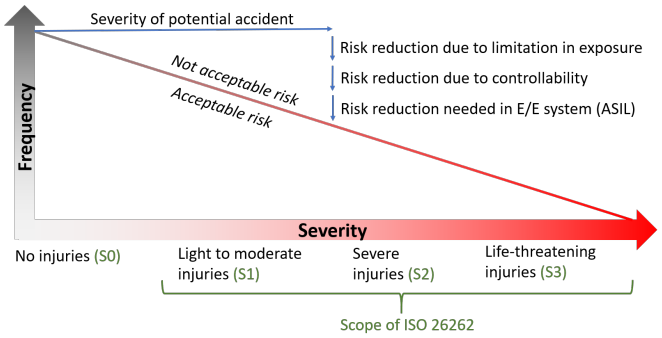


Fig. 1. Acceptable risk for accidents of different severity - ISO 26262.

in the operational situation, the *exposure* to this situation, and the *controllability* or possibility for persons involved to take actions to avoid an accident. Fig. 1[1] illustrates how the acceptable accident frequency (y-axis) is lower for more severe injuries (x-axis). Given the severity of a potential accident, controllability and limitation of exposure reduces the frequency of accidents for a certain HE. A discrete safety integrity level (ASIL) is assigned to the HE if further risk reduction is needed to reach the area of acceptable risk, and top-level safety requirements, safety goals, covering all HEs with an ASIL must be formulated.

The primary output from the HARA activity is a complete set of safety goals, where the connotation of 'complete' is that if all SGs are fulfilled, this implies that the item is functionally safe. There is an explicit request for a work product arguing for the completeness and consistency of the SGs, which is also the subject of a confirmation review with the standard's highest defined degree of independence. The reason for this rigor is that the rest of the reference lifecycle activities are dependent on a complete set of SGs with appropriate ASILs in order for the total integrity of the item to hold.

### B. HARA for an ADS

The objectives for a HARA are still valid when considering an ADS, even if we propose a significant tailoring of the activity. This means that we are still seeking for a set of vehicle-level safety requirements, which if all are fulfilled the ADS would be concluded as safe. However, there are at least four reasons why the details of the HARA activity as described in ISO 26262:2018 make it less suitable for an ADS. The assumption is that all relevant situations shall be considered, which means they are used as input to the process of generating SGs, which will in turn give input to the realization. In the ADS case, these relations are not evident. It becomes problematic to: *(1)* argue completeness of identified situations; *(2)* regard exposure to situations as given input; and *(3)* separately identify hazards from hazardous events as the source of harm. Furthermore, it might imply a too conservative design because of: *(4)* all considered frequencies of situational properties are seen as globally valid.

[1]The figure is adapted from [7] p.18, background material for the ISO 26262 standardization, explaining the proposed HARA and ASIL concept.

*1) Argue completeness of identified situations:* An ADS feature is a very complex function. When analyzing the potential consequences of different situations, it becomes evident that the number of situations to consider is virtually infinite, unless the feature has a very limited ODD. For a conventional E/E functionality used by a human driver, the number of relevant situations can be seen as bounded thanks to the limited complexity of what the user expects from it.

*2) Regard exposure to situations as given input:* What situations the ADS will be exposed to will depend on its decisions in previous situations. Therefore we can argue that an important part of an ADS feature's safety strategy is to avoid hazardous situations instead of making sure they can be handled. The fact that its exposure for certain situations will be design choice dependent needs to be considered in the analysis. This differs from a manually driven vehicle where the exposure used as input to the HARA can be based on, e.g., statistics on typical usage patterns, which is not affected by the analysed function itself.

*3) Separately identify hazards from hazardous events as the source of harm:* For a manually driven vehicle, the E/E functionalities are used by a driver who has the overall responsibility for safe operation. A violation is typically evident, and depending on the circumstances this may or may not lead to an unsafe situation. For example, a vehicle-internal fault leading to a reduced braking capacity of only 4 $m/s^2$ on dry asphalt, can be regarded as a hazard of a brake-by-wire functionality. If the ego vehicle at high speed is approaching a vulnerable road user (VRU), this violation can be unsafe because the driver expects a higher retardation when performing full braking. In the HARA, the objective is to determine how often there is a situation in which the driver needs to brake significantly harder than 4 $m/s^2$ to avoid an accident. For an ADS this is not an appropriate analysis. Firstly, we don't have to consider any absolute and constant braking capability as safety critical, as might be the case for a manually driven vehicle. We could say that as long as the tactical decisions know about the current actual braking capability, it should be possible to safely adjust the driving style accordingly. Secondly, how often we would need a certain braking capability depends on our tactical decisions. Regarding the situations as an input to the HARA, would then run the risk to introduce a circular proof reasoning. Impact on what situations that ego vehicle will be exposed to, might be the consequence of both safety-critical and non-safety-critical tactical decisions. If we for example give a general instruction to the ADS that braking harder than 3 $m/s^2$ is considered uncomfortable, we should assume that the need for braking significantly harder than 4 $m/s^2$ will be very rare. Furthermore, the design choices can elaborate a balance how much responsibility to achieve safety is put on reactive vs. proactive capabilities. E.g. more focus on proactive capability would result in less frequent situations where we need to brake significantly harder than 4 $m/s^2$ (i.e. apply a reactive measure).

*4) All considered frequencies of situational properties are seen as globally valid:* The frequency of many situational conditions of the real world are very dependent on time and place. For example the exposure to snow on the road is typically dependent on the season, and the frequency of pedestrians running across a street is most likely something that varies in time and space. It would be natural to allow the ADS to get applicable data for its current context, rather than statically do such coding in a HARA.

Further, for the activity of formulating safety goals, there are two principles we should follow:

- The complete set of SGs shall guarantee a safe behaviour.
- Each SG shall be formulated such that it can be efficiently refined and verified in the implementation.

The first criterion is strict, as being able to argue for completeness is essential in any safety case. The second is about finding a pattern to enable an efficient design. Note that the outcome of the HARA is partly a design choice, in the sense that how the resulting SGs are formulated can make the safety measures more or less easy to design.

Based on the discussion above, one can conclude that the promise of functionality for an ADS, i.e. the specified intended function including quality attributes such as safety, is of a different kind compared to features intended to aid a human driver. Rather than providing a specific functionality with well-defined performance, e.g. a certain braking capability, the promise of an ADS is more something like 'drive safely from point A to B'. Furthermore, this promise is not only towards the humans inside an ADS equipped vehicle, but also towards other road users. Given that the SGs shall reflect this promise, it is more difficult to find a well-defined unsafe behaviour to encode as SGs. One could imagine using e.g. rule-based goals such as 'not driving faster than the speed limit', which would certainly be relevant for perceived safety, but not necessarily for the frequency of accidents. It would both be difficult to show that a set of such proxy goals is complete, and to prove that they will actually result in safe behaviour.

## III. A Proposed Tailoring for ADS

Based on the observations in Sec. II-B, we propose to formulate safety goals to restrict certain accident/incident types instead of using a traditional HARA. Furthermore, each SG shall have an integrity attribute in the form of a guaranteed frequency, i.e. what is the maximum tolerated occurrence of violating this SG. We do this by first defining a quantitative risk norm, as described in Sec. III-A, containing defined limits of tolerated frequencies of different consequences related to their severity. Then a classification of incidents and their contribution to the QRN is made as detailed in Sec. III-B; each defined incident type will result in one SG. Carefully choosing the classification of incident types will enable showing both completeness and efficiency of the set of SGs.

### A. A Quantitative Risk Norm

The risk norm defines what is regarded 'sufficiently safe' in the design-time safety case top claim. As opposed to the risk-model in Fig. 1, QRN is about actual outcomes in terms of frequencies of consequences such as fatalities or severe injuries. Such a norm also has the potential to connect what

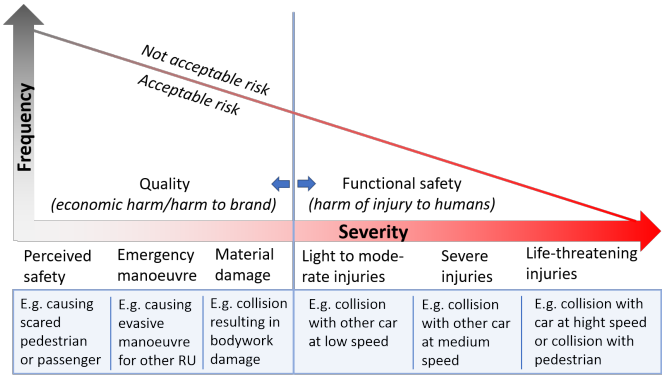


Fig. 2. Safety and incident quality - acceptable risk.

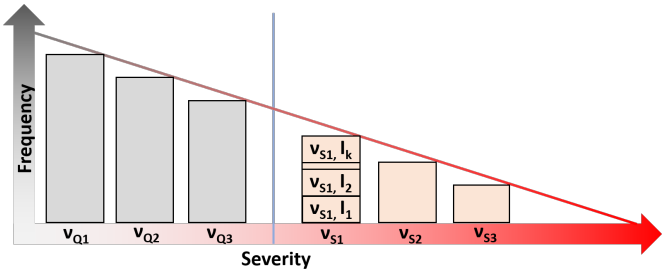


Fig. 3. A risk norm based on consequence classes and incident types.

is traditionally the concern of functional safety (tolerance frequencies related to severity of injuries, compare to the S-factor in ISO 26262) with perceived safety and quality requirements (tolerance frequencies of non-safety related consequences), as illustrated in Fig. 2. The blue box in the figure exemplifies with some types of incidents that would typically result in different consequences along the severity (x) axis. Including consequences in a wider sense reflects the fact that e.g. light rear-end collisions resulting in bodywork damage, or careless driving causing other road users (RU) to perform emergency manoeuvres, are also about avoiding unwanted traffic events, though in severity different than those resulting in injuries. In the figure, this is reflected by the fact that we will likely accept higher frequencies (y-axis) of quality-related consequences than those involving injuries, i.e. quality will be found on the left-hand side of the risk acceptance diagram.

In order to create a useful risk norm, the severity/criticality dimension is divided into a manageable number of discrete levels, or *consequence classes*, where each class receives a total norm frequency telling how often, at most, this kind of consequence is allowed to occur. This is illustrated in Fig. 3, where the consequence classes are denoted $\nu$. We do not suggest a specific number of consequence classes, it can be defined as deemed appropriate. For the example in the figure we have chosen three classes for safety, $\nu_{S_1}$ to $\nu_{S_3}$, and three for quality-related consequences, $\nu_{Q_1}$ to $\nu_{Q_3}$. Sec. III-B will show how these classes are used when defining SGs.

We use the same risk norm for the entire safety case. As we do not restrict the use of the ADS other than the ODD limits, the safety case needs to be valid inside the entire ODD regardless of where, when, and how the feature is used. What is safe enough for an ADS is yet an open question, and we don't want to hard-code any specific criteria for that in our proposal. On the one hand it will be a political upper limit of acceptance from the society and customers; and on the other hand, it should not contradict the lower claim limits understood as the state of the art in the industrial and scientific community.

### B. Incident Types and Safety Goals

Having established a QRN that defines what sufficiently safe means in our safety case, the next task is to create SGs that can both be shown to meet the defined limits for all consequence classes of the QRN, and be possible to create a technical solution for. We propose to use classification of incidents into a set of incident types, where each type will result in one SG. In the following, we use *incident* as the generic term when discussing both quality-related incidents and safety-related accidents[2].

If we base our safety goals on incident classification rather than a list of hazards and situations, the completeness criteria will, analogous to the HEs, apply to this classification. If there exist incidents which are not included in the classification, the safety argument will be flawed. However, we can guarantee completeness by making the classification scheme complete by definition, i.e. every theoretically possible incident belongs to one of the defined incident types. Fig. 4 shows an example of how such a classification may look. The types are defined based on two criteria; for each type it shall be possible to:

- show the contribution to each consequence class, and
- provide meaningful input to refined safety requirements.

Each type of incident ($I$) will contribute to one or several of the consequence classes ($\nu$), e.g. some collisions between the ego (ADS equipped) vehicle and a VRU will lead to a fatality, some to severe injuries, and some to light injuries. The first criterion means that to show that the QRN is fulfilled, it must be possible to assign the frequency contribution of each defined $I$ to every applicable $\nu$. One aid in achieving this should be to separate incidents according to their severities, so that each $I$ contribute to as few of the defined $\nu$ as possible. If we consider incidents where an ego vehicle collides with a VRU (we denote an incident involving these two actors Ego↔VRU), it would make sense to categorize them according to how the impact speed is likely to have different consequences. So for example[3], separating a collision between ego vehicle and VRU with collision speed at 17 km/h from a similar collision at 19 km/h might be too fine grained, but having two incident types for collision speeds below or above 10 km/h may be appropriate if the likelihood of severe injuries rises quickly above this limit. As a correctly assigned contribution

[2]While sometimes used with other connotations, several definitions treat accidents as a subset of incidents, e.g. [8], where incidents also include undesired events not leading to injury. In this paper we use this definition.

[3]Please note that all examples in this paper are made up for illustrative purposes only and not based on actual statistics, hence they should not be used in a real safety case!

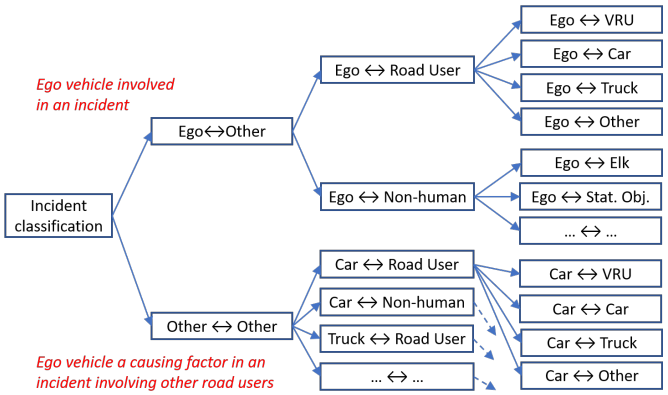


Fig. 4. Example incident classification.

of incidents to consequence classes is a vital part of the safety argument, it must be well substantiated; however this is a topic where much data and domain knowledge is available, e.g. from research and national traffic analysis databases such as [9]. In mathematical terms, the following must hold to show that the QRN is fulfilled:

$$\sum_{k=1}^{n} f_{\nu_j, I_k} < f_{\nu_j}^{(acceptable)} \quad , \forall \nu_j$$

where $\nu$ denote the consequece class, $I$ the incident type, $n$ the number of such types, and $f_{\nu_j, I_k}$ the frequency of incident $k$ within a consequence class $j$. $f_{\nu_j}^{(acceptable)}$ denotes the acceptable total incident frequency for consequence class $j$. This is illustrated for $\nu_{S_1}$ in Fig. 3.

The second criterion says that making the classification in a way that fit how requirements are expressed in the refined implementation makes it easier to find a reasonable argumentation that the SG is fulfilled. For example, distinguishing between object types would be appropriate only if the refined requirements on perception and prediction can be expressed by means of such object categorization. E.g. trying to separate collisions with pedestrians with haemophilia from collisions with all other pedestrians would not make sense, as it will likely not be possible to refine as requirements on perception. Another example could be that separating a frontal collision with a car from a side impact at the same speed may be useful given that we know the side impact is likely to be more severe, but it is only useful if we can design the ADS so that it can make use of this difference.

The expression shown above suggests that we can regard determination of the incident types and their integrity attributes (the limit frequencies) as a allocation process, where we must make sure that the budget we set on each $I$ must be such that the total allowed frequency is fulfilled for all $\nu$. An important aspect of this allocation is that defining the incident types to a certain extent will entail ethical considerations. For instance, even if the total acceptable frequency of fatalities is low in an ADS risk norm compared to accidents caused by human drivers, it will hardly be acceptable to create a set of SGs where all of these fatalities are assigned to an $I$: Ego↔Child, if it turns out to be more difficult to design for avoidance of collisions with children compared to adults.

Fig. 5 shows an example where Ego↔VRU has been elaborated to three concrete incident types:

- $I_1$ is where Ego approaches the VRU with $>$ 10 km/h when closer than 1 m (i.e. not a collision); a situation which is undoubtedly scary for the VRU, and potentially leading the VRU to perform some emergency action. The total frequency of this incident is determined to contribute with a certain percentage each to $\nu_{Q1}$ and $\nu_{Q2}$.
- $I_2$ is a collision with an impact speed $<$ 10 km/h, leading to either light ($\nu_{S1}$) or moderate ($\nu_{S2}$) injuries.
- $I_3$ is a collision with an impact speed of between 10 and 70 km/h, and therefore also has a contribution to the consequence of fatalities ($\nu_{S3}$).

The figure also illustrates how each $I$ contribute to the total budget for each $\nu$. If the total budget for a consequence class is not fulfilled, the budgets of some of the contributing incidents must be reduced. E.g. assuming it has been determined 70% of $f_{I_2}$ will contribute to $\nu_{S1}$ and 30% to $\nu_{S2}$, an improvement of $f_{I_2}$ will reduce the total incident frequency for these two consequence classes correspondingly, but result in an SG for $I_2$ which will be more challenging for the implementation.

We suggest that many of the incident types can be defined as an interaction between ego vehicle and `<object_type>` within `<tolerance_margin>`. The `<object_type>` is a complete and unique set. The `<tolerance_margin>` is for accidents telling the impact speed, and for quality-related incidents limits for distance and corresponding relative speed. Note that beside incidents involving ego vehicle, there are induced incidents involving at least one other traffic actor (lower part of classification in Fig. 4) which may be more difficult to clearly define. We so far surmise that incidents with more than two involved parties can be treated as the sum of several one-to-one incidents.

We can now formulate the safety goals for each of the defined incidents. For instance, the SG for incident $I_2$ from Fig. 5 would look like this:

| SG-$I_2$: |
| --- |
| Avoid collision Ego↔VRU, |
| with $0 < \Delta v_{collision} < 10$ km/h, to below $f_{I_2}$. |

## IV. The Solution Domain

Using the QRN approach to create SGs enables us to both show completeness of the safety goals, and make sure they do not inadvertently contain assumptions on e.g. exposure that may unnecessarily restrict the solution. Instead of hard-coding an assumed exposure in the HARA, we let the ADS have the ability to get relevant information about the environment and adapt the decisions accordingly. Such information can be from several kinds of origins like own sensors, remote sensors in infrastructure, cloud-based services, own stored statistics valid for specific locations and times, etc. The point is that

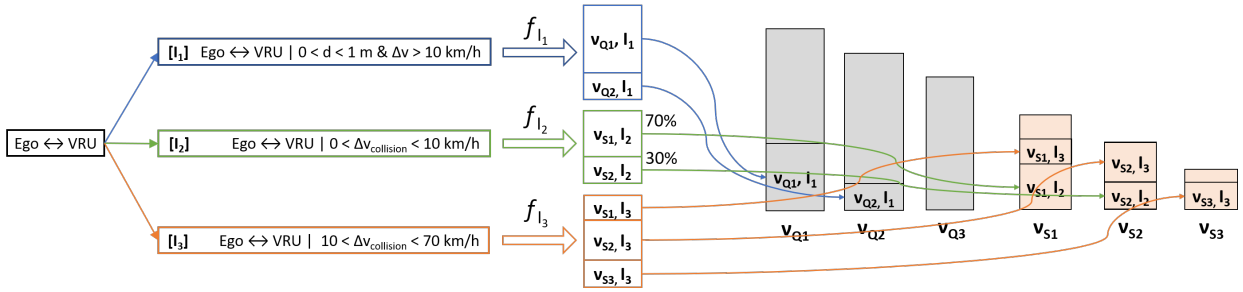


Fig. 5. Assignment of incident frequencies to consequence classes in the risk norm.

the adaptation to situational conditions can be much more precise than is the case if everything has to be expressed as one exposure in design-time HARA. However, SGs created with our approach will not contain information such as the situation and a failure mode, which is easily relatable to the architecture. E.g. safety goals from traditional HARA may contain concrete physical characteristics, such as in our earlier example a minimum allowed braking capacity, and also a fault tolerant time interval specifying the minimum time from the occurrence of a fault to a possible hazardous event. Determination of such characteristics needed to fulfil the SGs will instead need to be done in the solution domain, i.e. when defining an architecture capable of fulfilling the risk norm.

The work of fulfilling the SGs in ISO 26262 starts with a functional safety concept (FSC) where functional safety requirements are defined and allocated to logical elements. It will hence be up to the FSC to translate what it means to fulfil the risk norm, as expressed by the SGs, to the solution. As part of this work, strategies how to adapt to different situations/scenarios will likely play an important role; however, now with the purpose of fulfilling the risk norm rather than defining the risks. Note that part of the FSC may be to prescribe the importance of getting information about situational awareness and prediction. The more precise information that is available in run-time, the more likely it is that the tactical decisions can enable higher speed etc, still being able to guarantee a safe driving style. And a precondition for this precise information to have an influence on safety-critical decisions, is that its integrity is high enough for this purpose. All this reasoning becomes an important part of the FSC for an ADS. This analysis is also confined by the ODD. The role of the ODD in the safety argument is discussed by Gyllenhammar et al. in [5].

This way of working gives considerable freedom to define a safety strategy using trade-offs between performance of sensors/actuators (e.g. range, or performance in different environment conditions), driving style (e.g. cautionary vs. performance) and verification effort (e.g. adjusting critical ODD parameters to ease difficult verification tasks).

## V. A Quantitative Assurance Framework

The proposed tailoring of the HARA also opens up for further possible tailoring that might be useful in the safety argumentation of an ADS. This tailored HARA produces as an output a number of safety goals each having a quantitative integrity attribute (numeric value of maximum frequency for each incident type). In the refinement of these safety goals into allocated safety requirements in the ADS realization, the integrity values can be determined by traditional mathematical quantitative rules, instead of the qualitative ordinary rules of ISO 26262 of ASIL inheritance and ASIL decomposition. Furthermore, it also opens up for one budget to be met by all contributing causes, regardless whether they could be described as systematic faults in design of system, software or hardware; or as random hardware faults; or as 'performance limitations' by sensors or actuators. This, together with the case that an ADS always can be defined as safe, enables an integrated HARA, refinement strategy, V&V approach, and safety case structure, that includes all problems and concerns that today have been split in ISO 26262 and ISO PAS 21448 (safety of the intended functionality) [10], respectively. Please note that having a quantitative framework still allows qualitative evidence, so for example all the ASIL-oriented criteria defined in ISO 26262 to argue freedom from systematic faults would still be applicable.

There are a few things that are specific for an ADS compared to many features having been in scope of ISO 26262 and ISO 21448 so far. Firstly, an ADS can be defined as safe. This means that the intended function is safe, and there is no need for a separate HARA concerning the safety of the intended functionality. What is left as the subject for hazard analysis are all the cases where the real behaviour deviates from the defined behaviour. In many ADAS features this is a real problem which is addressed by the ISO 21448 HARA. For example, an automatic emergency brake (AEB), can introduce risks if fulfilling its specification. E.g. with a pedestrian in front of ego vehicle and a heavy truck behind, there is no obvious way of specifying a safe function avoiding both the possible collisions. Because the ADS is not a promise to the user how to reactively handle a certain situation as is the case with for example the AEB, we can define it as safe. An ADS can be specified as 'drive safely from A to B', which then can be deemed safe by definition. Hence we do not consider the separate concern of safety of the intended function, and for the ADS just one HARA for all risks can then be used.

Secondly, an ADS is much more complex than any tra-

ditional automotive feature. This means that the redundancy patterns to achieve a safe feature might also become quite complicated, and the rules of ASIL decomposition might not be fully applicable. For example, a common problem in ADS is to determine a drivable area in front of ego vehicle free from VRUs. A safety requirement on the aggregated block of sensing and prediction could then be not to overestimate such an area, with a very tough integrity attribute. In a quantitative framework this can be expressed as a frequency covering all possible causes why such an area would not become free from VRUs in reality. When decomposing this in several redundant sensing and prediction blocks, these can each get frequency attributes of a value that in traditionally ISO 26262 only would be in the QM range. By being agnostic to fault causes and being able to take into account redundancy contributions of just a few orders of magnitudes, also high integrity on vehicle level can be reached and argued for in architectures that are appropriate for the design of ADS.

Thirdly, the high complexity might make it hard to argue for the validity of the rules of ASIL inheritance. According to ISO 26262, a safety goal with attribute ASIL A can in theory be refined to thousands of software elements, each having dependent safety requirements which will inherit the ASIL rating. This means we can still claim ASIL A for the SG, despite having thousands of potential contributing ASIL A fault causes. This is of course not the intention behind the standard, which has the implicit assumption that the total complexity of the design contributing to one safety goal is limited. For most traditional automotive E/E features this might be true, and it is obviously a good design principle to separate complexity from integrity which makes this assumption on complexity valid. However, in ADS design it is far from evident that it is possible to limit the overall complexity of elements contributing to one safety goal. One way to make a safety case robust against violating such implicit assumptions, is to instead apply traditional mathematical rules for combining the effects of violating the safety requirements of the different elements in an architecture.

## VI. Related Work

Shortcomings of the ISO 26262 HARA for an ADS has been discussed before; e.g. how the lack of a human driver affects determination of controllability in risk assessment, since human passengers would not be ready and able to mitigate a failure [2], [11], [12]; and that the complexity of an ADS feature makes it more difficult to divide vehicle functionality into simpler items that can be analyzed independently [11].

Wardziński [13] describes two approaches to risk assessment: predetermined - based on analysis of possible accident scenarios, and dynamic - where the vehicle control system evaluates the risk of possible actions in real-time and selects an appropriate action based on the current situation. The HARA of ISO 26262:2018 would fall into the former category. Some of the authors of this paper previously proposed an iterative approach to predetermined hazard analysis for an ADS [12]. In this method combinations from situation and hazard classification trees are used to elicit HEs, followed by function refinement to redefine the scope of the function if the realization task is determined to be too difficult. This is repeated until a stable set of HEs is obtained. However, this method does not effectively address the problem of completeness of situations, and the attempt to define hazards on a tactical level proved difficult. Stolte et al. [14] use a similar iterative method. For their relatively simple use-case (in terms of both environment and functionality of the ADS) such methods may be feasible, but we argue that they will not scale well to more advanced ADS features. Dynamic approaches have been proposed by several authors. For instance, Khastgir et al. [15] proposed a framework where tactical decisions are used to reduce risks and seen as a redefinition of the function, leading to a real-time update of the HARA; Gleirscher and Kugele [16] formalized a way to construct planning (decision) models from hazard analysis; and Trapp et al. [17] presented a framework for dynamic risk analysis where the system dynamically can determine which safety goals can be fulfilled at any point in time. A missing piece in these approaches is the lack of a clear goal for what a safe system is. The QRN approach instead begins with determining risk acceptance, and we rather see the fulfillment of the established risk norm as part of the solution than a HARA; however, clearly the realized system needs dynamic risk awareness in order to meet the SGs, i.e. comparable to Wardziński's dynamic risk approach.

Similarly to us, Neurohr et al. [18] observes that HARAs as described in ISO 26262 and ISO PAS 21448 have weaknesses when it comes to accounting for the behavior of the ADS as well as the problem of completely characterizing the potential situations. They also propose an alternative approach addressing what is covered in both standards and takes into account the potential of using ODD restrictions and requirements on behaviour to reduce risks; these are called risk mitigating measures (RMMs). Their approach is scenario-based and builds on known techniques (HAZOP) where experts, with the help of data-driven scenario analysis, derive tables of relevant hazardous scenarios. Rather than decoupling the problem and solution domains they advocate an iterative process where the HARA and design (RMMs) are repeatedly improved until the risk is judged to be tolerable. They propose acceptable risk should be bounded by regulatory requirements on exposure to such hazardous scenarios. However, the difficulty of showing completeness of scenarios and thus the top-level safety requirements remain with this method. We believe the clear separation of problem and solution domains given by establishing a risk norm gives several important advantages, i.e. we can show completeness of the safety goals as well as make clear what the expectations with regard to total frequencies of consequences of different severity are. As mentioned in Sec. IV we rather see scenario analysis such as the one presented in [18] as an important tool in the realization phases.

Another approach is the responsibility sensitive safety (RSS) model [19], which defines a mathematical model formalizing a driving policy with the goal that the ego vehicle should never cause an accident, and also drive with care to be

able to compensate for reasonable mistakes of other road users, without being overly cautious. This model, essentially, specifies a solution based on what is assumed to be acceptable driving behaviour, but without considering incident frequencies or different severities. Schöner [20] similarly takes the route of defining what safe driving means in terms of 'good enough' behaviour for an ADS. While such work is relevant to consider when implementing rules for the ADS strategical/tactical planning, we believe it is imperative to both take the different consequences into account and to consider the number of incidents we can tolerate. In other words, these models may be part of the solution, provided that they achieve the safety goals determined e.g. using the QRN approach.

## VII. Conclusions

Main obstacles to HARA for an ADS when using operational situations as part of the analysis are: that the virtually infinite number of operating conditions requiring action to minimize risk makes a completeness guarantee difficult; and that the behaviour of the ADS can affect the determination of exposure to such risks. Therefore we propose the QRN approach, which instead focuses on incidents and acceptable frequencies of consequences with different severity. As shown in this paper, the need to define operational situations to create safety goals is eliminated with this method, making a completeness guarantee possible. The need to analyze situations/scenarios is confined to the solution domain, which seems appropriate given that what are relevant situations is, to a large extent, implementation-dependent. Further advantages to the approach are that: both safety-related and other unwanted traffic incidents can be included in the same framework; and since the risk norm is decoupled from the implementation the approach is advantageous for handling variability (e.g. in product lines) since the same risk norm can be used for many variants. I.e., while there may be some variability in the frequency allocation for each incident type (as solutions for variants may have different characteristics) the total acceptable risk for each consequence class will be the same. Several of the co-authors represent companies developing ADS features. We believe the approach is suitable for such systems, and could be used as a tailoring of the ISO 26262 lifecycle, replacing the HARA described in that standard as well as the one of ISO PAS 21448.

We are aware that some aspects of the QRN approach may seem controversial, perhaps specifically to explicitly set goals on the frequencies of accidents of different severity (essentially saying *we're allowed to kill and injure these many persons per operational hour*). In current standards, this is more implicit in the HARA framework and integrity levels, but it may hide weaknesses, e.g. the problems with ASIL decomposition and inheritance rules discussed in the paper, or potential inconsistencies when using several standards to ensure safety, e.g. ISO 26262 and ISO PAS 21448. In fact, we believe moving towards a quantitative assurance framework altogether (with some qualitative parts such as process arguments), and not just for the safety goals, may be good idea.

## References


[1] SAE, "SAE J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2018.

[2] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE International Journal of Transportation Safety*, vol. 4, no. 1, pp. 15–24, Apr. 2016.

[3] M. Martínez-Díaz and F. Soriguera, "Autonomous vehicles: theoretical and practical challenges," *Transportation Research Procedia*, vol. 33, pp. 275–282, Jan. 2018.

[4] ISO, "ISO 26262:2018 Road vehicles – Functional safety," 2018.

[5] M. Gyllenhammar, R. Johansson, F. Warg, D. Chen, H.-M. Heyn, M. Sanfridson, J. Söderberg, A. Thorsén, and S. Ursing, "Towards an operational design domain that supports the safety argumentation of an automated driving system," in *Proceedings of the 10th European Congress on Embedded Real Time Systems (ERTS)*, Toulouse, France, Jan. 2020.

[6] IEC, "IEC 61882 hazard and operability studies (HAZOP studies)—application guide," 2016.

[7] ISO, "N028: ISO TC22 SC3 WG16: Functional Safety - Presentation of Part 3 'Concept Phase'," Nov. 2005.

[8] Safeopedia, "Definition - what does incident mean?" [Online]. Available: https://www.safeopedia.com/definition/384/incident-occupational-health-and-safety

[9] M. Melkersson and B. Tano, "Trafikanalys - road traffic injuries 2018," 2019.

[10] ISO, "ISO/PAS 21448:2019 Road vehicles – Safety of the intended functionality," 2019.

[11] H. Martin, K. Tschabuschnig, O. Bridal, and D. Watzenig, "Functional safety of automated driving systems: Does ISO 26262 meet the challenges?" in *Automated Driving*, D. Watzenig and M. Horn, Eds. Cham: Springer International Publishing, 2017, pp. 387–416.

[12] F. Warg, M. Gassilewski, J. Tryggvesson, V. Izosimov, A. Werneman, and R. Johansson, "Defining autonomous functions using iterative hazard analysis and requirements refinement," in *Proceedings of SAFECOMP 2016 Workshops, ASSURE, DECSoS, SASSUR, and TIPS*. Springer, Cham, Sept. 2016, pp. 286–297.

[13] A. Wardziński, "Safety assurance strategies for autonomous vehicles," in *Proceedings of the 27th International Conference on Computer Safety, Reliability, and Security (SAFECOMP)*. Springer, 2008, pp. 277–290.

[14] T. Stolte, G. Bagschik, A. Reschka, and M. Maurer, "Hazard analysis and risk assessment for an automated unmanned protective vehicle," in *Proceedings of IEEE 2017 Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1848–1855.

[15] S. Khastgir, H. Sivencrona, G. Dhadyalla, P. Billing, S. Birrell, and P. Jennings, "Introducing ASIL inspired dynamic tactical safety decision framework for automated vehicles," in *Proceedings of IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[16] M. Gleirscher and S. Kugele, "From hazard analysis to hazard mitigation planning: The automated driving case," in *Proceedings of NASA Formal Methods Symposium*. Springer, 2017, pp. 310–326.

[17] M. Trapp, D. Schneider, and G. Weiss, "Towards safety-awareness and dynamic safety management," in *Proceedings of the 14th European Dependable Computing Conference (EDCC)*. IEEE, 2018, pp. 107–111.

[18] C. Neurohr, B. Kramer, M. Büker, E. Böde, M. Fränzle, and W. Damm, "Identification & quantification of hazardous scenarios for highly automated driving," 2020. [Online]. Available: https://www.doi.org/10.13140/RG.2.2.26704.66564

[19] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[20] H.-P. Schöner, "'How good is good enough?' in autonomous driving," *Reasearchgate preprint:330452811*, 2019.