Gyllenhammar, M., Rodrigues de Campos, G., Sandblom, F., Törngren, M., Sivencrona, H. (2022)
Uncertainty Aware Data Driven Precautionary Safety for Automated Driving Systems Considering Perception Failures and Event Exposure
In: *Proceedings of IEEE Symposium on Intelligent Vehicle* Aachen, Germany

Uncertainty Aware Data Driven Precautionary Safety for Automated Driving Systems Considering Perception Failures and Event Exposure

Magnus Gyllenhammar [1,2], Gabriel Rodrigues de Campos [1], Fredrik Sandblom [1],
Martin Törngren [2] and Håkan Sivencrona [1]

***Abstract*— Ensuring safety is arguably one of the largest remaining challenges before wide-spread market adoption of Automated Driving Systems (ADSs). One central aspect is how to provide evidence for the fulfilment of the safety claims and, in particular, how to produce a predictive and reliable safety case considering both the absence and the presence of faults in the system. In order to provide such evidence, there is a need for describing and modelling the different elements of the ADS and its operational context: models of event exposure, sensing and perception models, as well as actuation and closed-loop behaviour representations. This paper explores how estimates from such statistical models can impact the performance and operation of an ADS and, in particular, how such models can be continuously improved by incorporating more field data retrieved during the operation of (previous versions of) the ADS. Focusing on the safe driving velocity, this results in the ability to update the driving policy so to maximise the allowed safe velocity, for which the safety claim still holds. For illustration purposes, an example considering statistical models of the exposure to an adverse event, as well as failures related to the system's perception system, is analysed. Estimations from these models, using statistical confidence limits, are used to derive a safe driving policy of the ADS. The results highlight the importance of leveraging field data in order to improve the system's abilities and performance, while remaining safe. The proposed methodology, leveraging a data-driven approach, also shows how the system's safety can be monitored and maintained, while allowing for incremental expansion and improvements of the ADS.**

I. Introduction

Providing evidence for the safety of Automated Driving Systems (ADSs) is a difficult challenge and, arguably, one of the key reasons for the lack of wide-spread market deployment of such technology. Before its deployment, one needs to show the fulfilment of the safety requirements of the system and further present evidence for a predictive conclusion of this fulfilment. In detail, this requires evidence and thorough safety argumentation, resulting in a predictive as well as reliable safety case. Some of the evidence required for this will be hard to gather before a larger scale deployment. Further, it is commonly agreed that it may be infeasible to collect all the necessary evidence from in-traffic testing alone [1], [2]. Therefore, in relation to gathering evidence, there is a need for new approaches (methods and tools) and architectures for ADSs with a high level of automation [3]. One way to overcome this issue is relying on mathematical/behavioural models, even though such solutions come at a price. Indeed, each model used requires a set of assumptions. Carefully selected assumptions could yield a safe system, but might result in an overly conservative system with inferior performance metrics. Alternatively, if inadequate models, assumptions or simplifications are considered, this could potentially result in fatal outcomes upon the deployment of the ADS in real traffic conditions. Let us consider some examples where assumptions and estimations impact the design and development of ADSs.

For instance, Mobileye's Responsibility-Sensitive Safety (RSS) approach [4] implicitly relies on assumptions regarding the vehicle's braking capabilities, as well as those of the surrounding vehicles [5]. While considering such capability models is important, not adequately accounting for the details in relation to real operational conditions could yield safety issues, as discussed in [5]. Conversely, using conservative estimates could, on the other hand, yield a system that is unable to operate due to an overly pessimistic view of the system's capabilities [6].

Similarly, scenario models, used for scenario-based testing, verification and validation (e.g. [7], [8], [9], [10]), also heavily rely on assumptions and estimations. On the one hand, making wrong or inaccurate assumptions will likely impact safety, since a non adequate consideration of corner cases and rare events for testing and validation purposes would result in overly optimistic assessment results and, worst case scenario, in the deployment of an unsafe system. Overemphasis on such rare events, on the other hand, would increase the risk of keeping a safe system off the streets, representing a missed opportunity of reducing crash rates and saving lives as a consequence.

Also, in the context of incremental deployment of the ADS to new use cases, that could be achieved, for instance, through matching the system's Operational Design Domain (ODD) with the required operating conditions of new potential use cases, as suggested in [11], it is essential to make an appropriate use of estimations so to maximise performance while operating safely. In-line with industry practises, the safety risks created by a given set of estimations can be mitigated by making use of *conservative* assumptions (in a statistical sense), and leveraging the available field data. Consider, for example, the requirements of proven-in-use arguments, which according to the ISO 26262 should be

This research has been supported by the Strategic vehicle research and innovation (FFI) programme in Sweden via the SALIENCE4CAV project (ref 2020-02946), as well as by the Wallenberg Al, Autonomous Systems and Software Program (WASP) of the Knut and Alice Wallenberg Foundation.

[1] Zenseact, Göteborg, Sweden. Email: {name.surname}@zenseact.com

[2] Mechatronics division, KTH Royal Institute of Technology, Stockholm, Sweden Email: gyllenha@kth.se, martint@kth.se

accepted only if the data supports the requirement with at least a 70% one-sided confidence interval [12, §14.4.5.2.4].

In this paper, we propose a design and monitoring methodology for determining safe driving policies while accounting for perception failure and event exposure rates, as illustrated in Fig. 1, and discussed in Sec. V. In particular, we focus on the impact that continuously updated estimates of these rates may have on the ADS's safe driving policy. This exploration is achieved by deriving a driving velocity, see Fig. 1, such that the prescribed safety requirements are fulfilled given the field data and prior knowledge. Here, we assume that the safety requirements are provided in the form of a Quantitative Risk Norm (QRN) [13], as discussed in Sec. II. The safe driving velocity for a given road segment is derived using the estimated arrival rate of an adverse event, in combination with the estimations of the failure rates of the perception system at different range sections (in-line with the proposal of [14]). The rates are modelled as random variables and estimates are calculated based on the posterior distribution of the rates given (simulated) field data. In this way, estimates change as a consequence of receiving more data, and so does the allowable velocity. However, the confidence that the QRN is met remains constant.

The contributions of this paper are five-fold, and detailed as follows:

- **A methodology** enabling continuous updates of the ADS's safe driving policy, given field data,
- **Maintaining confidence in safety claims** while making continuous updates of the ADS,
- **An uncertainty-aware** formulation of the ADS functionality providing a basis for realising such functions, where i) perception system failures (e.g. false positive detection of free space) and ii) exposure rates with respect to an adverse event, are considered for deriving a safe driving policy fulfilling the safety claims,
- **An analysis** of the impact on the safe driving policy of the ADS, from the continuously incorporation of field data to update the statistical models underpinning the uncertainty awareness, and
- **A numerical evaluation** of the proposed methodology on an end-to-end use case, including the derivation of quantitative safety requirements from accident statistics.

The remainder of the paper is organised as follows. Sec. II introduces the definition of safety considered in this work. This is followed by Sec. III, where the considered perception and sensing models are introduced. Field monitoring of the considered quantities, i.e. event exposure rate and perception failure rates, is proposed in Sec. IV, while Sec. V presents the proposed methodology (depicted in Fig. 1) for deriving a safe driving policy given the considerations from the previous sections. Finally, the considered use case and numerical results are presented in Sec. VI, and the insights from the paper are summarised in Sec. VII.

II. A Definition of Safety in an ADS Context

Safety can be defined as the absence of unreasonable risk or, alternatively, as the absence of adverse consequences. The acceptable safety of an ADS can also be considered in relation to a positive-risk balance [15] and the driving performance of human drivers [16], [2], [1]. Furthermore, one could use the term safety in a broader context, in order to also include quality considerations, e.g., avoiding not only events that result in injuries (safety driven) but also events with material damage or perceived risky situations (quality driven), just as discussed in [13, p. 2]. While the safety definition and the different related aspects is a challenging topic by itself, we will not however address these details here. Rather, we allow ourselves to work on one abstraction level above this, and consider that concrete and quantitative safety requirements are expressed as a Quantitative Risk Norm (QRN), as introduced in [13]. In particular, the QRN establishes quantitative requirements on the allowed probability of undesired consequences of operating the system. This would go beyond functional safety standards and their integrity requirements on hardware, with the difference that the QRN provides a system safety perspective. The QRN can be split into quantitative requirements on different accident types and different consequence classes (e.g. severity). Rather than requiring one specific integrity, with which collisions should be a avoided, the QRN also accounts for the capabilities of the passive safety of the operated vehicle, as well as allowing probabilistic allocation to different consequence classes given an incident. This is increasingly important considering the stochastic nature of the operating environment of the ADS, as well as some of the deployed algorithms such as neural networks, for example, which can drastically change their outputs given small fluctuations to the inputs [17]. For the use case considered in this paper, and for the sake of the clarity and simplicity of the discussion, we only consider a deterministic transfer function from incident to consequence class, where the resulting consequence is derived from the impact velocity directly. In other words, each impact velocity contributes to one, and only one, consequence class. The concrete QRN levels and consequence classes considered here are presented in Sec. VI-A.

III. Sensing and Perception Modelling

The necessity for the ADS to provide information about the surrounding free-space, at different integrity levels, has already been highlighted and discussed in [14]. In particular, is has been suggested that the detected free space should be given dynamically together with an estimated integrity, thus promoting dynamic risk assessment, as proposed in [18]. Such a model could be determined by using the real-time operating conditions and status of the ADS in order to estimate its perception capabilities. For example, the perception integrities could be derived in relation to the operating conditions as, for example, conditional probabilities. However, one could also consider a perception system, the performance of which is determined at design-time, where different combination of sensors and algorithms provide free-space at different ranges. To avoid unnecessary complexity in the discussion in this paper, we will consider the latter. Hence, we consider a perception system, potentially formed

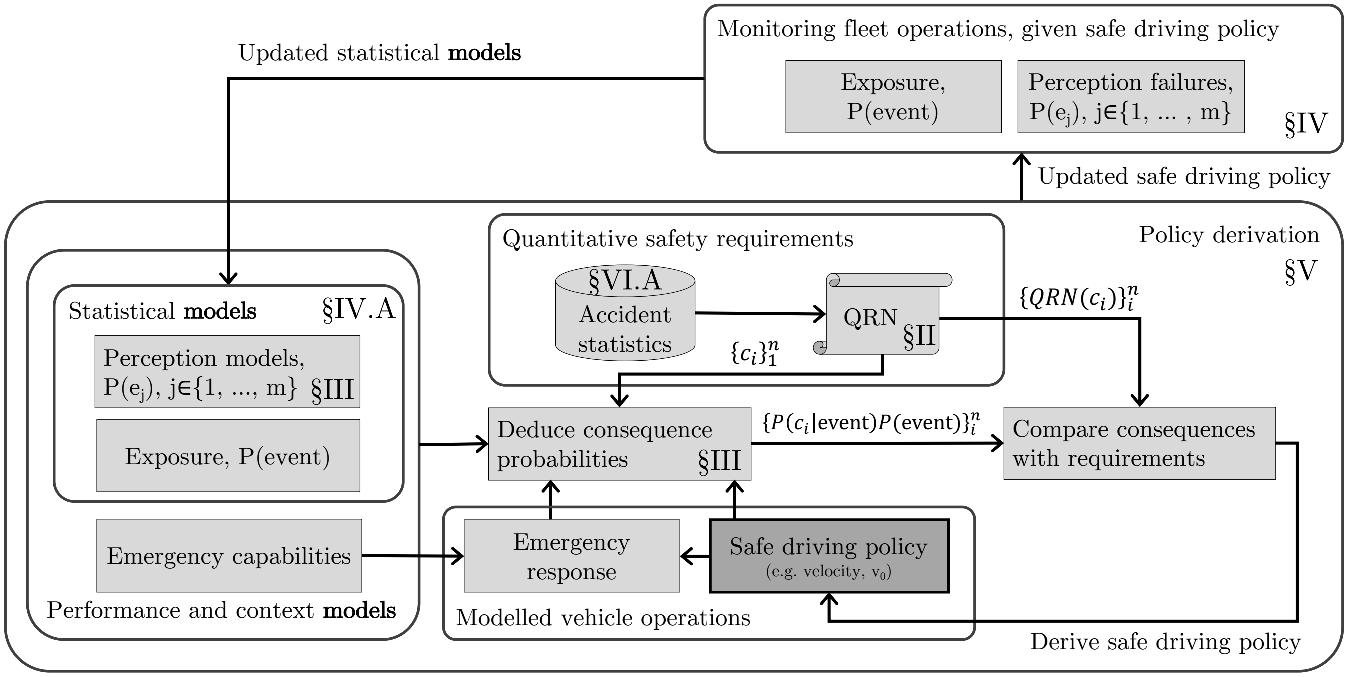

Fig. 1: Illustration of the proposed methodology for deriving a safe driving policy. The objective is the maximisation of the safe driving velocity v_0 with maintained confidence that the QRN is met. $P(e_j),\ j \in \{1,\dots,m\}$ denote the perception failure probabilities for different range sections and $c_i,\ i \in \{1,\dots,n\}$ the different consequence classes. References to the corresponding sections in this paper are denoted by the symbol "§".

by a set of different subsystems, that is responsible for the free-space detection in different ranges sections. Each range section is associated with a performance integrity, i.e., a failure rate in terms of false positive reporting of free space. Further, we limit the discussion to direct detection of free-space and omit any effects arising from incorporation of predictive estimates of the free-space.

A. Impact Speeds and Detection Distance

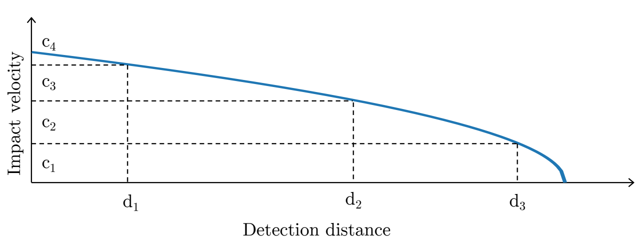

Fig. 2: Illustration of the impact velocity's dependence on the threat detection distance. When the object is detected, an evasive action, established in terms of an immediate braking manoeuvre, is assumed to be triggered. Depending on the vehicle's original speed, different consequence classes (denoted here by c_i (for $i \in \{1,\dots,4\}$)), representing for instance different severity levels, can be deduced.

Let us consider an adverse event where an object appears in front of the vehicle. Assuming full maximum braking capability ($a_{max} < 0$) at the time of detection and zero response delay, the vehicle speed will be reduced, from a given initial speed (v_0), according to the following expression:

$$v_{\text{impact}}(d) = v_0\sqrt{1 + \frac{2d\ a_{max}}{v_0^2}},$$

for different detection distances d, as depicted in Fig. 2. Subsequently, assuming that a mapping between the impact speed and the consequence classes of the QRN exists, one can simply read off the detection distances that would result in a specific consequence class. For example, in Fig. 2, any object detected within the range $[d_1, d_2]$ would result in consequence c_3.

B. Modelling of Perception Failures

For the sake of clarity of the discussion, let us assume that the object can appear, with an equal probability, anywhere in front of the vehicle up to a maximum distance (d_{max}), i.e., the object detection distance d is uniformly distributed such that $d \sim Uni(0,\ d_{max})$. Thus, without accounting for failures of the perception system, the probability of reaching one of the consequence classes of the QRN is equal to the length of the corresponding section (as exemplified in Fig. 2) divided by d_{max}. As an example, following the illustration of Fig. 2, given an unexpected event (i.e., unexpected appearance of an obstacle), the probability of reaching the consequence class c_3 is given as $(d_2-d_1)/d_{max}$.

However, to understand the impact of perception failures (i.e. inaccurate reporting of free-space) over the different range sections, one needs to consider how much of the probability mass that is shifted from each range section, and where it ends up. For illustration purposes, Fig. 3 depicts this process for three range sections, denoted by r_1, r_2 and r_3. The perception failures are assumed to be correlated across all range sections, in such a way that if the perception system fails at a closer range section (with higher integrity), then so it does for all range sections farther away as well. For r_1 this means that a thin slice from the entire range up to d_{max}, with thickness corresponding to the associated failure rate e_1, would result in a crash at the vehicle's initial velocity, as illustrated by the vertical blue bar marked with * in Fig. 3. More generally, upon a failure of the perception system to detect an object in one given range section (and thus falsely reporting free space when it is not), we consider that the object will instead be detected at the boundary of the previous, closer range section. See, for example, that the horizontal, red-white striped box starting from and including r_3, is translated to a vertical box at the delimiting boundary of r_2 and r_3. The size of the shifted probability mass corresponds to the failure rate of the corresponding range section. Any object appearing beyond the perception range limit, corresponding to the grey area to the right of r_3, will be detected at the perception range limit, corresponding to the grey vertical bar at the perception limit in Fig. 3. Note that the filled boxes, in the bottom of the figure, correspond to the remaining probability mass that is uniformly spread across the range section.

IV. Field Monitoring for Model Updates

Using customer driven vehicles, rather than dedicated test vehicles, as means to provide relevant and sufficient data for proving safety is a desirable approach from a cost and efficiency point-of-view. Further, there is, arguably, a strong need for monitoring the performance and capabilities of the ADS while in operation. This is an important topic not only for risk mitigation purposes, but also for the validation and updating of the models used for the system's design and development, as discussed in Sec. I. In the use case considered here, the models are selected in such a way that they can be updated through *open loop* monitoring, i.e., there is no need for the ADS to plan and control the paths of the vehicle in order to update the models. Rather, both the models of the perception failure rates and the exposure rate to adverse events can be inferred from all vehicles with the appropriate perception system and sensor capabilities. Despite its pertinence, the topic of closed-loop monitoring will not be discussed further in this paper. It remains, though, an important corner stone for the design of a performant ADS, and an enabler for the construction of a reliable and predictive safety case.

In this work, we consider two underlying aspects:

i) the exposure rate to adverse events, and
ii) the failure rates of the perception system, the model of which is discussed in Sec. III.

The first type i) requires a means to measure: (a) accidents due to a given event type, and (b) exposure to the given event type. Statistics of accidents resulting from the event should be straightforward to gather, as the event resulted in a clear consequence. On the other hand, capturing each event exposure would require a dedicated detection system. If such a detection system is in place, however, any occurrence of the event can thus be shared across the fleet of vehicles in order to update the (global) exposure model (across the vehicle fleet). Regarding the perception failures ii), there is a need for a different solution. Stationary objects will eventually be close to the vehicle and thus the failure rates of the farthest range sections can be validated using the higher integrity of the closer range section(s). For moving objects, one could consider a similar approach, where the scenes are post-processed to align the detections and tracks both backward and forward in time, as detailed in [19]. Thus, it is possible to create a reference, towards which the perception system can be compared, and consequently to infer the failure rates in each range section. The integrity of the closest range section can also be deduced from the accident rate of the system, as failures to report detections near the vehicle would likely lead to a collision.

In this paper, we assume that solutions to make accurate updates to both i) and ii) are available. Furthermore, we assume that the number of events that the vehicle is exposed to, but fails to detect, is negligible.

A. Modelling Occurrences – Arrival Processes and Bayesian Inference

A common way to model the number of independent occurrences of an event, in a given period of time, is to assume that the number of occurrences follows the Poisson distribution. The probability mass function, $p(y|\lambda)$, characterised by the rate parameter λ, describes the probability for observing y occurrences in that time-interval.

Assuming a large number of observations, $\mathbf{y} = [y_1, \ldots, y_N]$, it is possible to calculate the maximum likelihood estimate of the rate parameter,

$$\hat{\lambda} = \frac{1}{N} \sum_{i}^{N} y_i.$$

However, in this paper, we are not only interested in an estimate of the rate λ. Rather, we seek to establish the influence that the uncertainties associated to the rate have on the driving policy. In that context, a *safe* policy is defined by an allowed safe speed that is conservative with respect to the risks deduced from the uncertainties associated with the estimated rates of perception failures and the exposure rate to adverse events.

Furthermore, assuming repeated observations from a fleet of vehicle using the aforementioned monitoring system, such observations are to added to the prior belief of each rate and will subsequently impact the allowed safe driving policy derived, as shown in Fig. 1.

In such a context, it is straightforward to use a Bayesian model, where the prior belief is updated with the observations to get the posterior belief through Bayes' rule [20]. With the appropriate selection of a prior model, the incorporation of

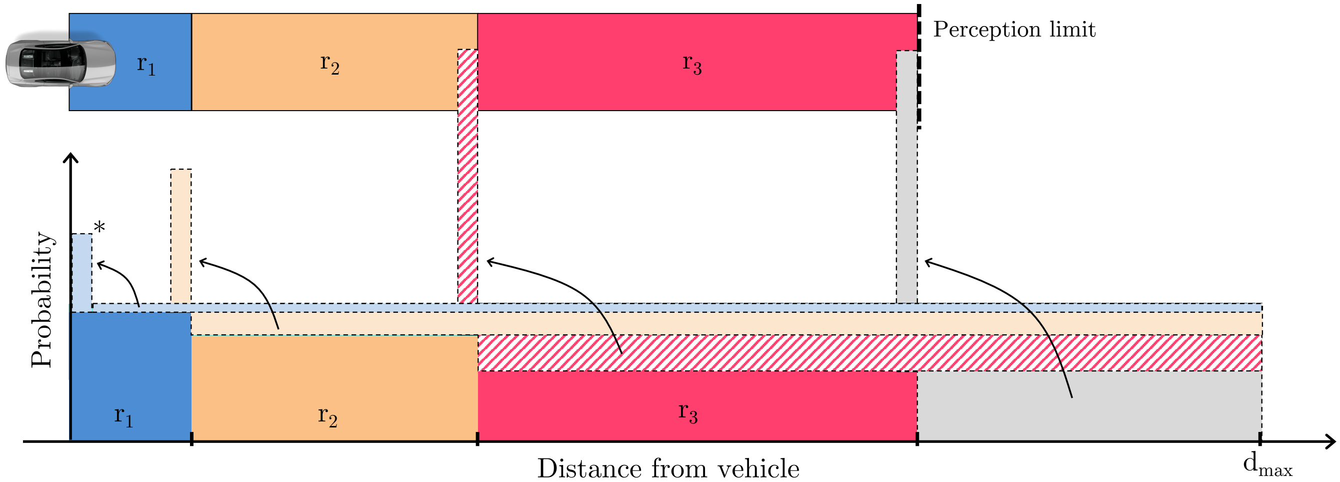

Fig. 3: Illustration of the perception failure modelling. The resulting distribution of probability mass, after accounting for the perception failures, is depicted in the bottom. Considering perception failures in each range section (r_1, r_2, and r_3), a portion of the probability mass is shifted from the (original) uniform distribution to the vertical boxes at the delimiting boundary between the current and the previous range section. Note that, for visual reasons, the areas of the vertical and horizontal boxes are not depicted to scale in the illustration. Furthermore, the vertical boxes should be point masses, but shown here with a width for illustrative purposes.

additional data through Bayesian inference yields a model of the same family as the prior, but with updated parameters. Such a prior model is called a conjugate prior and, for the Poisson distribution, corresponds to the gamma distribution defined by parameters α and β. One can therefore describe the distribution of the rate parameter as

$$\lambda \sim \Gamma(\alpha, \beta),$$

which after the incorporation of n new observations, $\mathbf{y} = [y_1, \ldots, y_N]$, yields the following updated parameters:

$$\hat{\alpha} = \alpha + \sum_{i}^{N} y_i, \qquad \hat{\beta} = \beta + N. \tag{1}$$

In general, one might want to leverage federated learning techniques to update these models. This means that models are asynchronously and locally updated, in each vehicle of the fleet, and that only the resulting model is shared with a central entity in order to preserve privacy and reduce transmission costs [21]. The resulting Bayesian update, formulated in equation 1, is particularly suitable for this purpose as there is no dependence to the order of how the samples are incorporated into the model. Thus, perfect inference can be achieved despite asynchronous and distributed updates of the model, e.g. following the approach of [20].

In this work, the inference is assumed to start without any previous data and it is therefore appropriate to use a non-informative prior in order not to skew the subsequent inference by inappropriate initial assumptions. The Jeffreys prior is an example of a non-informative prior, and for the Poisson distribution it is $p(\lambda) = \sqrt{\lambda^{-1}}$ [1]. The posterior probability distribution can then be shown to be gamma distributed $\lambda \sim \Gamma(\hat{\alpha}, \hat{\beta})$, where the parameters are calculated from observations $\mathbf{y}$, as shown in equation (1), using $\alpha = 1/2$ and $\beta = 0$. The corresponding posterior probability density function, $p(\lambda|y)$, can be used to account for uncertainties in the estimation of the rate λ. For deriving a safe driving policy and obtain the results in Sec. VI, we choose an estimate $\hat{\lambda}$ such that

$$\int_0^{\hat{\lambda}} p(\lambda|y)\; d\lambda = c_\lambda, \tag{2}$$

for $c_\lambda = 75\%$. While no particular c_λ is recommended, the authors want to point out that a higher value represents a more conservative approach and higher confidence that the QRN is met.

Note that for the results presented in the sequel, the Jeffreys prior is assumed for each of the three perception failure rates $P(e_j)$, $j \in \{1, 2, 3\}$ (corresponding to r_1, r_2, and r_3, see Fig. 3), and for the event exposure rate (see Sec. IV). A reader interested in learning more on the concepts outlined above is referred to [22].

V. Deriving a Safe Driving Policy

Safe planning is one of the most challenging tasks in the development of ADS, and naturally a cornerstone for the development and deployment of such technology on the roads. In the scope of deriving a safe driving policy, many different approaches have been proposed in the literature, each with different focuses and principles. Of particular interest for this paper is the notion of Precautionary Safety introduced in [23], where a set of principles and a design methodology for safety policy design are proposed. Under the proposed

[1]Note that although $\sqrt{\lambda^{-1}}$ is not a proper distribution, the posterior distribution is proper given at least one data point.

design principles, safe driving policies are established and adapted with respect to the vehicle's capabilities, external conditions and knowledge of human mistakes, in order to satisfy ambitious quantitative safety requirements.

Here, the approach proposed in [23] for the derivation of a safe driving policy of an ADS is used to derive a safe *driving velocity* which, without loss of generality, increases the clarity of the discussion of this paper. In a complementary way to [23], we focus our attention on the impact of perception failures rates, exposure rates to adverse events and field monitoring thereof. The actuation aspects are not explicitly considered in this work, and a simpler actuation model is assumed (i.e., instantaneous braking without accounting for any response time of the system). Note, though, that such assumptions are only taken for the sake of space and clarity of the paper, and do not compromise the generality of the proposed methodology and the results shown. The reader can, however, refer to [23] for further details on how actuation capability models can be considered for the design of safe driving policies.

Based on the principles of [23], we propose the methodology illustrated in Fig. 1, for the derivation of a safe driving policy. For a given an initial velocity v_0, and the estimated perception failure rates (described above in Sec. III), one can deduce the fraction of the events $P(c_i|\text{event})$ that, if the corresponding event occurs, contributes to each consequence class c_i of the QRN. Subsequently, the resulting accident probabilities are multiplied with the event exposure $P(\text{event})$, and compared with the prescribed QRN levels. Following this process the maximum safe driving velocity that maintains safety with respect to the prescribed QRN can be identified.

The proposed methodology provides a means to evaluate the safety of a given driving policy v_0. By using this evaluation as a feedback signal, it is possible to derive of a safe driving policy v_0, with respect to the estimated perception failure rates and the estimated event exposure rate. In Fig. 1, the methodology of deriving a safe driving policy, together with that of deducing the QRN-levels from accident statistics, is presented. Further, statistical models of the perception failure rates and the event exposure rate are updated through the monitoring of the system, as shown in the top of Fig. 1.

VI. Results

In this section we present a numerical example illustrating the applicability of the proposed methodology. The considered use case is presented in Sec. VI-A, followed by a discussion of the results and findings in Sec. VI-B.

A. Use Case Description – Moose Intrusion on the Road

To illustrate the notion and important points discussed in previous sections, the authors have chosen to consider a concrete use case. We consider here an ADS with a perception system that provides free space detections at different range sections and with different integrity levels, as described in Sec. III. More concretely, the perception system is able to ensure absence of objects with a lower failure rate (i.e., false positive rate) for the range section closer to the vehicle, and a higher failure rate for farther range sections. The considered failure rates and range sections as considered in this paper are presented in TABLE I. It is important mentioning that these numbers do not correspond to a real perception system, but rather correspond to an engineering judgement of the authors, established for the purpose of this work and believed to be realistic. Indeed, it is not unrealistic to assume that the high integrity levels for shorter range section stem from relying on a purely physical models, based on e.g. LiDAR reflections. While for the longer range sections, one may leverage segmentation or object detection, requiring neural network-based approaches, for which lower integrity levels could be expected.

Range section	r_1	r_2	r_3
Perception range [m]	$[0, 20]$	$[20, 50]$	$[50, 150]$
Failure rate, $\lambda_{\text{true}}^{r_i}$ [h^{-1}]	10^{-6}	10^{-4}	10^{-3}

TABLE I: Failure rates for the different perception range sections of the considered ADS.

Furthermore, and for the sake of illustration only, we assume that the ADS operates on rural roads and is exposed to a particular adverse event, i.e., collisions with wild moose. Note that considering moose accidents in this work is not a random choice, and motivated by the fact that:

(a) there exist available statistics for such types of accidents in different datebases, such as Strada (Swedish Traffic Accident Data Acquisition)[2] and Nationella Viltolycksrådet (National council for road accidents with game)[3] (as considered here),
(b) it provides a relevant example of a traffic event, where the outcome is relatively independent of the interaction between the object and the ADS. This is for instance not the case for the interaction of vehicles and jaywalking pedestrians, where the behaviour of the pedestrian is subject to the traffic context and the behaviour of the ADS,
(c) unlike other human traffic participants, the injuries caused to the moose can, from a legal, societal, and insurance perspective, be neglected to a certain extent, which simplifies the analysis of the consequence classes of the QRN.

For the purposes for our use case, we use $\lambda_{\text{true}}^{\text{moose}} = 2 \cdot 10^{-5}$ h^{-1} as the *true* exposure to moose. This value is derived based on the following arguments:

- there are about 5500 accidents with moose in Sweden every year[3];
- there exist roughly 5.5 million vehicles in Sweden[2];
- each vehicle is assumed to operate 50 hours a year on rural roads where moose could appear.

[2]More details on the Strada database: www.transportstyrelsen.se/sv/vagtrafik/statistik/olycksstatistik/om-strada/

[3]Data can be found at *Nationella Viltolycksrådet (National council for road accidents with game)*: https://www.viltolycka.se/statistik/

There are two caveats that should be mentioned though:

i) the exposure rate to moose is likely much higher than the reported accident rate. However, for simplicity in our modelling, we consider a moose appearing to be equivalent to it staying on the road until the ADS could crash with it. Hence, slightly underestimating the exposure rate would arguably be counteracted by the fraction of cases where the moose simply run across the road without the vehicle even being close to it,

ii) the actual number of accidents may also be slightly under reported, as some accidents might never show up in authorities' records. This factor should, however, be relatively small and it is only relevant for non-fatal and mild injury accidents, as both police and hospital records are included in the Strada statistics.

Finally, regarding the ADS actuation principles, we restrict ourselves to, but without loss of generality of the overall methodology discussed earlier, consider longitudinal control only. In other words, the task of finding a safe driving policy becomes equivalent to finding a safe driving velocity, where the higher level tactical performance objective is to maximise this value.

1) QRN for the Use Case: The QRN concept, as proposed in [13], provides quantitative vehicle level requirements in the form of allowable frequencies for different incidents and associated outcomes. In the considered use case, one can characterise the incident of interest, i.e. collision with a moose, into different consequence classes depending on the impact velocity. In the interest of keeping the focus of the discussion here, we have not considered any cross transfer from velocity ranges into different consequence classes, as discussed in [13], but rather assume a deterministic relationship between impact speed and injury level, as discussed in Sec. II. Thus, a velocity range is selected, based on the authors' engineering judgement, in order to capture roughly the same injury level.

Considering the statistics of accidents with game from Strada[2], in combination with the total number of accidents with moose identified by *Nationella Viltolycksrådet (National council for road accidents with game)*[3] and the assumptions on the driven hours by each vehicle (as given above), we are then able to derive the accident numbers as presented in TABLE II. It is worth noting that one could consider adding a factor to account for an increased performance requirement for the ADS when compared to the average human performance (assumed to correspond to the numbers of TABLE II). However, without impacting the validity of the results presented in the sequel, we omit such a factor and consider that the ADS is expected to operate as good as the average human driver.

B. Results

As detailed in Section VI-A, the consequences from colliding with a moose are deduced deterministically given the estimations of the exposure rate and the perception failure rates. However, to get a realistic understanding of the updating process of these estimates, in relation to the derived safe velocity, the occurrences of each sub process are simulated by sampling from Poisson distributions with the respective rate parameters set according to the values given in Sec. VI-A.

Impact velocity [km/h]	Consequence, c_i	Accident rate [h^{-1}]
$0-10$	Damage to vehicle	$2 \cdot 10^{-5}$
$10-30$	Minor injuries	$2 \cdot 10^{-6}$
$30-50$	Moderate injuries	$2 \cdot 10^{-7}$
$50-70$	Severe injuries	$4 \cdot 10^{-8}$
$70 <$	High fatality risk	$1.5 \cdot 10^{-8}$

TABLE II: Considered QRN values for scenarios with moose. The numbers of injuries and fatalities are taken from Strada[2], whereas the damage to vehicle value ($0-10$ km/h range) are retrieved from *Nationella Viltolycksrådet*[3].

The results from these simulations are presented in Fig. 4 with respect to the estimated moose arrival rate. From Fig. 4 one can see that it takes close to 10^6 hours for the estimate to converge to the true arrival rate. In general terms, the estimates converge to the same order of magnitude when the system has operated the same number of hours as the inverted true arrival rate ($\lambda_{\text{true}}^{-1}$), but the last notches within that magnitude are only accurately estimated with an additional order of magnitude. For the moose arrival rate estimates, it only start to saturate in its convergence at 10^6 hours, which is one order above the true value of $\lambda_{\text{true}}^{\text{moose}} = 2 \cdot 10^{-5}$ h^{-1}. Note that the resulting development of the estimates are qualitatively the same for the perception failures but have been, for the sake of conciseness of the paper, excluded from the presentation.

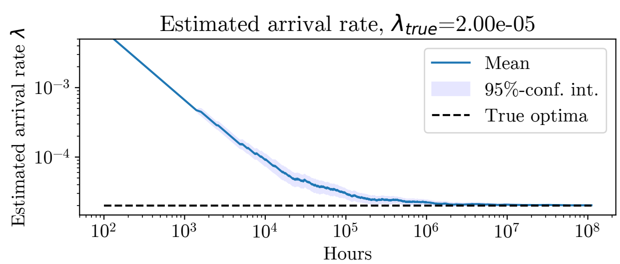

Fig. 4: Mean estimated exposure rate of the moose using 50 simulation runs and using $\hat{\lambda}$ according to Sec. IV-A. The true arrival rate $\lambda_{\text{true}}^{\text{moose}} = 2 \cdot 10^{-5}$ h^{-1} is depicted by the dashed line, and the light shaded area identifies the 95% confidence interval across the 50 simulations. Note that both axes are in log-scale.

Let us now analyse the resulting safe velocity profile, considering the updated estimates of the moose exposure rate and perception failure rates. The safe driving velocities are depicted in Fig. 5. There are four interesting things to note:

(1) Before having sufficient data, the system will not be able to safely operate at all. Indeed, a safe driving velocity profile (denoted by the blue line) is only available after roughly $2 \cdot 10^2$ hours. This is largely due to the strict

QRN requirements for the speed range $[0, 10]$ km/h, and the fact that we do not have any prior knowledge when starting the inference process,

(2) The safe driving velocities approach the true value in a similar fashion as the convergence of the estimated arrival rate of the moose, as depicted in Fig. 4,

(3) The safe velocity transitions relatively quickly from 10 to 30 km/h, and from 30 to 50 km/h. This is probably due to the combined effect of the discretisation of the prescribed QRN requirements and the marginal effect on the impact velocity of braking actions for short detection ranges (c.f. d_1 in Fig. 2).

(4) The allowable velocity does not monotonically increase. Instead, it is impacted by the oscillations of the estimated rates.

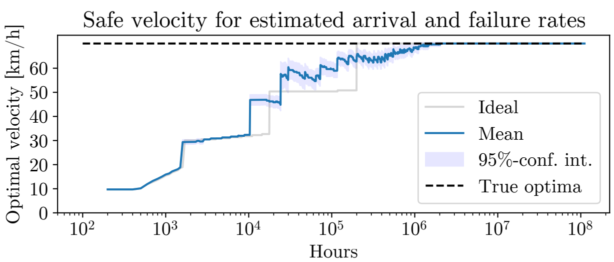

Fig. 5: Allowed safe velocity in relation to the amount of gathered data. The dashed line depicts the safe velocity derived using the true rates of the arrival processes, and the light blue area corresponds to a 95% confidence interval around the mean velocity, calculated from a set of 50 simulations. The light grey line corresponds to the ideal safe velocity given an ideal sampling from the arrival processes, i.e. without randomness. Note that the x-axis is in log-scale over the operated hours of the system, whereas the y-axis is linear and established in km/h. The safe velocity is derived using the estimated exposure rate and failure rates, considering a $\hat{\lambda}$ according to equation 2 in Sec. IV-A.

VII. Conclusions and Future Work

In this paper we present a methodology for designing a safe driving policy where uncertainties from underlying statistical models are taken into account to achieve an uncertainty aware system. We have analysed the impact of perception failures on the safe driving velocity of an ADS, and considered an example end-to-end use case focusing on collision scenarios with moose. The prescribed safety requirements on the system have been defined according to the Quantitative Risk Norm (QRN) concept [13]. The process of deriving representative values for the QRN is exemplified using real accident statistics. The process provided here is only for illustration purposes and further research on this topic is necessary. Statistical (Poisson) models for both the exposure probability of an adverse event (appearance of a moose), as well as the failure rates of the perception system, have been considered. The estimated exposure and perception failure rates are continuously updated from (simulated) field data and the safe driving velocities, fulfilling the prescribed QRN, are derived based on the updated exposure and failure rates. Using a confidence interval on the statistical models, establishing a conservative estimation, the safety of the system can be assured based on the present knowledge of the system's performance and its operational context. The resulting allowable safe velocity is shown to rely on the estimated exposure and failure rates, such that increased knowledge (i.e., increased quantity of gathered data) allows for a more performant ADS, i.e. one with a higher allowed driving velocity, while maintaining the confidence in the safety claim. The considered use case provides some first considerations on how fleet monitoring can be used to support continuous updates of an ADS. In particular, it exemplifies the use of evidence produced at run-time for design-time activities. The shown results also stress the importance of incorporating field data into the design and development process of an ADS. It is worth mentioning that while the proposed methodology in this paper only considers design-time updates to an ADS, i.e. where all ADSs are implicitly assumed to drive with the same policy until the next update, the methodology could be extended to also consider run-time aspects. For example, the proposed methodology could support dynamic risk assessment [18], or be used to perform safe risk-aware, tactical decisions, as discussed in [24]. In such a context, the considered statistical models here could be extended so to include the probabilities conditioned on different operational conditions. While not investigated in-depth in this paper, this interesting topic is left for future work.

The considered use case is limited to a *discrete* adverse event, where a moose intrudes the road. Unlike vehicle behaviour, such as cut-ins or overtakes [25], for which crash situations could be extrapolated from a set of near crash scenarios, the considered *discrete* event requires actual measurements of it happening, e.g., a sighting of a moose. As a future work, the authors propose to investigate the incorporation of data from events where it is also possible to extrapolate from near crashes. For example, one could model a threat measure using extreme value theory [26] and connect that metric to the impact on the performance and safety considerations of the ADS, following the methodology outlined in this paper.

References

[1] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" Transportation Research Part A: Policy and Practice 94, pp. 182–193, 2016.

[2] R. A. Young, "Automated driving system safety: Miles for 95% confidence in "vision zero"," *SAE Int. Journal of Advances and Current Practices in Mobility*, vol. 2, no. 2020-01-1205, 2020.

[3] M. Törngren and P. T. Grogan, "How to deal with the complexity of future cyber-physical systems?" *Designs*, vol. 2, no. 4, 2018.

[4] 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.

[5] P. Koopman, B. Osyk, and J. Weast, "Autonomous vehicles meet the physical world: RSS, variability, uncertainty, and proving safety," in *Int. Conference on Computer Safety, Reliability, and Security*. Springer, 2019.

[6] A. Shetty, M. Yu, A. Kurzhanskiy, O. Grembek, H. Tavafoghi, and P. Varaiya, "Safety challenges for autonomous vehicles in the absence of connectivity," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103133, 2021.

[7] A. Pütz, A. Zlocki, J. Bock, and L. Eckstein, "System validation of highly automated vehicles with a database of relevant traffic scenarios," in *ITS European Congress*, 2017.

[8] G. Bagschik, T. Menzel, and M. Maurer, "Ontology based scene creation for the development of automated vehicles," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[9] J. Bach, S. Otten, and E. Sax, "Model based scenario specification for development and test of automated driving functions," in *IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[10] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 595–607, 2016.

[11] 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 *European Congress on Embedded Real Time Systems (ERTS)*, 2020.

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

[13] 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 Int. Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2020.

[14] R. Johansson and J. Nilsson, "The need for an environment perception block to address all asil levels simultaneously," in *IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[15] M. Woods et al., "Safety first for automated driving, White paper," 2019. [Online]. Available: www.connectedautomateddriving.eu/wp-content/uploads/2019/09/Safety_First_for_Automated_Driving.pdf

[16] M. Lindman, I. Isaksson-Hellman, and J. Strandroth, "Basic numbers needed to understand the traffic safety effect of automated cars," in *IRCOBI Conference*, 2017.

[17] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] J. Reich and M. Trapp, "SINADRA: Towards a framework for assurable situation-aware dynamic risk assessment of autonomous vehicles," in *European Dependable Computing Conference (EDCC)*. IEEE, 2020.

[19] M. Gyllenhammar and C. Zandén, "Performance monitoring and evaluation of a vehicle adas or autonomous driving feature," Apr. 22 2021, US Patent App. 17/075,181.

[20] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan, "Streaming variational bayes," in *NIPS*, 2013.

[21] M. Gyllenhammar, C. Zandén, M. K. Vakilzadeh, and A. Falkovén, "Methods and systems for automated driving system monitoring and management," Oct. 21 2021, US Patent App. 17/232,331.

[22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman and Hall/CRC, 2004.

[23] G. Rodrigues de Campos, R. Kianfar, and M. Brännström, "Precautionary safety for autonomous driving systems: Adapting driving policies to satisfy quantitative risk norms," in *Int. IEEE Conference on Intelligent Transportation Systems*, 2021.

[24] M. Trapp, D. Schneider, and G. Weiss, "Towards safety-awareness and dynamic safety management," in *European Dependable Computing Conference (EDCC)*. IEEE, 2018.

[25] M. Gyllenhammar, C. Zandén, and M. Törngren, "Defining fundamental vehicle actions for the development of automated driving systems," SAE Technical Paper, Tech. Rep., 2020.

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