Modelling road hazards and the effect on AV safety of hazardous failures

Cornelius Buerkle[1], Fabian Oboril[1], Michael Paulitsch[1], Peter Popov[2] and Lorenzo Strigini[2]

***Abstract*— Autonomous vehicles (AV) are about to appear on our roads within the next few years. However, to achieve the final breakthrough, not only functional progress is required, but also fundamental safety questions must be solved. Among those, a question demanding special attention is the need to assess the overall safety of an AV and quantify that it is safe enough to take part in normal traffic despite its inherent imperfections. Therefore, this paper describes a probabilistic model, which allows to study how imperfections of an AV perception system and of mechanisms responsible for AV safety (e.g., Safety Monitors), can impact AV safety in the presence of road hazards. We also demonstrate how the model can be used to validate if the AV is safe enough, to understand the criticality of (perception) errors, and to identify areas/parameters that have more influence on safety than others.**

I. Introduction

Assessing safety of autonomous vehicles (AV) poses new challenges. The known approaches to AV safety assessment, e.g., functional safety [1], have been tried in the past, but dealing with components based on machine learning remains problematic. Consequently, new alternatives are evolving, among them is the standard on "Safety of the Intended Functionality (SOTIF)" [2] (ISO 21448), or the upcoming standard on safety of AV decision making components [3]. In addition, governmental authorities are currently setting up the required legislative boundaries for the public use of AVs. For example, the German government announced that an AV has to operate at least as safely as a human driver to receive certification [4]. However, while these standards define guardrails in which manufacturers must operate, an important question remains: *How to prove that an AV is safe enough to receive certification to participate in regular traffic?*

To demonstrate that an AV is safe enough, manufacturers need to validate that the rate of catastrophic failures (i.e., accident) is below the required threshold. Targeting the performance of a human driver, this means that the accident rate must be below 10^{-5} (1 severe accident every 10^5 hours of driving). One solution to this problem is to derive safety arguments from the large number of miles driven by AVs [5]. This, however, seems impractical for future mass deployment with frequent software updates. Instead, formal approaches are required, that can be either mathematically verified or validated on a smaller amount of data.

As illustrated in Figure 1, a catastrophic failure (i.e., an accident caused by the AV) is the consequence of two events that happen "simultaneously". First, a road hazard is required (dangerous driving situation), and second, a system failure (e.g., a failure of the AV's perception system to detect a leading vehicle) has to occur as well. Thus, it is important that formal safety assessment approaches consider the situational awareness together with the failure rates of the AV subsystems (e.g. perception, planning, safety monitors, etc.).

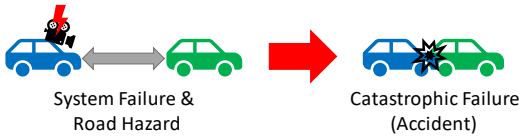

Figure 1 Illustration of road hazards.

In this regard, it is worth noting that for example even the best state-of-the-art perception systems have only a modest accuracy. As indicated by some, e.g., [6], detection accuracy of objects (lights, pedestrian, etc.) by a good machine-learning (ML)-based perception device is in the range of 80% - 99% per object. Although these numbers can be improved by combining several perception approaches and using multi-modal sensor technologies, one question remains [1]: *What is the required failure rate of a perception system, to achieve the necessary level of vehicle safety* (given the situational-awareness mentioned before)?

In this work we address the key questions raised before by a formal approach to safety assessment. We propose a comprehensive model to estimate the probability of a catastrophic failure of an AV over a period of time (thousands of hours of driving), that connects road hazards to the aforementioned failures in the AV processing pipeline (e.g., perception or planning). For this purpose, we model the failures and their occurrences as non-homogeneous stochastic processes. As we demonstrate, the model can be fed with naturalistic driving data on the distribution of road hazards (their rate of occurrence and durations as recommended by ISO 26262) and yield important insight on acceptable failure rates of AV components. Our evaluations reveal that perception stack and safety monitors must meet stringent reliability requirements. Yet it is not required that their failure rates are on par or below the target catastrophic failure rate, as non-hazardous traffic situations act as masking factors. We also report that some of the parameters used in our analysis have negligible impact on system safety, thus they need not be estimated with high confidence. In summary, our proposed model captures SOTIF performance limitations with situational awareness and their impact at system level (catastrophic failures), which can be used early in the design process to make appropriate design decisions.

In the rest of this paper, Section II defines essential concepts used in our proposed model. The model itself is presented in Section III. Section IV covers our evaluations, followed by a discussion in Section V and conclusions in Section VII.

II. Preliminaries

In our formal model, failures are seen as "random" with occurrence rates likely to vary with the variation of road

[1] Intel Labs, [2] City University London

conditions, i.e., their occurrence can be modelled as a non-homogeneous stochastic process. In this paper we discriminate between two states of the operational environment on the road:

- "normal conditions": an AV is operating in the presence of *no road hazards*. In this mode of operation, failures are very unlikely to occur, a view which is clearly extreme. A more realistic view would be to assume that catastrophic failures occur in "normal conditions", too, with lower intensity.
- "hazardous conditions": the AV is faced with hazardous road conditions, e.g. high traffic density (i.e. close proximity of surrounding vehicles), poor weather conditions or rare traffic situations limiting the space for maneuvering on the road, etc. In these circumstances, the likelihood of a catastrophic AV failure is different, typically significantly higher, than the (very low) likelihood of catastrophic failure in "normal conditions".

A *hazard* on the road is a "temporary" state of the environment [1]. Once it occurs, it will have a *duration* (e.g., a few seconds, possibly longer) after which it will cease to exist. Hazard duration is modelled as a random variable with a specified probability distribution. The hazard duration, of course, is not affected by whether the hazard is recognized correctly by the perception system or not.

A *catastrophic failure* may or may not occur while the vehicle is in "hazardous condition" or indeed in "normal condition". When a hazard occurs, it seems plausible to assume that the likelihood of catastrophic failure is no better than the likelihood in "normal conditions". In this work, we interpret "catastrophic failures" as "severe accidents caused by the AV".

An AV should be able to detect a road hazard and adapt its control accordingly (e.g., if an obstacle is seen on the road, then the AV speed may need to be reduced [13, 14]. Hazard recognition is part of the AV perception system and as such may be subject to failures. We consider the following possibilities for the hazard perception system:

- Correctly perceived hazard (CPH), given the hazard has occurred.
- Overlooked hazard (OLH) – failure by the AV perception system to recognize a road hazard. In this situation the AV will continue to operate as if the hazard did not exist. Clearly, in this situation, the vehicle becomes more likely to fail catastrophically than if the hazard has been correctly recognized and the control duly adapted to the hazard (e.g., to either stop the AV or undertake a suitable maneuver). This event is also often called a "false negative" or "perception miss" in the literature.
- Falsely perceived hazard (FH) (or false alarm, "false positive") occurs when the AV perception system incorrectly perceives the situation as hazardous. This, in turn, may trigger actions, which are strictly not necessary (e.g., reduce the AV speed). The response to FH may not affect safety at all, or may indirectly improve it (for instance, reducing velocity due to a false alarm may actually make catastrophic failures less likely than in "normal conditions"). But also, could lead to a hazard, as, e.g., sudden unnecessary braking may surprise the vehicles behind (see also case 4 in IV.B).

We explore different situations about the relationship between the likelihood (rate) of catastrophic failure in the different situation (conditions) listed above:

- The likelihood (rate) of a catastrophic failure in "normal conditions" may be set to 0, i.e., the AV cannot possibly fail catastrophically in "normal conditions". Clearly, this is an extreme scenario, which rules out a failure due to equipment failure or due to rare circumstances where the rules used by AV *safety monitors* turn out to be insufficient to avoid an accident [7]. However, the scenario can be seen as a "best case" scenario as in reality AV safety will be no better.
- Rate of catastrophic failure in CPH is greater (or at least no lower) than the rate of failure in "normal conditions". In the absence of empirical data to suggest otherwise, this assumption seems quite plausible, since even though the AV can maneuver to evade escalation of hazards to accidents, this does not nullify the added risk. E.g., braking hard to avoid a vehicle intruding in one's lane may avoid collision but is still a more dangerous condition than if the intrusion (hazard) had not occurred.
- Rate of catastrophic failure in OLH should be greater than both the rate of failure in "normal conditions" and the rate of failure in CPH.

III. The Model

Our envisioned stochastic model to assess AV safety and estimate the rate of catastrophic failures, based on the formalism of stochastic activity networks (SAN) [7] is shown in Figure 2.

This model can be in one of the following 5 states (modelled as places in Figure 2):

- OK, which models "normal conditions"
- FalselyPerceivedHazard, models "normal condition" perceived incorrectly by the AV perception system as hazardous ("false alarms" state).
- CorrectlyPerceivedHazard (CPH), a "hazardous condition" perceived correctly as hazardous by the AV perception system.
- OLH (OverlookedHazard) is a "hazardous condition", which the AV perception system failed to classify as hazardous. The AV is unaware that it is in a "hazardous situation".
- CPH_Late, a hazardous condition is eventually correctly perceived as hazardous, but only after overlooking it for a while, i.e., after AV has spent some time in the OLH state. This state is usually different

from CPH, as the vehicle is already in an unsafe state and may not avoid an accident even if it applies countermeasures (e.g., braking).

- Accident – this is an "absorbing state", which models the occurrence of a catastrophic failure. If the system reaches this state, further changes of the state in the model are impossible.

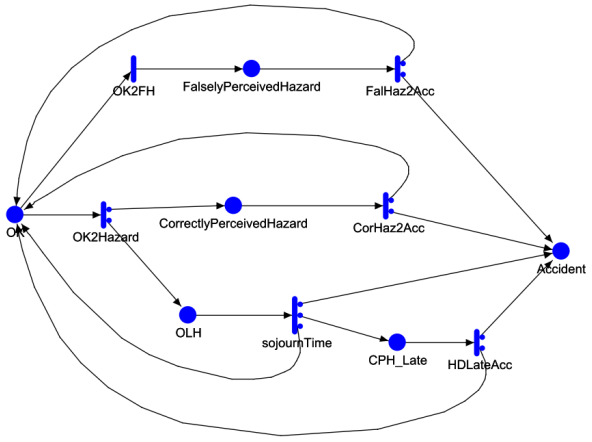

Figure 2. The model of an AV with road hazards and imperfect hazard perception system: hazards may or may not be recognised correctly.

Transitions between these states are governed by timed "activities", which are parameterized accordingly. Each timed activity will have a parameter characterizing its duration and, optionally, 2 or more transition probabilities which define the new state in which the model will be at the end of the activity. For instance, activity OK2Hazard on Figure 2. is parameterised as shown in Table 1 below. Its duration is assumed an exponentially distributed random variable with a parameter HazardRate. The "case distribution" defines the probabilities associated with the two "cases" attached to the timed activity, which in turn define the probabilities of transitions to either the state CorrectlyPerceivedHazard or to the state OLH upon completion of the timed activity duration. The two case probabilities define a distribution (their sum should equal 1) and in this example rely on another parameter, missHazardProb, a constant which allow the modeler to define the probability of perception to overlook a hazard, should the hazard occur.

The "global" variables HazardRate and missHazardProb can be assigned different values to conduct "sensitive analysis", i.e., how sensitive AV safety is to the particular parameter(s).

Table 1. Parameters of timed activity ***OK2Hazard***

Distribution type	Exponential
Duration	return(HazardRate);
Case Distribution	case 1: return(1.0 - missHazardProb); case 2: return(missHazardProb);

All transitions in Figure 2. are defined in a similar manner; their parameters are discussed next. It is clear from the model that a hazard can either escalate to an accident or not. The transition from a hazardous state – either correctly recognized or not – to an accident is governed by the parameters of the respective timed activity leading to the state Accident.

A failure of the perception system to perceive a hazard is modelled by a transition to state "OLH", from which a transition to "Accident" may occur after a delay captured by the duration of the timed activity ***sojournTime*** and with probability defined in the case distribution of this timed activity. Another possible transition from OLH is to state CPH_Late with probability defined for case 2 of ***sojournTime***. Finally, the overlooked hazard may simply disappear without any visible consequences and the model may return to OK state with probability defined for case 3 of ***sojournTime.*** Should the model enter state CPH_Late, it will stay in this state for the duration of the timed activity ***HDLateAcc***, from which the model can move to either state Accident of to state OK. The transitions take place according to the case distribution defined for ***HDLateAcc***.

Similarly, a transition to either state "Accident" or state OK from state "CorrectlyPercievedHazard" according to case distribution defined for timed activity ***CorHaz2Acc*** following a delay defined by the Duration of ***CorHaz2Acc***. It may be worth noting that safety concept for AV planning, like RSS [8], aim to ensure that, if implemented in all vehicles, a transition from "CorrectlyPerceivedHazard" to "Accident" is impossible. In practice, this ideal condition may not always hold as described in [15]. Therefore, we have presented model results for both the idealized case (the transition probability is 1 to state OK and 0 to "Accident") and for non-zero transition rates to "Accident" (cases 4 and 6 in Section IV.B) .

Finally, in this model the possibility of falsely perceived hazards is modelled via the timed activity ***OK2FH*** leading to FalslyPercievedHazard state. This state triggers the time activity ***FalHaz2Acc***, which models the time the AV may remain in the state before it moves to either state Acc or to OK according to the case distribution of ***FalHaz2Acc***.

The model is built on a set of assumptions about the stochastic relationships between the various random variables used in the model: i) the duration of ***OK2Hazard*** is assumed *stochastically longer* (i.e., more likely to exceed a threshold, for any threshold duration one wishes to consider) than the duration of ***sojournTime***. The rationale for this is that ***sojournTime*** represents only a part of the time the model is in hazard – the time the hazard is overlooked – followed by the time the hazard is correctly recognized. If the hazards – recognized from their occurrence and those which are initially overlooked but later recognized – are stochastically similar, this assumption is quite plausible. We acknowledge, however, that the hazards which are initially overlooked may be a "different" category of hazards, which apart from being more difficult for the perception system to recognize may have different *temporal characteristics*, i.e., the distribution of their duration may be very different from the distribution of the duration of the other, "easy to detect" hazards. In this paper we do not explore this possibility.

The initial marking of the places (states) assigns a token to OK state and 0 tokens to all other states. Thus, the model always starts in a state without hazards in the environment, a modelling choice which seems quite plausible. All timed activities used in the model are assumed exponentially[2] distributed with parameters as defined below.

The model is solved to compute the probability distribution of the TimeToAccident, i.e., the time until the model enters the absorbing state Accident. This distribution is captured via the values the cumulative distribution function (cdf) takes on a set of predefined points of time: 100 hours, 1100 hours, 2100 hours, …, 9100 hours of operation. The largest value, 9100, represents a "mission of observation" longer than a year.

IV. Evaluation

In this section we address two concerns: i) value estimation for the parameters used in the model shown in Figure 2. and demonstrate that some of them are estimable using naturalistic driving datasets; ii) for those model parameters that cannot be estimated using existing data sets, we applied sensitivity analysis, allowing the parameter values to vary to establish ranges of parameter values, which will lead to acceptable AV safety.

A. Parameter Estimation

As described earlier, some model parameters are related to traffic conditions, some others to the quality of the perception system, or the safety monitors used in the vehicle. Even if suitable datasets were available for all parameters, it seems that applying sensitivity analysis is still useful, e.g., to establish how robust the AV safety assessment results are with respect to the different values used in the model. Parameters that affect AV safety significantly should be subjected to further scrutiny to ensure that despite some fluctuations of model parameters, system safety will remain within reasonable bounds.

To estimate the parameter values, we used two different data sources. HighD dataset [9] was used to obtain the parameter values related to road hazards (duration intervals between adjacent hazards). We also used the classical perception dataset Lyft [16] to obtain statistical information related to perception errors: duration of perception errors.

The HighD dataset is a drone-recorded dataset for naturalistic human driving behavior on German highways, containing more than 100 hours of driving data. The HighD dataset covers various vehicle speeds from almost standing still to more than 200km/h. To get coherent driving situations, we restricted the dataset to the speed range of 100km/h - 130km/h, for all evaluations in this paper, as this was the speed range for which most datapoints are available and provided reasonable vehicle-to-vehicle road hazards (no trucks, no pedestrians, etc. were involved). Of course, the model can be fed with additional speed ranges as well. For simplicity this is omitted for the scope of this paper.

To obtain the distribution duration of hazardous situations and the distribution of intervals between them, we assume that an AV will behave similarly to human drivers, and thus will be exposed to similar situations, with similar probability distributions. Further, we assume that a hazardous situation can be expressed through a time-to-collision (TTC) of less than 5 seconds, considering that the front vehicle may decelerate with its current deceleration or – 2m/s², and that the rear vehicle may accelerate with 2m/s². Based on these assumptions, we parsed the restricted HighD dataset to obtain all hazardous situations from the dataset, and the distributions used in the model: i) of the intervals between adjacent hazards, and ii) the of durations of the individual hazards. An excerpt of the resulting data is shown in Table 2.

Table 2. Excerpt from the data derived from HighD

Duration [frames]	Start frame id	Interval between Hazards [frames]
195	1053	2501
73	3749	837
18	4659	1667
261	6344	36
…	…	…

The first column represents the duration of the hazardous situation, measured in frames. The cameras used to collect HighD uses scanning frequency of 25 frames per sec, which allows for an easy transformation of data shown in Table 2 into seconds (and hours, as required in the model). The full dataset with hazards contains 16,000+ records, which allowed us to check how well the dataset fits some of the popular analytic probability distribution. The results of fitting can be seen in Figure 3. Visual inspection reveals that neither of the distribution curves fits very well (a formal goodness-of-fit test would fail for all candidate distributions). The fitness, however, is not too bad and we use the parameters from fitting an exponential distribution to parameterize the duration of the timed activity ***CorHaz2Acc***. After obvious transformations (from frames to hours) the parameter value of the exponential distribution timed activity was estimated ~856.3 hour^{-1}.

With the rationale summarized above that the duration of the ***HDLateAcc*** timed activity under plausible assumptions is likely to be shorter than the duration of ***CorHaz2Acc***, and thus the parameter of the exponential distribution associated with ***HDLateAcc*** was set to 1000, e.g., the mean duration of this timed activity is 1/1000 hours = 10^{-3} hours against 1/856.3 = 1.16×10^{-3} hours for the mean duration of ***CorHaz2Acc***.

Using the second column of Table 2 we can compute the intervals between adjacent hazards. These are shown in the 3rd column of Table 2 and are used to estimate the duration parameter of the timed activity ***OK2Hazard*** which is assumed

[2] The choice of distribution is motivated by *convenience*. Assuming exponentially distributed activity durations allows us to use numeric solvers for CTMC (continuous-time Markov chains), which are fast and give exact solutions. Using other distribution types would rule out the use of these solvers: instead, Monte Carlo simulation will be the only option as a solver. Monte Carlo simulation with high accuracy of the solutions will require very long simulation campaigns (days, vs. seconds with the numeric solvers).

to be exponentially distributed. After obvious transformations (from frames to hours), this parameter was estimated as 197.4 hour^{-1}, i.e. the average interval between hazards is 1/197.4 ~0.005 hours, i.e., about 18 sec.

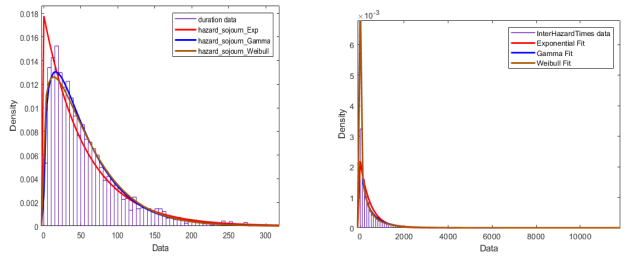

Figure 3. Fitting the dataset with hazard duration (plot on the left) and intervals between hazards (plot on the right), derived from HighD using several standard analytic probability distributions. The histograms represents the raw data.from Table 2.

The model parameter related to the length of perception failures (i.e., how long a hazard may be overlooked) were estimated using the Lyft dataset, which is a public perception dataset. We used this dataset together with a standard object detector for LiDAR sensors, PointPillars [7] is an AI-based 3D object detection solution, which we trained on a subset of the Lyft dataset. We used the remaining part of the dataset to measure the number of perception errors (i.e., not detected objects), and the duration of these perception errors. The two model parameters thus estimated were: i) the case probability leading to state OLH of the timed activity ***OK2Hazard***, which captures the conditional probability of hazard omission given a hazard occurred, and ii) the duration of time activity ***sojournTime***, which, as the name suggests represents the distribution of how long a hazard remains overlooked. The values of the first parameter was subjected to "sensitivity analysis" in the range [0, 0.01] with intermediate various of 10^{-4}, 5×10^{-4}, 0.001 and 0.002, which represent a significantly better detection rates than the empirically estimated value of 0.01 reported in [6]. The mean duration of the overlooked hazard was estimated to lay in the range of a handful of frames up to ~50 frames. We assumed that the length is exponentially distributed and applied sensitivity analysis to the parameter of this distribution assuming a mean duration of [20, 40 and 80] frames. After applying the necessary transformations (from frames to hours as required in the model) the parameter of the assumed exponentially distributed duration of the ***sojournTime*** timed activity was estimated to be 4500, 2250 and 1125, respectively. These values were used in our studies.

The other parameters used in the model are defined in Table 3[3]. A number of parameters are set to 0 in the first 3 cases reported in section B (cases 1 - 3): the case probabilities CH2Acc_prob FH2Acc_prob of timed activities ***FalHaz2Acc*** and ***CorHaz2Acc***, respectively, both leading to Accident. falseHazardProb defines the probability of detecting incorrectly a false hazard and for cases 1 – 3 was assumed equal to 0. Later, in cases 4 - 6 these assumptions are relaxed.

Table 3. Excerpt from the data derived from HighD

Parameter	Values	Description
CH2Acc_prob	0	Case probability 2 of ***FalHaz2Acc*** timed activity.
CorHaz_sojournTime FH_sojournTime	856.3	Rate of timed activities ***CorHaz2Acc*** and ***FalHaz2Acc***
FH2Acc_prob	0	Case probability of ***CorHaz2Acc*** timed activity.
HDLateAcc_prob	[10^{-4}, 2×10^{-4}]	Case probability of timed activity ***HDLateAcc*** leading to Accident.
HDLate_sojournTime	1000.0	Rate of timed activity ***HDLateAcc***.
HazardRate	197.4	Rate of timed activity OK2Hazard (intervals between hazards).
OH_sojournTime	[1125, 2250, 4500]	Rate of timed activity ***sojournTime*** (duration of overlooked hazard)
OLH2Acc_prob	[10^{-5}, 2×10^{-5}, 5×10^{-5}]	Case probability of timed activity ***sojournTime*** leading to Accident
OLH2CHLate_prob	[0.99, 0.991, 0.995]	Case probability of timed activity ***sojournTime*** leading to CPH Late.
falseHazardProb	0.0	Probability of false hazard detection.
missHazardProb	[0, 10^{-4}, 5×10^{-4}]	Probability of omitting a hazard given hazard occurred

B. Results

Below we provide some evaluations of AV safety for different model parameterizations. The first 3 cases illustrate the model's behavior under the parameterization shown in Table 3 whereby: i) accidents are assumed impossible if a hazard is recognized correctly immediately upon its occurrence, and ii) no false detection of hazards occurs. In other words, these three cases are limited to a single failure mode – failure of the perception system to recognize a hazard correctly immediately upon its occurrence.

Case 1: Probability of overlooking a hazard varies

With this case we demonstrate the impact of the probability of overlooking hazards on system safety, which is illustrated for different values of this probability in Figure 4.

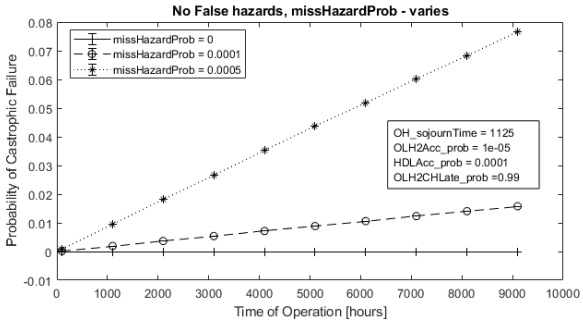

Figure 4. Effect of probability of overlooked hazards on system safety.

The effect is quite significant. Perfect perception system (i.e., when the probability of overlooking a hazard is 0) would

[3] The full SAN model is available from the authors on request.

make the system "perfect" – the probability of catastrophic failure will remain 0 for any number of hours driven. Even tiny values of the probability of overlooking a hazard (10^{-4} and 5×10^{-4}), however, lead to a visible deterioration of safety. The probability of AV experiencing a catastrophic failure for 9,100 hours of driving (more than 1 calendar year and more than 1,000,000 km, at the chosen speed of 100+ km/h,) will increase the probability of accident to almost 2% and 8%, respectively. Note that the chosen values of 10^{-4} and 5×10^{-4} are two orders of magnitude better than the estimated value of 0.01 suggested by [6]. Using 0.001 would lead to a probability of an accident at the end of 9,100 hours in excess of 0.3.

Case 2: Duration of overlooked hazard varies

Now we demonstrate the effect of the duration for which a hazard is overlooked on AV safety: we vary this duration as described above using for it an exponentially distributed random variable with a parameter 1125, 2250 and 4500, which represents mean duration of 3.2, 1 and 0.8 seconds, respectively. Note that all these values represent very short periods for which the hazard is overlooked.

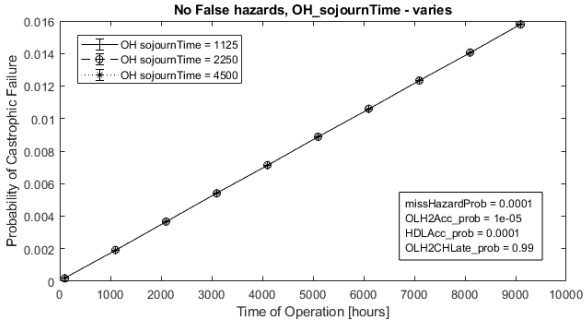

Figure 5. Effect of length of hazard on system safety.

It is clear from Figure 5. that the effect of the chosen variation is negligible: the three curves overlap and are indistinguishable in the plot. We checked also how perception delays that are *longer by an order of magnitude* will impact safety and did not register any visible impact.

Case 3: Effect of the probability of accident following an overlooked hazard

The results from this study are illustrated in Figure 6.

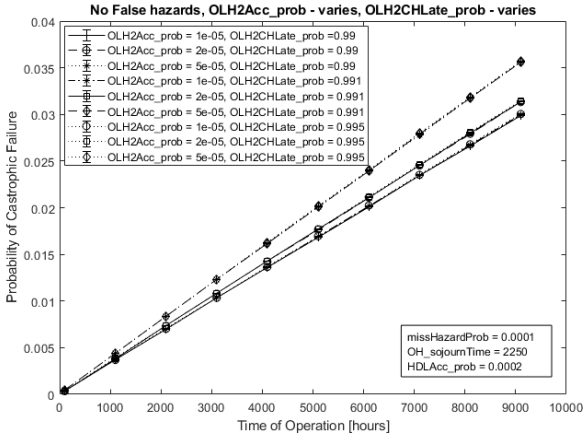

Figure 6. Effect of length of hazard on system safety.

It seems clear that this parameter (one of the case probabilities associated with the timed activity ***HDLateAcc*** in Figure 2.) has a visible impact on AV safety: even a small increase of the probability OLH2Acc_prob from 10^{-5} to 5×10^{-5} - impacts noticeably AV safety. In absolute terms, however, the impact is quite limited: after 9,100 hours of operation, it remains within the range just below 3% to 3.5%.

The next 3 cases illustrate the impact on safety of the parameters that so far have been set to 0. These are CH2Acc_prob, which defines the probability of accident given a hazard has been recognized correctly upon its occurrence, falseHazardProb, the parameter dealing with falsely perceived hazards, and FH2Acc_prob, the parameter which specifies the probability of an accident given a false hazard. The values of these parameters used in Case 4 and 6 are defined in Table 4. Case 4 illustrates the impact on system safety of failures of the *safety monitors* only, i.e., failures which may occur even when the perception system does recognize hazards immediately upon their occurrence. Case 5 demonstrates the effect of "false alarms" only raised by the perception system (i.e., recognizing non-existing hazards). Finally, Case 6 illustrates the combined effect of false alarms and imperfection of safety monitors.

Table 4. Values of CH2Acc_prob and falseHazardProb parameters used in Case 4 – Case 6. The values of the other parameters used in these 3 cases are as in Table 3 above.

Parameter	Values
CH2Acc_prob	$[0, 10^{-6}, 2\times10^{-6}]$
falseHazardProb	$[0, 10^{-4}, 5\times10^{-4}]$
FH2Acc_prob	$[0, 10^{-6}, 10^{-5}]$

Case 4: Effect on AV safety of imperfect safety monitors.

It is clear from Figure 7. that imperfection of safety monitors has a very significant impact on AV safety. Even very small values of the probability of imperfect response (of order of 10^{-6}) lead to a visible increase of the probability of accident: for 9100 hours this probability reaches values greater than 0.8 and almost 1.0, suggesting that an accident is very likely to imminent within 9100 hours. Even for much shorter periods (e.g., 2000 – 3000 hours) the probability of an accident reached values of 0.4 – 0.5. These values of the probability of accident are significantly greater compared with the cases assuming that safety monitor is perfect (CH2Acc_prob = 0). This observation makes it very clear that safety monitors must be very reliable (better that 10^{-6}).

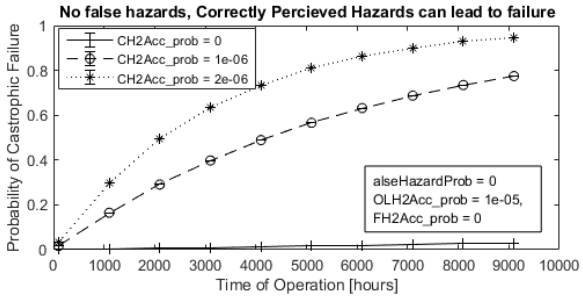

Figure 7. Effect of monitors imperfection on system safety.

Case 5: Effect of false alarms on system safety

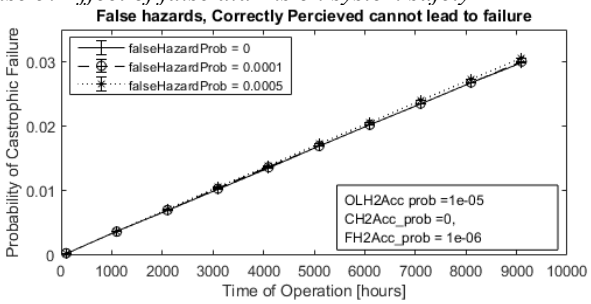

Figure 8. Effect of false alarms on system safety.

This case looks at the impact of falsely perceived hazards (a new failure mode of the perception system) on AV safety. We outlined earlier how this may affect safety. In the studies under this case, we assume that false alarms are hazardous and in rare cases may escalate to an accident. The plots on Figure 8. indicate that such additional hazards may reduce system safety (the probability of accident increases), but the reduction is barely noticeable for the chosen parameters. We note that similar lack of impact applies also to a much larger values of the probability of false alarm of up to 0.1% and 1% and conditional probability of accident given false hazard of 10^{-6}. In all cases that we have studied, the AV safety is practically indistinguishable from the safety estimated under the assumption that false alarms are not hazardous.

Case 6: Combined effect of false alarms and imperfect safety monitors.

While in case 4 and 5 we assumed a single failure mode (either imperfect safety monitor or false alarms in hazard perception) in this case we studied the combined effect of false hazards and imperfect safety monitors. The plot in Figure 9. is similar to Figure 7. – imperfections of safety monitors significantly affect AV safety, while the impact of false alarms is barely visible – we now include in the plots in Figure 9. all values of FH2Acc_prob listed in Table 4. In other words, the effect of safety monitor imperfection on AV safety is much stronger than that of false hazards.

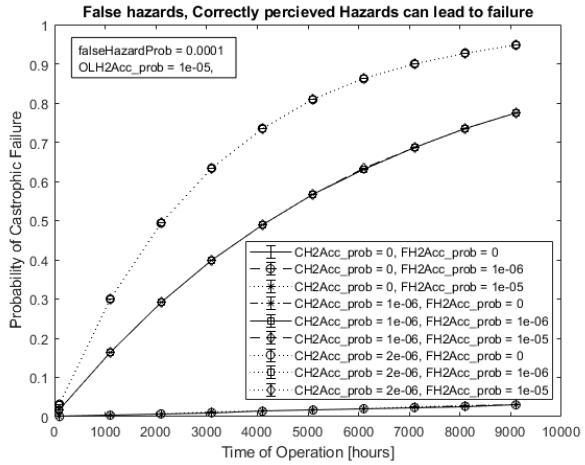

Figure 9. Combined effect of false alarms and imperfect safety monitors on system safety.

V. Discussion

We have proposed a model-based approach to safety assessment of AV, based on probabilistic modelling. The model allows one to spell out the phenomena which affect AV safety and then explore in quantitative fashion the impact of model parameters on AV safety.

We demonstrated that some of the model parameters can be estimated using public datasets. For some other model parameters, no suitable datasets existed, and we had to apply sensitivity analysis varying their values within plausible ranges. It turned out that, with these ranges of parameter values, some of the model parameters had little to no impact on AV safety, e.g., the delay in perceiving initially overlooked hazards, and duration of false hazards. Other parameters, however, e.g., the probability of overlooking a hazard, given that the hazard exists, affect AV safety very significantly. The results obtained with the model suggest that even with a very good perception system (probability of overlooking a hazard of 0.01%), the AV safety is modest – the probability of an accident of over a year (9100 hours non-interrupted driving) is non-negligible. While 9100 hours of driving at 100 – 130 km/h means over 1 million km of driving, these figures still compare unfavorably with human drivers' accident rates (of the order of hundreds of millions kilometers between fatalities and several millions between injuries [10]). If this safety level is unacceptable, developers need to improve some of the parameters under their control, by improving: i) the AV perception, which affects the critical model parameter, the probability of overlooking a hazard, which we already assumed very good; and/or ii) the safety mechanisms, which affect the probabilities of transition to Accident state.

The role of the sensitivity analysis in the study is noteworthy. It allows one to identify those parameters that affect AV safety significantly and concentrate on estimating them accurately or even conservatively. Parameters that have little/no impact on model behavior (e.g., in our case, the delay in perceiving hazards that are not identified immediately) require less estimation precision. Demonstrating that these parameters lie within a "ballpark" range for which models show adequate safety will be sufficient, saving the effort and resources that would be needed for more accurate estimates.

Another aspect worth mentioning is that the values of the various parameters are usually not constant over time. However, in this work we assumed that the distributions of the various intervals captured in the model by timed activities, do not change over time. This is clearly a simplification: the distributions typically will vary over time. The impact of this variation will pose no technical problems – the theory of non-homogeneous Markov processes is well developed and should be easy to apply in our proposed model.

A related problem concerns variation of driving conditions in operation. Operational Design Domain (ODD) captures the idea well but focuses on ways of capturing the differences between operation conditions ("modes of the environment") within a given ODD. The dynamic aspects of how these operating conditions manifest themselves in operation, e.g.,

the rate of change of these operating conditions is paid little attention but can be easily integrated by a) using the dynamic approach explained above and b) sub-models capturing different parameter values for different ODDs.

Finally, the influence of hardware errors vs. software errors should be noted. The hardware failure rate impact on perception is low as the hardware impact is typically managed by hardware development and production process, which are captured in safety standards like ISO26262 [1]. Additionally, mechanisms for modern perception machine learning architectures are available that can limit the hardware impact on system level perception failures (like false positives and false negatives) like activation range supervision [11], which may lead to hazards. Also, typically machine learning models are quite robust to hardware failures [12]. As a result, hardware errors, which are to be expected in the ballpark of 10^{-7} to 10^{-8} are less important than all other error sources (e.g., mal-trained perception network).

VI. Conclusion

Assessing and proving safety of an AV quantitatively is an open research challenge. The trends in standardization and legal regulation highlight a goal that the probability of an AV causing an accident needs to be significantly lower compared to a human driver. Therefore, it is of uttermost importance to be able to quantitatively estimate the probability of a catastrophic failure.

The challenge is that this failure probability is impacted by different factors. On the one hand, there are imperfections in the AV processing stack, e.g., errors in the perception or planning system, that can cause an accident. On the other hand, there are hazards imposed by the environment in which the vehicle operates. In addition, the two factors might influence each other, as e.g., a hazardous environment might be especially challenging for the AV perception system.

In this work we proposed a formal model that allows to estimate the final system failure probability (i.e., of catastrophic failures) based on the aforementioned influential factors. In other words, our model quantifies SOTIF performance limitations of the AV with situational awareness. We acknowledge that it is still an open question how the required parameters can be estimated in a reliable manner. Nevertheless, we showed that initial estimates are possible for some of the parameters from publicly available data. We are confident that some other parameters can be accurately estimated via tailored data acquisition, e.g. using fleet information of OEMs. At that point, the model[s] will help to show whether any parameters remain hard to estimate empirically because they relate to very rare, or hard to observe, events.

Besides the actual estimation of the failure rate, we also showed the usefulness of the model for sensitivity analysis. This allows one to quantify the impact of certain failures on the overall system, to identify critical parameters and adapt data collection, analysis, and potentially the AV design, accordingly.

Our evaluation shows that some of the parameters used in our analysis have only minor effects. Perception and planning errors have, as expected, an important influence on the accident rate. This information can be used by AV manufactures that employ such models early in their design process to adapt their architectures accordingly to achieve the desired level of safety.

Acknowledgment

This work was supported in part by the Intel Collaborative Research Institute on Safety of Automated Vehicles (ICRI-SAVe).

References

1. ISO/IEC, *Road vehicles — Functional safety*, in *ISO 26262 - (1-12):2018*. 2018, ISO/TC 22/SC 32 Electrical and electronic components and general system aspects; Available from: https://www.iso.org/standard/68383.html.
2. ISO, *ISO/PAS 21448:2019, Road vehicles — Safety of the intended functionality*, in *Electrical and electronic equipment*. 2019, ISO. p. 54.
3. IEEE, *IEEE P2846, IEEE Draft Standard for Assumptions for Models in Safety-Related Automated Vehicle Behavior*, in *VT/ITS - Intelligent Transportation Systems* 2020, IEEE SA; Available from: https://standards.ieee.org/ieee/2846/10361/.
4. U. Di Fabio, M. Broy, and R. J. Brüngger, *Ethics commission automated and connected driving*. 2017, Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany: Germany.
5. K. Nidhi and S. M. Paddock *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability*. JSTOR, 2016. 16.
6. R. E. Bloomfield, et al. *Safety Case Templates for Autonomous Systems*. 2021. 136.
7. P. Bourque and R.E. Fairley, *SWEBOK (v.3.0): Guide to Software Engineering Body of Knowledge*, P. Bourque and R.E. Fairley, Editors. 2014, IEEE Computer Society. p. 335; Available from: **Error! Hyperlink reference not valid.**.
8. S. Shalev-Shwartz, S. Shammah, and A. Shashua *On a Formal Model of Safe and Scalable Self-driving Cars*. 2018. **CS**, 37 DOI: https://doi.org/10.48550/arXiv.1708.06374.
9. R. Krajewski, et al., *The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems*, in *21st ITSC*. 2018, IEEE: Indianapolis, US. p. 2118-2125.
10. UK Department of Transport, *Reported road casualties Great Britain, annual report: 2020* 2021, National statistics; Available from: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2020/reported-road-casualties-great-britain-annual-report-2020.
11. F. Geissler, et al., *Towards a Safety Case for Hardware Fault Tolerance in Convolutional Neural Networks Using Activation Range Supervision.* , in *Proceedings of the Workshop on Artificial Intelligence Safety* 2021: online.
12. G. Li, et al., *Understanding error propagation in deep learning neural network (DNN) accelerators and applications*, in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC '17)*. 2017,
13. Eggert, Julian, and Tim Puphal. *Continuous risk measures for ADAS and AD*. Future Active Safety Technology Symposium. 2017.
14. C. Buerkle, F. Oboril, et al., *Safe Perception: On Relevance of Objects for Vehicle Safety*. 2021 IEEE ITSC.
15. F. Oboril, et al., *RSS+: Pro-Active Risk Mitigation for AV Safety Layers based on RSS*. In 2021 IEEE Intelligent Vehicles Symposium (IV).
16. R. Kesten, et al., *Level 5 perception dataset 2020,* https: //level-5.global/level5/data/, 2019