The 12th IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT'2022
9-11 December, 2022, Athens, Greece

Modelling Safety of Connected and Autonomous Vehicles (CAVs) under Cyber-Attacks on Perception and Safety Monitors

Aria Kosari
Department of Computer Science
City, University of London
United Kingdom
Aria.Kosari@city.ac.uk

Peter Popov
Centre for Software Reliability
City, University of London, United Kingdom
p.t.popov@city.ac.uk

Rajkumar Roy
School of Science & Technology
City, University of London
United Kingdom
r.roy@city.ac.uk

***Abstract*— CAVs have recently attracted interest both for researchers and automotive industry.**

Among the various issues with the design of Autonomous vehicles (AV) including CAV, safety and security assurance as well as dealing effectively with the trade-off among these have been recognized as very important. The debate regarding what level of safety and security of (C)AV is socially acceptable is very active at the moment [1], [2].

In this paper, we present a probabilistic modelling approach to dealing with the problem of safety assessment of CAV under cyber-attacks, and demonstrate its plausibility and usefulness in ranking various *modes of vulnerability* of the essential components of (C) AV such as the AV perception system and safety monitors.

Keywords—Digital CAV, Road hazards, Cyber-attacks, Safety, Probabilisitic modelling, Perception system, Safety monitors.

I. INTRODUCTION

In a world where connectivity is ever-increasing, cyber-security is of the utmost importance. This is even more important for safety-critical cyber-physical systems, i.e. systems whose failures can cause physical harm to people (including death) or to the environment.

The vehicles have evolved in recent years from manual transmission, to driver assistance with very significant effort nowadays being expended on bringing the level of autonomy to levels which require little to no input from the driver.

Naturally, as the control is reduced and given to an autonomous system, there will be flaws in the design and implementations of various components used in (C)AVs. Their vulnerabilities, too, create new risks for malfunction caused by malicious agents.

CAVs exacerbate the problem of cybersecurity. CAVs are attractive as they equip each vehicle with ability to communicate with the outside world through connections that allow it to have improved awareness of its surroundings, not merely relying on the human driver or on the sensors of an individual AV. However, the benefits come with new risks – compromises of the road systems intended to improve the situational awareness of the individual CAVs can be used by adversaries as a single point of failure to broadcast inaccurate or outright misleading information about the CAV surroundings.

This ability of an AV to 'see' the environment is delegated to the AV perception system, consisting of the sensors, communications and a machine learning system that makes decisions based on the information it receives [3]. Having the perception and other relevant systems based on machine learning is the key difference between regular vehicles and AVs.

When the perception system fails accidentally or is interfered with by deliberate adversarial actions, the perception will instruct the AV control decisions to take incorrect and potentially dangerous actions, which can include or lead to various lethal scenarios.

Likewise, as with any safety-critical system, in order for CAVs to have vehicle safety, they rely on Safety Monitors (SM). These are independent devices which control the actions taken by the CAV and, in case of danger, would take corrective action (e.g., would stop the CAV safely). An example of SM are devices implementing a set of rules consistent with the Responsibility Sensitive Safety (RSS), proposed by MobilEye, an Intel Corp company [4]. SM, however, too, may be subjected to cyber-attacks and the rules they implement can be altered by an adversary. Successful attack on SM thus, would deny the CAV an essential element of its safety. A (C) AV, certified as safe due to a good SM, may become unsafe should the SM integrity be compromised by a successful attack.

This paper models the effects of successful cyber-attacks on (C)AV perception system and on (C)AV SM. We identified several vulnerability modes of these two subsystems and compare these by looking at their impact on (C)AV safety.

The paper is organized as follows. Section 2 describes in more detail the problem that we study. Section 3 present a probabilistic model used in the paper. Section 4 presents the findings, in section 5 we discuss their implications. Section 6 concludes the work and outlines areas for future study.

II. RELATED RESEARCH

In this section, we provide a summary of the recent relevant research on safety and security, particularly in relation to vehicle perception.

Various CAV communications across short and long distances are discussed in [5]. These communications include V2V (Vehicle-to-Vehicle) and V2I (Vehicle-to-Infrastructure) as the main forms of vehicular communication on the outside.

V2V is low latency, short-range communication that requires messages to be sent quickly when vehicles are in close proximity to inform one another of their status as well as their

own information on the road. The protocols for this are typically Dedicated Short Range Communications (DSRC) and sometimes cellular and Wi-Fi [5].
Communications from the vehicle end typically happen through the Onboard Unit (OBU) of the vehicle, which sends and receives data to other nodes on the vehicular network.

V2I usually happens for longer range information transfer, such as for traffic data being sent across different areas, using the OBU of the vehicles to communicate with the Roadside Unit (RSU), which is the main infrastructure component of CAV networks.
Vehicular communications are typically vulnerable to attacks such as Denial of Service (DoS) and impersonation, namely an attack performed by pretending to be another node such as a vehicle or RSU.

In [6], we see a detailed safety and security review, where it is stated that connected safety-critical systems tend not to be secure and that the solutions implemented in such systems do not always support each other as they could be applied to separate sections, causing issues such as added delay in other subsystems.

Important outlined areas for looking at safety include the methods of evaluating it, including the need to evaluate safety and security at the same time. Risk analysis is a key part of safety and security as modelling could help identify the greatest risks.

An important part of perception is discussed in [3], which is adversarial examples, one of the most serious CAV vulnerabilities related to machine learning. Adversarial examples are when different mathematical inputs are used to perturb images received by the perception system so that the machine learning model classifies them incorrectly. By doing this, the definition of what is being seen is changed and can mislead the CAV control thus endangering (e.g., a pedestrian is overlooked or misclassified).

This is further evaluated in [7], where features such as transferability are considered, meaning that another, similar machine learning model can be built and tested to create the ideal perturbations that are harder to detect.

In [8], we get an evaluation of both attacks and defences in CAVs, starting off by describing the popularity of CAVs, which also make them attractive to those wishing to perform cyberattacks.

The CAV itself is described as being a series of sensors and communication mechanisms when it comes to the perception needed to be autonomous, after which it goes into their vulnerabilities.

All sensors of the types LiDAR, radar and GPS are vulnerable to spoofing and jamming. Spoofing is where previously collected information in the form of sensor signals is sent to the sensors instead of legitimate information, which they will consider correct, having no means of knowing otherwise. This happens in GPS by using stronger signals than usual to overpower the ones belonging to the satellites in space, either to spoof or to completely jam the GPS receiver.

Cameras simply take images, so they can be blinded with bright light, and even damaged with laser light. In other cases, adversarial examples can be fed through the cameras to confuse the machine learning system.

Solutions to attacks on the sensors include things like sensor fusion, where all sensors are combined to have an overall view of the surroundings, or the introduction of redundant sensors so that an attack on one sensor does not compromise that entire section of the perception system. Other solutions include filtering of laser light and even some detection systems.

For the in-vehicle network, the bus is one of the bigger issues, particularly for the CAN (Controller Area Network), described in [9], which indicates the following vulnerabilities in the CAN bus:

- Broadcast transmission allows for eavesdropping.
- Lack of authentication allows injection of false, potentially malicious messages.
- The CAN is able to be flooded with high priority frames so other ECUs (Electronic Control Unit) cannot transmit, essentially DoSing the CAN.
- No encryption on CAN frames makes them readable to anyone connected to the bus.
- Units connected to the attack surface can provide direct access to the CAN, potentially granting total vehicle control.

Various solutions have been proposed to this problem, including methods to detect intruders, introduction of authentication to the bus or components connected to it and even encryption, though this could have unwanted overheads depending on where it is deployed (i.e., on the ECUs).

Finally, a recently developed attack is described in [10], where it is possible to use a form of LiDAR spoofing to make pedestrians and potentially other obstacles completely invisible to the sensor. By injecting invisible echoes of the LiDAR signal near the sensor, legitimate cloud points are discarded so that the obstacle is no longer detected, with the injected LiDAR signals being fired at the angle of detection of the vehicle LiDAR. Simulations in the article showed that the vehicle can accelerate and collide with hidden objects despite the fact that they are only being obscured for short periods of time, though if the planning of the vehicle is designed to be more careful then it is less likely to do so. Methods can also be applied to detect false signals that are made for this purpose in order to prevent it being as easy to implement.

III. METHODOLOGY

The methodology used in this work is based on the approach developed in [11]. The behaviour of an AV is modelled using a stochastic activity network (SAN) model (Fig. 2), in which road hazards are captured as a stochastic process – they occur at random and their duration is also a random variable – an approach consistent with ISO 26262. Examples of road hazards are situations on the road which *may* lead to accidents, but not all hazards lead to accidents. The likelihood of a hazard escalating to an accident is affected by the seriousness of the hazard, and by the quality of the AV *perception* and on AV *safety monitors*. Quality of perception, typically based on machine learning, is limited. Some report rates of failure [12] in excess of 2%, often significantly higher. This level of quality in practice may lead to a significant probability of overlooking road hazards (e.g., objects on the road) which, in turn, may lead to a significant risk of an accident, much higher than if the hazard is detected as soon as it occurs. AV safety is also affected by the quality of safety monitors, which by design should prevent AV accidents even in case failures of some of the AV components. Achieving very high reliability of safety monitors, however, is also problematic [13], which may lead to accidents.

The prior work is extended by adding a model of *cyber-attacks* and modelling in detail the possible effects of successful cyber-attacks on the behaviour of the AV sub-systems critical for AV safety – the perception system and safety monitors. The approach taken here is based on our prior work [14]–[16]. The essence of the adopted modeling approach is that successful attacks can reduce reliability of the compromised software component. We compare the impact on AV safety of different *modes of compromising* different critical for safety AV sub-systems. This comparison allows us to rank the modes of compromise and thus establish the most serious AV cyber-vulnerability.

The probabilistic model relies on the formalism of Stochastic Activity Networks (SANs) [17] supported by the Mobius tool developed by the University of Illinois at Urban Champaign. SAN is an extension of Petri nets, a formalism popular in Computer Science.

The structure of the model is shown on Fig. 1. It includes two "atomic" models[1] known as "StateMachine" and "AttackModel", which will be discussed in turn below.

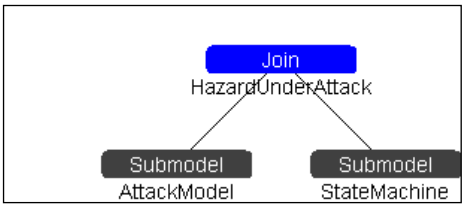

Fig. 1. Structure of the SAN model used in the study

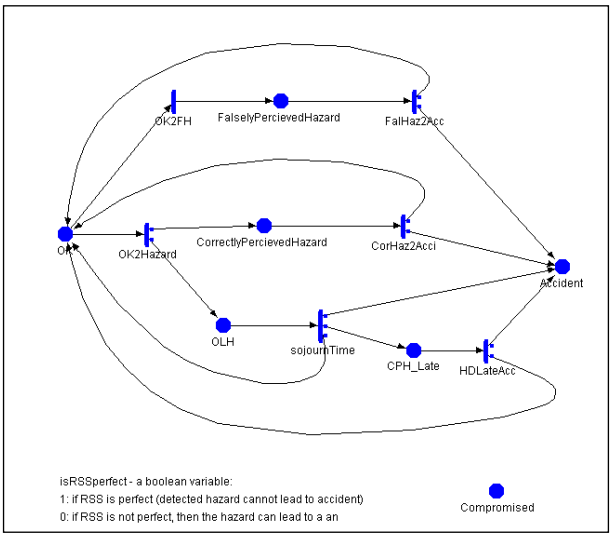

Fig. 2. Vehicle Hazard Model

Fig. 2 shows a stochastic state machine which models the AV operation in the presence of road hazards in a "trusted environment" (e.g., when no cyber-attacks take place but also captures the effect on this behaviour of successful cyber-attacks).

The state of the operational environment oscillates between the state OK, which model the road conditions free of road hazards, and a number of states where a road hazard either occurs or is falsely perceived to have occurred. Below the states in Fig. 2 are explained:

- **OK:** This state models road conditions free from road hazards.
- **Accident:** The vehicle has now had an accident. This is an absorbing state for the model.
- **FalselyPerceivedHazard:** This state models the situation when the AV incorrectly perceives the current road conditions as hazardous (false hazard).
- **CorrectlyPerceivedHazard:** This state models the hazardous state on the road which is correctly detected as hazardous by the AV.
- **OLH:** This state models the hazardous situation on the road which is overlooked by the AV.
- **CPH_Late:** This state models the situation with a road hazard which is eventually detected by the AV as such, but with some delay. For some time since the occurrence of the hazard, the AV has remained unaware of the hazard.
- **Compromised:** This state captures the fact that a cyber-attack on AV has succeeded and, as we will see below, affects the various modelling parameters (the transitions between the states) or the "case" probabilities, which we explain next. The state is in fact a "shared place" in the terminology of SAN: this is a state which is present in a different atomic model "Vehicle Attack Model".

In addition to the states, Fig. 2 includes a number of timed transitions between states, which are modelled as timed activities. These are:

- **OK2FH:** This timed activity captures the intervals between false alarms (i.e. between events of perception system flags the road condition as hazardous when no hazard is actually present).
- **OK2Hazard:** This time activity models the intervals between the hazardous situations on the road.
- **CorHaz2Acci:** This timed activity models the duration of a road hazards which is correctly and timely detected (i.e., as soon as it occurs). There are two alternative ways for a road hazard to finish – either it escalates to an accident, or instead the hazard "goes away" (i.e., the situation on the road is not dangerous any longer). These two options are captured by the two "cases", which can take place at the end of the hazard: one returning the model back to the "OK" state, or the second which leads to the state "Accident". The cases occur with probabilities the sum of which must be 1 (i.e. with certainty one of the options will take place).

- **FailHaz2Acc:** This timed activity models the duration of the false hazard and can have outcomes similar to **CorHaz2Acci** – either the false hazard goes away without any visible consequences and the model returns to "OK" state, or the false hazard escalates to an accident. The two options occur with

[1] "Atomic models" in SAN are used to deal with complexity of large models. A complex model can be split into parts using several atomic models, which are put together using one or more composed models (as the one shown in Fig. 1). The link between atomic models is achieved via a set of "shared places" – these are places which appear in several atomic models and provide semantic links between the parts (atomic models). The reader interested in details, should consult the user guide provided by the Mobius vendors.

probabilities, the sum of which is 1. The possibility of transitioning to an accident is quite plausible in different scenarios – erratic AV driving, when the AV decreases speed without an obvious reason may take the other vehicles in the vicinity by "surprise", thus leading to an accident.

- **sojournTime:** Models the duration of overlooking a hazard, given the hazard has occurred. This interval can end in one of the following possibilities: escalation to "Accident", moving to state "OLH", when the hazard is ongoing and is eventually detected but with some delay, or returning back to "OK" state, if the hazard goes away. These three options are captured by 3 cases and their respective probabilities.
- **HDLateAcc:** This transition models the duration of hazard after its successful detection (with a delay), which may result in an accident or going back to the "OK" state. This transition is conceptually the same as **CorHaz2Acci**, but the length of the hazard which is difficult to detect may differ stochastically from the length of the hazards which are detected immediately upon their occurrence. The model allows for systematic exploration of the differences between lengths of hazards which are difficult or easy to detect.

  In addition to the model of AV on the road, we model attacks on the AV vehicle. The model is shown in Fig. 3.

The second atomic model – "AttackModel" is shown in Fig. 3, which, as the name suggests, captures the cyber-attacks on the AV. Despite its simplicity, the model requires two parameters – the intensity of the attacks and the probability of attack success. Both aspects are captured by the timed activity "Ready2Attack": the rate of this activity (modelled as a global variable "AttackRate", see Table 1 below) models attacks intensity; the case probabilities at the output of timed activity "Ready2Attack" represent the two possible outcomes of an attack instance – a successful or unsuccessful attack. The probability of success is a defined as a modelling parameter ("attackSuccess", see Table 1 below). Under normal conditions the model is in OK state, meaning that the state of the software which might be subjected to attacks has not been compromised. Once an attack is launched and it succeeds, the model moves to a new state "Compromised". The effects of successful attacks are handled in the other atomic model "StateMachine". The atomic model "AttackModel" also captures the possibility of recovery from successful attacks – this happens with intensity defined in the timed activity "CompromisedReset".

The model is such that after the first successful attacks the state Compromised is reached and no further successful attacks will have any impact.

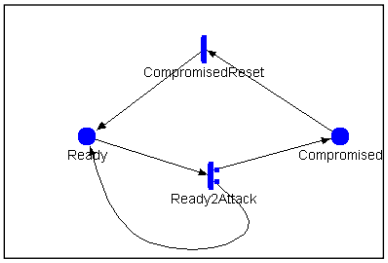

Fig. 3. Structure of Atomic Model AttackModel

The effect of successful attacks may vary depending on the payload of a successful attack (e.g. what the successful attack may try to achieve). We call the hypothetical harm that successful cyber-attacks can lead to "vulnerability modes", following the spirit of Failure Mode and Effect Analysis (FMEA), a popular safety technique, and its recent extension FMFEA [18]. We have conducted an informal analysis of AVs and identified several "vulnerability modes" which an adversary can exploit to harm the operation of AV. We focused this analysis on potential vulnerabilities of AV perception and safety monitors:

- Vulnerabilities of the perception system. The perception system may be harmed in a number of ways:
  - **V_mode 1**: Omissions/misclassifications of road hazards can be exacerbated by successful attacks by reducing the likelihood of hazard detection immediately upon its occurrence. Examples of such attacks are the "adversarial examples" widely documented in the literature.
  - **V_mode 2**: Another related vulnerability mode would be maliciously extending the duration of "hazard blindness" beyond the accidental "blindness" which under normal road conditions is relatively short in the order of no more than a few seconds.
- Vulnerability modes which may affect the reliability of AV safety monitors (SM) is of particular interest. Often safety monitors are designed conservatively to guarantee AV safety even at the expense of availability (false alarms). Successful adversaries may alter the behaviour of SMs (e.g., those defined by RSS [4], [13]) by modifying the integrity of the safety rules that SMs implement. There are three separate aspects that we explore:
  - **V_mode 3**: Increased SM failure rate given the road hazard is detected correctly.
  - **V_mode 4**: Increased SM failure given the road hazards is overlooked.
  - **V_mode 5**: Increased SM failure rate given road the hazard is eventually detected, but with some delay (e.g., in when the road hazard is a "difficult" one).

In this paper we compare the seriousness of the identified vulnerability modes by conducting a "what-if" analysis: hypothesizing that a successful attack may exploit one of the identified vulnerability modes and proceeding compare the safety impact of such exploits in turn. We compare the effect of such exploits against the "base-line", when attacks are disabled and also against the "worst case scenario" whereby the attacker may decide to exploit all vulnerability modes simultaneously.

Other vulnerability modes may be present in AVs and CAVs (e.g., software implementing (C) AV control) such as in the case of vulnerabilities which affect the communication (e.g., speed or integrity) inside an AV or in the case of CAVs, either V2V and/or V2I. Although the approach presented in the paper can be applied to these vulnerabilities, their analysis is outside the scope of this paper.

IV. RESULTS

A. Model parameters

Given the vulnerability modes we introduced a number of model parameters which allow us to model the effect of exploiting the different vulnerability modes. These parameters are described next:

- **HazardScale:** This parameter allows us to study the impact of V_mode 1 on AV safety. By varying it (e.g., increasing it by an order of magnitude) we can establish how sensitive AV safety is to deterioration of hazard detection rate caused by an adversary.
- **DelayScale:** This parameter allows us to look at the impact of V_mode 2, i.e. on how sensitive AV safety is to malicious extensions of road hazard "blindness".
- **AccidentScale:** This parameter allows us to study the impact of exploiting V_mode 3 (e.g., how compromising integrity of SM rules may affect AV safety).
- **OH2CH_scale:** This parameter allows us to assess how sensitive AV safety is V_mode 4.
- **LateCH2AccScale:** This parameter allows us to assess how sensitive AV safety is V_mode 5. This parameter is similar to **AccidentScale** but for those difficult road hazards which are initially overlooked, but subsequently correctly detected (with some delay).

Given the 5 vulnerability modes we conducted 7 experiments: i) the "base line" experiment when attacks are assumed "disabled", ii) 5 experiments when we hypothesize that a successful attack will exploit one of the vulnerability modes only (thus, we vary only one of the 5 scale parameters listed above), and iii) an experiment when we hypothesize that the successful attack will exploit all vulnerability modes simultaneously.

We have tested 7 experiments where we start with all the above parameters set to 0, activating only one for the following five experiments and then activating all of them for the final experiment. From this, we will look at the average time to an accident in each case. The values below are the model parameters adopted for the "base line".

Parameter	Value	Function
CH2AccRate	0.001	Accident rate of correctly perceived hazard [hours^{-1}]
CH2OKRate	856.3	Rate of recovery from correctly perceived hazard [hours^{-1}]
FH2AccRate	1.0E-5	Rate of accident due to falsely perceived hazard [hours^{-1}]
FH2OKRate	856.3	Rate of recovery from falsely perceived hazard [hours^{-1}]
HazardRate	197.4	Rate of hazard occurrence [hours^{-1}]
MH2AccRate	0.01	Accident rate due to delayed hazard perception [hours^{-1}]
OH2OKRate	1000.0	Overlooked hazard recovery [hours^{-1}]
attackRate	0.1	Rate of attack [hours^{-1}]
attackReset	1.0E-4	Rate of attack model reset [hours^{-1}]
attackSuccess	0.5	Rate of successful attack [hours^{-1}]
falseHazardProb	0.0	Rate of false hazards [hours^{-1}]
isRSSperfect	0	Is the safety system perfect or not
missHazardProb	0.05	Rate of missing/overlooking a hazard [hours^{-1}]

Table 1: Fixed Experimental Parameters

All distributions used in the model are assumed to be exponentially distributed. The parameter values listed in Table 1 are derived from publicly available datasets. Details on parameter estimation are provided in [11].

The values of the scale parameters used in the experiments with attacks enabled are given in Table 2.

Experiment	1	2	3	4	5	6	7
AccidentScale	-	10	-	-	-	-	10
DelayScale	-	-	1000	-	-	-	1000
HazardScale	-	-	-	10	-	-	10
LateCH2AccScale	-	-	-	-	10	-	10
OH2CH_scale	-	-	-	-	-	10	10

Table 2: Activated Parameters of Compromised Hazard Perception

For these results, particularly in Table 1, more realistic results were obtained from the ITSC paper where drones had been used to monitor vehicle behaviour on the road, with the exception of the parameters being varied in Table 2, which were chosen with the assumption that any effect on the different parts of the hazard perception system would be bad irrespective of the nature of the attack. The parameters aim to increase the rate or probability of an accident by at least one order of magnitude.

B. Rewards

A measure of interest in all experiments was the probability of accident occurring at pre-defined instances of time in hours of operation. These instances are: 100, 1100, 2100, 3100, 4100, 5100, 6100, 7100, 8100 and 9100

C. Model Solutions

There are two ways to solve SAN models and the relevant experiments:

i. Monte Carlo simulations
ii. Numeric Solvers

Our experiments were solved using a numeric solver for transient analysis, provided by the Mobius tool. The solver was set to obtain values of the rewards with an accuracy of 10^{-9}, which is effectively an exact solution. Monte Carlo simulations would have had to run for millions of repetitions to get results that were anywhere near the level of accuracy afforded by the numeric solver.

The findings are shown in Fig. 4.

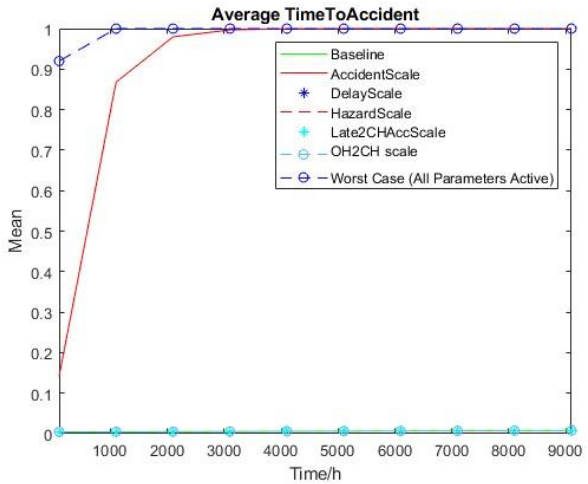

Fig. 4. Comparison of the impact on AV safety of exploiting vulnerability modes

Time [hours]	Ex.1	Ex.3	Ex.4	Ex.5	Ex.6
100	0.004165	0.004189	0.004165	0.004165	0.004165
1100	0.004606	0.004633	0.004606	0.004606	0.004606
2100	0.00502	0.005049	0.00502	0.00502	0.00502
3100	0.005433	0.005465	0.005433	0.005434	0.005433
4100	0.005847	0.005881	0.005847	0.005847	0.005847
5100	0.00626	0.006297	0.00626	0.00626	0.00626
6100	0.006674	0.006712	0.006674	0.006674	0.006674
7100	0.007087	0.007128	0.007087	0.007087	0.007087
8100	0.007499	0.007543	0.0075	0.0075	0.007499
9100	0.007912	0.007958	0.007912	0.007912	0.007912

Table 3: Numerical comparison of vulnerability modes on AV safety.

It is clear from Fig. 4 that the impact on AV safety of exploiting different vulnerability modes differs considerably between vulnerability modes. The impact of maliciously altering the SM rate of failure (AccidentScale) is much greater than the impact of all other vulnerabilities. Increasing this rate by an order of magnitude translates into a dramatic increase of the probability of accident from ~10^{-3} without attacks to 14% with an attack after 100 hours of operation. The probability of an accident quickly escalates for longer times of operation (of 1100 … 9100). In contrast, the impact of exploiting other vulnerability modes is invisible on Fig. 4. The actual values of the chosen reward at different times for those vulnerability modes are detailed in Table 3. The observation that reliability of SM is critically important parameter is consistent with the results reported in [11]. The additional insight that our model provides is the magnitude of the impact under the particular model of cyber - attacks.

Another striking observation from Fig. 4 is that the worst-case scenario (an attack which exploits all vulnerability modes simultaneously) is not merely a sum of the impacts of the different vulnerabilities (most of which are negligible in comparison with the impact of increased AccidentRate). The combined exploit of all vulnerability modes simultaneously leads to almost a certain accident after 100 hours – the probability of accident after 100 hours of operation is greater than 90%. This combined effect is quite surprising indeed.

V. DISCUSSION

While the findings are quite clear from Fig. 4, the recorded observations need further elaboration.

The dramatic impact of AccidentRate is probably due to the fact that the increase of this rate is not countered by any defence. All other vulnerability modes have to overcome *additional barriers*. For instance, the lower branch of the atomic model StateMachine where most of the vulnerability modes are, is relatively rare. Based on the adopted model parameters, under normal circumstances a road hazard is overlooked with a probability of 5%, which is relatively low. An increase of the probability of overlooking a hazard as a result of a malicious activity is possible, of course, but then the "single vulnerability mode exploited" assumption means that if the hazard omission is increased by an order of magnitude (to 50%) the other parameters leading to accident are still low and the model remains "stiff" with a significant likelihood of the road accident going away before an accident occurs. If another vulnerability mode is exploited there is only a 5% chance for overlooking the hazard, which limits the scope for these other exploits to manifest themselves.

Exploiting all vulnerability modes simultaneously changes the situation considerably. A malicious increase of the probability of overlooking a hazard by an order of magnitude will make the two branches of StateMachine model – the branch of correctly and timely detecting a road hazard and the branch of overlooking the hazard – equally likely (each will occur with a probability of 50%). The increased likelihood of overlooking a road hazard, however, now will be exacerbated by the fact that the overlooked hazard will lead to an increased rate of accident before and after the sojournTime. The additional barriers of low accident rates are now removed under the "worst-case" scenario. As a result, the probability of having an accident when the hazard is overlooked becomes very dramatic.

These results summarized in Fig. 4 suggest that the combined effects of multiple vulnerability modes may be significant and protecting against exploiting only some of them (e.g., making the safety motor very reliable and intrusion tolerant [19]) may be insufficient. More detailed analysis of exploiting simultaneously multiple vulnerabilities may reveal additional insight as to how one can allocate defences against exploiting different vulnerability modes under the additional constraint of having a limited budget of making an AV safe and cyber-resilient.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a method of ranking *vulnerability modes* of an autonomous vehicle using a probabilistic model. The model that we used is an extension of the previous work [11] by adding to it a model of a *generic cyber-attack* which may lead to a "compromise" of different components in an AV. We call the location of the compromise vulnerability modes. We concentrate on two important for AV sub-system – the perception system and the safety monitors.

We conducted an informal safety analysis to identify several vulnerability modes and then study the possible impact on system safety of attacks exploiting these vulnerability modes one at a time or all vulnerability modes simultaneously.

Our finding demonstrate that the safety impact of exploiting different vulnerabilities may vary significantly: some vulnerability modes have a negligible impact, while the effect of exploiting others – is quite dramatic. These observations, although expected, reinforce one of the main outcomes of this work that conducting quantitative analysis of the style we outline in this paper may provide an important insight and guide the AV designers where to spend their time, effort and resources on making an AV cyber-resilient.

The second observation worth highlighting is that the combined effect of exploiting several vulnerability modes may be "non-linear" and very dramatic. Even when quantifying the impact of exploiting a single vulnerability is judged "negligible", one should be prepared to quantify the combined effect of exploiting several vulnerability modes simultaneously. Conclusions based on analysing individual vulnerability modes separately and on assumptions that the combined effect could be "predicted" as a linear sum of the individual effects may be grossly inaccurate. This last observation seems particularly important since in safety engineering often the analysis is limited to dealing with a single failure at a time (the aforementioned FMEA is an example). While accidental simultaneous failure can

justifiably be assumed rare[2], the assumption that a successful attack will only exploit a single vulnerability mode is very hard to justify and quantifying the worst possible impact of a successful attack should be based on considerations that a successful attack may exploit more than one vulnerability mode.

The model that we used in the study relied on a significant number of parameters, these were estimated elsewhere [11] using publicly available datasets.

The parameters of the attacks, however (intensity and probability of success) that we used in this work are not based on any empirical data. These are likely to vary significantly between different deployment and operational environments. We intend to extend our survey of literature and look for datasets to help with the parameterization of the AttackModel. An alternative approach would be to conduct sensitivity analysis on the values of the AttackModel parameters. In this case, rather than looking for specific values of attacks intensity and probability of success we will need to establish bounds on the range of values for intensity/probability of success and conduct sensitivity analysis, another area for future research.

Finally, we did not consider falsely perceived road hazards in this work. We plan to extend the presented work in the future and account for the impact of successful attacks that may lead to false alarms.

REFERENCES

[1] IEEE Standards Association, 'IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems', IEEE, Apr. 2022. doi: 10.1109/IEEESTD.2022.9761121.

[2] BSI, 'Road vehicles - Safety of the intended functionality'. BSI Group, 2022.

[3] A. Kumar and S. Mehta, 'A Survey on Resilient Machine Learning'. arXiv, Jul. 11, 2017. doi: 10.48550/arXiv.1707.03184.

[4] S. Shalev-Shwartz, S. Shammah, and A. Shashua, 'On a Formal Model of Safe and Scalable Self-driving Cars'. arXiv, Oct. 27, 2018. Accessed: Nov. 14, 2022. [Online]. Available: http://arxiv.org/abs/1708.06374

[5] Z. El-Rewini, K. Sadatsharan, D. F. Selvaraj, S. J. Plathottam, and P. Ranganathan, 'Cybersecurity challenges in vehicular communications', *Vehicular Communications*, vol. 23, p. 100214, Jun. 2020, doi: 10.1016/j.vehcom.2019.100214.

[6] E. Lisova, I. Sljivo, and A. Causevic, 'Safety and Security Co-Analyses: A Systematic Literature Review', *IEEE Systems Journal*, vol. 13, no. 3, pp. 2189–2200, Sep. 2019, doi: 10.1109/JSYST.2018.2881017.

[7] X. Yuan, P. He, Q. Zhu, and X. Li, 'Adversarial Examples: Attacks and Defenses for Deep Learning'. arXiv, Jul. 06, 2018. doi: 10.48550/arXiv.1712.07107.

[8] M. Pham and K. Xiong, 'A Survey on Security Attacks and Defense Techniques for Connected and Autonomous Vehicles'. arXiv, Jul. 15, 2020. doi: 10.48550/arXiv.2007.08041.

[9] E. Aliwa, O. Rana, C. Perera, and P. Burnap, 'Cyberattacks and Countermeasures For In-Vehicle Networks'. arXiv, Apr. 22, 2020. Accessed: Aug. 22, 2022. [Online]. Available: http://arxiv.org/abs/2004.10781

[10] Y. Cao, S. H. Bhupathiraju, P. Naghavi, T. Sugawara, Z. M. Mao, and S. Rampazzi, 'You Can't See Me: Physical Removal Attacks on LiDAR-based Autonomous Vehicles Driving Frameworks'. arXiv, Oct. 27, 2022. Accessed: Nov. 02, 2022. [Online]. Available: http://arxiv.org/abs/2210.09482

[11] P. T. Popov, C. Buerkle, F. Oboril, M. Paulitsch, and L. Strigini, 'Modelling road hazards and the effect on AV safety of hazardous failures', presented at the The 25th IEEE International Conference on Intelligent Transportation Systems (IEEE ITSC 2022), Macau, China, Jun. 2022. Accessed: Nov. 14, 2022. [Online]. Available: https://openaccess.city.ac.uk/id/eprint/28344/

[12] R. Bloomfield, G. Fletcher, H. Khlaaf, L. Hinde, and P. Ryan, 'Safety Case Templates for Autonomous Systems'. arXiv, Mar. 11, 2021. doi: 10.48550/arXiv.2102.02625.

[13] F. Oboril and K.-U. Scholl, 'RSS+: Pro-Active Risk Mitigation for AV Safety Layers based on RSS', in *2021 IEEE Intelligent Vehicles Symposium (IV)*, Jul. 2021, pp. 99–106. doi: 10.1109/IV48863.2021.9575731.

[14] P. Popov, 'Models of Reliability of Fault-Tolerant Software Under Cyber-Attacks', in *2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE)*, Oct. 2017, pp. 228–239. doi: 10.1109/ISSRE.2017.23.

[15] P. T. Popov, 'Stochastic Modeling of Safety and Security of the e-Motor, an ASIL-D Device', presented at the 34th International Conference on Computer Safety, Reliability, and Security, SAFECOMP 2015, Delft University of Technology, Netherlands, Sep. 2015. Accessed: Nov. 14, 2022. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24255-2_28

[16] O. Netkachov, P. T. Popov, and K. Salako, 'Quantitative Evaluation of the Efficacy of Defence-in-Depth in Critical Infrastructures', Berlin, Germany: Springer International Publishing, 2019, pp. 89–121. doi: 10.1007/978-3-319-95597-1_5.

[17] W. H. Sanders and J. F. Meyer, 'Stochastic Activity Networks: Formal Definitions and Concepts⋆', in *Lectures on Formal Methods and PerformanceAnalysis*, vol. 2090, E. Brinksma, H. Hermanns, and J.-P. Katoen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 315–343. doi: 10.1007/3-540-44667-2_9.

[18] C. Schmittner, Z. Ma, and P. Smith, 'FMVEA for Safety and Security Analysis of Intelligent and Cooperative Vehicles', in *Computer Safety, Reliability, and Security*, Cham, 2014, pp. 282–288. doi: 10.1007/978-3-319-10557-4_31.

[19] P. Sousa, A. N. Bessani, M. Correia, N. F. Neves, and P. Verissimo, 'Highly Available Intrusion-Tolerant Services with Proactive-Reactive Recovery', *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 4, pp. 452–465, Apr. 2010, doi: 10.1109/TPDS.2009.83.

[2] Except in the case of common cause/common mode failures, of course.