Analyzing Influence of Robustness of Neural Networks on the Safety of Autonomous Vehicles

Jin Zhang[1,2,3]

[1]*Computer Science Department,Norwegian University of Science and Technology(NTNU), Norway.*
[2]*Engineering Systems Design Group, Technical University of Denmark (DTU), Denmark.*
[3]*School of Information Science and Technology, Southwest Jiaotong University(SWJTU), China.*
E-mail: jin.zhang@ntnu.no

J.Robert Taylor

Independent consultant and researcher, Denmark. E-mail: roberttayloritsa@gmail.com

Igor Kozin

Independent consultant and researcher, Denmark. E-mail: igor.o.kozin@gmail.com

Jingyue Li

Computer Science Department, Norwegian University of Science and Technology(NTNU), Norway.
E-mail: jingyue.li@ntnu.no

Neural networks (NNs) have shown remarkable performance of perception in their application in autonomous vehicles (AVs). However, NNs are intrinsically vulnerable to perturbations, such as occurrences outside of the training sets, scene noise, instrument noise, image translation, and rotation, or small changes intentionally added to the original image (called adversarial perturbations). Incorrect conclusions from the perception systems (e.g., missing objects, wrong classification, and traffic sign misdetection or misreading) have been a major cause of disengagement incidents in AVs. In order to explore the dynamic nature of hazardous events in AVs, we develop a range of methods to analyze AV safety and security. This work is part of the project and is devoted to analyzing the influence of robustness in the NN-based perception system by using fault tree analysis (FTA). We extend the traditional FTA to represent combinations of failure causes in the multi-dimensional space, i.e., two variables that influence whether the image is classified correctly. The extended FTA is demonstrated on the traffic sign recognition module of AV theoretically and in practice.

Keywords: safety, neural network, autonomous vehicles, robustness, failure mode, hazard identification.

1. Introduction

The development of Autonomous Vehicles (AVs) is proceeding rapidly and promises safer and more efficient roads. However, safety and security problems remain, and disengagement incidents, that is, the handover of vehicle control to a human driver, present a major problem Banerjee et al. (2018). ISO 26262:2011 (2011) and ISO/PAS21448:2019 (2019) intended to address the growing complexity of vehicle systems. However, ISO 26262 does not clearly specify the methods for safety analysis. In the automotive domain, traditional hazard analysis techniques such as Fault Tree Analysis (FTA) and Failure Mode and Effects Analysis (FMEA) or Hazard and Operability Analysis (HAZOP) are generally used for the complex system. In this study, the methods are extended to cover problems arising particularly in Neural Networks (NNs).

One of the major problems in analyzing AV controllers is that of NN components. Deep Neural Networks (DNNs) have been widely used for object detection, image recognition, navigation, and control in AVs. Although DNNs are powerful methods for performing complex tasks compared to humans, they are extremely vulnerable to natural noise Hendrycks and Dietterich (2019) and to small perturbations intentionally added to the input to cause mispredictions Szegedy et al. (2013). A DNN is different from traditional human written programs with certain intended behaviors. Risk analysis of the use of DNNs is at present challenging due to its black-box nature. Analyzing the internal working of a NN with no underlying design is computationally hard Shalev-Shwartz et al. (2017); Johnson (2018). This sets a limit on what can be achieved by hazard identification.

Kalra and Paddock of Rand Corporation made

Proceedings of the 31st European Safety and Reliability Conference
Edited by Bruno Castanier, Marko Cepin, David Bigaud, and Christophe Berenguer
Copyright © ESREL 2021.*Published by* Research Publishing, Singapore.
ISBN: 978-981-18-2016-8; doi:10.3850/978-981-18-2016-8_518-cd

a statistical assessment on the number of miles of driving that would be needed for AV safety Kalra and Paddock (2016). Their results show that demonstrating with 95% confidence that the AV failure rate is 20% better than the human driver failure rate would require 11 billion miles of on-road driving (equivalent to 500 billion vehicle years to complete the requisite miles). This level of testing is impractical. Therefore, it is desirable to analyze safety in the same way that other rare hazards are analyzed, that is, by risk analysis based on component reliabilities and by in-depth assessment of defense. This does not mean that on-road testing would not be needed. On-road testing is an evidence-based way of performing this validation. The risk analysis provides a way of amplifying the value of on-road testing, allowing near miss and partial failure cases to be included in the evidence base while providing a framework for assessing such less serious incidents Taylor et al. (2021).

This paper describes the part of the study that investigates the influence of perturbations in NNs in the context of AVs from an integrated perspective. We consider both safety hazards due to natural perturbations and security threats due to adversarial perturbations as part of an entire system risk assessment. We analyze the failure modes of perturbations in the NN-based perception system by using various hazard identification methods and a combination of methods, i.e., the use of dynamic fault tree methods to explicit reliability analysis of NNs. We also use the Systems Theoretic Process Analysis (STPA) of control loops but include emergent hazards Taylor and Kozin (2021a) as well as component functional failures and the semi-automated fault tree construction to help obtain completeness and consistency in the FTAs. The proposed methods are tested using a design for a 1/4 scale AV. The physical model enables the effects of "real world" problems such as camera resolution, processing response times, the field of view, camera alignment etc., to be investigated in the context of NN performance. An FTA was made for the entire vehicle, including physical, control, and sensor components. The design used as an example for the analysis includes vision algorithms and NNs for control of steering, acceleration, and braking. Due to the space limits, we present the whole FTA in a technical report Taylor et al. (2021).

Our main contribution is to show that NNs and vision algorithms can be included in overall risk analysis in the form of a Fault Tree (FT) by using the concept of exceeding robustness of NNs as FT events alongside the traditional component failure probabilities. The second contribution is that we demonstrate how an FT can include failure events that stem from multiple small deviations of parameters influencing image recognition. These failure events are a very special class of failures that are difficult to identify and quantify. The difficulty is rooted in the phenomenon that arises when all parameters - considered one by one - lie in operational regions. While multiple small variations occur together, they cause performance to fall in a region where the image can be misclassified.

The remainder of the paper is organized as follows: In section 2, we introduce background related to the AV hazard analysis. Section 3 summarizes the hazard identification methods we used for this study. In section 4, we identify both safety and security threats to the NN performance. Section 5 discusses robustness determination and robustness enhancement. Section 6 demonstrates our extended FTA for the traffic sign recognition network both theoretically and in practice. Section 7 concludes the study.

2. AV hazards analysis

AVs are composed of many functional modules – physical, electronic, and software. Since the most important safety issues involve crashes, FTAs provided the overall framework for hazard identification. Still, FMEA was used to provide details of mechanical and electrical component failure, STPA was used to analyze the control hierarchies, and emergent hazard analysis was used for control loop failures. Dynamic methods, including cause consequence analysis and dynamic FTs, were needed, especially for the navigation procedures, such as lane changing navigation functions and emergency response functions. A major problem has previously been that risk analysis of the NNs used for the vision systems and some control functions could not be included in the overall risk analysis. It is, therefore, necessary to extend FTA to incorporate NNs into the overall hazard identification and risk analysis.

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep reinforcement learning (DRL) are the three most common deep learning methodologies used in AVs Grigorescu et al. (2020). CNNs are widely adopted for AV perception. The perception algorithms are the most critical module to detect objects and make image classification. Any incorrect conclusions from the perception algorithms, such as missing objects, wrong classification, and traffic sign misdetection, may lead to potentially fatal incidents. RNNs are suitable for trajectory prediction, and DRL is for path planning, for example, learning driving trajectories.

A vital impact factor for NN hazards is the selection of the training set. Any omission of essential phenomena in the training set will result in a system that may fail to recognize critical cases. This results in the strategy of using massive training sets. Waymo, for example, trains its vision systems for AVs with millions of real traffic scenarios and billions of simulated scenarios Schwall et al.

(2020). However, meeting new phenomena can lead to accidents. The existing AV incidents indicate the difficulties in developing safe AI systems. Even if a system is empirically demonstrated to be safe with millions of tests, there is no guarantee that it will not fail when new situations arise. The selection of test cases needs to consider the wide range of challenges to performance identified by explicit hazard identification.

3. Methodologies for hazard identification of AVs

The overall risk assessment for the AV was made using FMEA for the components and sequential and dynamic FTA Taylor (1975). Sequential FTs are needed to deal with the sequence and timing of responses to hazardous situations versus the dynamic development of the accident situation. If the performance of the NN only depended on independent variations in input parameters, conventional FTs with discrete events could be used, such as ”perturbation exceeds the performance threshold.” Hybrid events are needed because, in many cases, NN’s performance depends on two or more continuously varying disturbance parameters. For this reason, we introduce hybrid events in FTs Taylor and Kozin (2021b) that can be interpreted as a point in a multi-parameter space belonging to the region where safety issues may occur with a rather high probability. The probability of failure is dependent on the probability of challenges to NN robustness. For example, a failure to function is often the result of deviations of two or more parameters, such as a braking force, vehicle speed, and distance to an obstacle at the start of braking. These must be determined empirically (as must failure rates in physical systems). The frequency of challenges can be observed by actually driving typical AV routes at different times and under different conditions. The NN robustness can be measured by the probability of correct image classification (i.e.,prediction accuracy) given perturbed inputs.

4. NN functional failures

One challenge of analyzing NNs is that of seeming randomness in the design of NNs. When the reverse analysis is performed on most NNs trained with a given set of test images, the features that are recognized seem to be distributed in inexplicable ways among the network layers Bengio et al. (2013).

4.1. *Safety threats to NNs*

There is a wide range of situations that can affect the performance of a neural network for AV control:

- Fundamental functional omissions (such as lack of training to recognize road diversion signs)
- Sensitivity to ambient conditions, especially low lighting
- Sensitivity to low contrast conditions
- Sensitivity to patterns (such as camouflage) or textures
- Obscuration due to intended objects hidden behind others or a blind curve or vegetation
- Obscuration by snow, blown sand, frost or ice
- Interference with well-trained recognition by extensions to the training set
- Orientation of the objects to be recognized (“pose”)
- Unusual elevation of objects to be recognized (such as lane markings on a transition to a steep hill)
- Road reflectance lights reflected from wet roads

A straightforward solution is to improve the vision system by data augmentation, sensor fusion, etc. Hendrycks and Dietterich (2019) evaluated NN robustness to common corruptions and perturbations, such as Gaussian noise, motion blur, and snow. They found that as accuracy of NN architectures improves, for instance, from AlexNet to ResNet, corruption robustness has no significant changes. All tested NN models are surprisingly vulnerable to common perturbations. Zhong et al. (2020) reported robustness of thirteen image classifiers and three object detectors to five real-world perturbations, i.e., luminance, spatial transformation, blur, corruption, and weather. Based on their results, some models outperform others for a particular perturbation, and a more complex NN architecture does not necessarily lead to a more robust model. Their results also showed that object detectors are more robust than image classifiers across various real-world perturbations.

4.2. *Security threats to NNs*

In an adversarial context, threats to the neural network could arise from:

- Training data poisoning
- NN model attack
- Adversarial example
- Physical adversarial attack
- Sensor sabotage

Training data poisoning refers to deliberately introduce false data during the training process. NN model attack takes advantage of the model flaws to fool the system. An adversarial example is small changes intentionally added to the original input that are invisible to human eyes. There is a long history of work on understanding, detecting, and mitigating impact of adversarial examples Zhang and Li (2020). Physical adversarial attack aims to fool NN models by creating perturbations on physical objects. Sensor sabotage can be conducted by using spotlights to blind cameras or laser-targeting of cameras. In this study, we focus on the practical consequences of adversarial examples on the design of AV perception models. Evaluating the security threats to NNs is a safety consideration, and adversarial examples can further be used to improve the model robustness.

5. NN robustness measures

Each of the threats to NN performance (introduced in Section 4) requires robustness testing and the probability that each threat will arise needs to be determined. For instance, the likelihood of poor illumination can be determined by driving representative routes at different times using a recording photometer.

5.1. *Robustness determination*

In a traditional risk analysis, the probability of an adverse consequence is determined by obtaining failure probabilities for components (generally by observing over a long period or looking them up in failure rate databases collected from observation). Here, failure probability for a component is derived by determining the robustness against perturbations or attacks, that is, the probability that the robustness limits will be challenged and exceeded. The probability of the AV failing must take account of redundancy in the whole AV system. The contribution of the NN to the AV FT will then be as shown in Fig. 1.

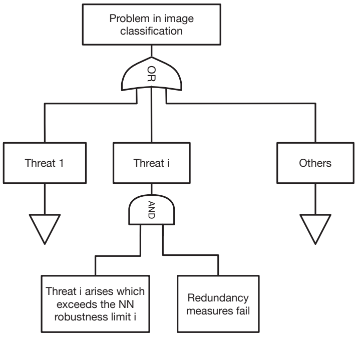

Problem in image classification
OR
Threat 1
Threat i
Others
AND
Threat i arises which exceeds the NN robustness limit i
Redundancy measures fail

Fig. 1. General template for an NN failure subtree in an FTA (for independent threats)

Functional failures of the NN can then be incorporated into fault trees in the form of multiple subtrees in an OR relationship. The probability of failure of the NN in any subtree is then:

$$P_{\text{functional failure i}} = P_{\text{robustness limit i exceeded}} \times P_{\text{redundancy measures fail}} \tag{1}$$

Robustness metrics can be developed to determine the functional range of NNs during testing. Most of the previous works propose accuracy-based metrics to measure NN robustness, i.e., the accuracy (fraction of intended targets recognized) of the NN when inputs are perturbed Hendrycks and Dietterich (2019); Zhong et al. (2020). In an adversarial setting, the minimum perturbation distance (i.e., size of deviation for a loss of function) and adversarial accuracy (i.e., the accuracy of the model when an attack takes place) are two standard metrics to evaluate NN robustness Moosavi-Dezfooli et al. (2016); Zhang et al. (2019). The AV we analyze in this study is relatively simple. Still, the perceptional module of our testing car has over 50 NNs and vision algorithms for different purposes and different navigation situations. There are tens of potential disturbances for each of these, which will affect performance, most being continuous factors rather than discrete yes/no influences. Each of these, and in many cases combinations of these, require robustness tests. Each test can involve hundreds or even thousands of test cases in order to obtain a stable measure of robustness. Laboratory testing is used for robustness determination because it seems doubtful that on-road testing could generate sufficient cases to explore the space of potential failures fully. Laboratory testing has been found to be practicable because the components can be set up and tested automatically.

5.2. *Robustness enhancement*

Data augmentation and increasing model complexity are commonly used approaches for improving NN robustness. However, robustness improvement is not uniform across perturbation types. For instance, increasing performance in the presence of Gaussian noise may cause reduced performance on other perturbations Hendrycks and Dietterich (2019). In Table 1, we identified robustness enhancements to perturbations based on perturbation types. We also map these robustness enhancements into appropriate safety strategies, i.e., inherently safe design, fail-safe design, and safety margins on components Varshney (2016). The inherently safe design aims to exclude potential hazards from the system. Fail-safe design is to keep the system in a safe state at the time of failure. Safety margins on a component are to reserve extra space for achieving safety.

Some defense mechanisms cannot enhance the robustness. For instance, Henriksson et al. (2019) used probability values from a normalized output layer of NNs as anomaly scores because they hypothesize that samples from an outlier distribution will have uncertain class results. This will not be true when the outlier is an adversarial example. Some methods (e.g., adversarial logit pairing Kannan et al. (2018) are less valuable to increase adversarial robustness. But they can be used to remarkably enhance common perturbation

Table 1. Robustness enhancements to perturbations

Perturbation type	Method/Example	Safety strategy
Natural perturbation	Multiscale networks Ke et al. (2017)	Inherently Safe Design
	Feature aggregating Xie et al. (2017)	Inherently Safe Design
	Adversarial Logit Pairing Kannan et al. (2018)	Inherently Safe Design
	Run-time out-of-distribution detection Henriksson et al. (2019)	Fail-safe design
	Histogram equalization Pizer et al. (1987)	Safety Margin
Adversarial perturbation	Adversarial training Madry et al. (2019)	Inherently Safe Design
	Randomized smoothing Lecuyer et al. (2019)	Inherently Safe Design
	Adversarial detection Smith and Gal (2018)	Fail-safe design

robustness Hendrycks and Dietterich (2019).

6. FTA for the traffic sign recognition network

The starting point and basis for safety analysis of AVs is a functional block diagram picturing all top-level functions and connections between them. A hazard identification analysis can be made by analyzing each function and indicating components/subsystems for their failure modes and effects (functional FMEA analysis). To identify more complex failure scenarios caused by several failures, degraded performances, and other internal and external factors, like weather and road conditions, causal models are needed. In this paper, we focus on FTs that, if properly analyzed, can generate a comprehensive set of hazard scenarios and provide the basis for the use of probabilistic reasoning to estimate the probabilities of the identified scenarios. However, constructing FTs for NN-controlled AVs is not a standard procedure and requires a substantial modification of classical FTA. This is due to two reasons. One is a possible malfunction of the NN and the difficulty of constructing the internal causal structure, resulting in outputting erroneous decisions. The second is that continuously changing processes (variables) influencing a vehicle's performance (possibly in combination) can result in safety issues and eventually in crashes. The second point motivates us to introduce failure events that manifest themselves when continuously evolving variables in a multi-dimensional space enter the "prohibited region". This is like in structural reliability – a failure occurs when stress exceeds the strength of the construction.

Given that the functional requirement placed on a NN is that of a simple function, such as recognizing a traffic sign, the NN can be considered a black box. The failure modes can be defined as failure to identify an image, incorrect classification of an image, or in some cases, wrong estimation of an image parameter. The NN will have a certain correct performance set and a certain level of robustness against image imperfections or distortions. The probability of failure of the NN is then the probability of the observed image lying in a domain outside the NN's capability or in a domain for which the NN is not robust. The hazard analysis can then be completed using standard methods (e.g., conventional FTA) to determine the possible causes of the inputs lying outside the NN's reliable domain. Our emphasis is placed on developing robustness measures for NNs against different types of threats.

6.1. *Problem formalism*

We propose a mathematical formalism to be able to calculate the probabilities of failure states. One of the possible hazardous events triggered by a decision made by the NN is the "Wrong classification of a traffic sign". This event can occur because of inadequate robustness of the NN, which in turn can be caused by naturally or intentionally perturbed inputs.

Robustness can be measured by the prediction accuracy given perturbed inputs. The prediction accuracy is unlikely to achieve unity, and there is a threshold of $T_r < 1$ where, if achieved, the NN decides that the image in question is recognized. Hence there is always a probability of misclassification that is greater than 0.

Assume that two variables influence whether the sign is classified correctly. One is contrast intensity, C, and the other is light intensity (i.e., brightness), L. If T_C stands for the lower limit for C, below which the sign cannot be classified correctly, we can define the event $E_C = \{E_C : c < T_C\}$ that is "too low contrast to recognize correctly". Similarly, $E_L = \{E_L : l < T_L\}$ is the event "too low lighting to recognize correctly". The third misclassification event is defined by the following condition: $E_{LC} = \{E_{LC} : (l, c) < f(c, l), c > T_C, l > T_L\}$. This should be understood as follows: while contrast and lighting both lie in the correct classification region, their combination may belong to the misclassification region. The border dividing the two regions is determined by function $f(c, l)$. Usually, this type of event occurs when variables (parameters) lie in the vicinity of the border points. That is to say, the effect of small deviations results in a failure. A possible region of misclassification is shown in Fig. 2.

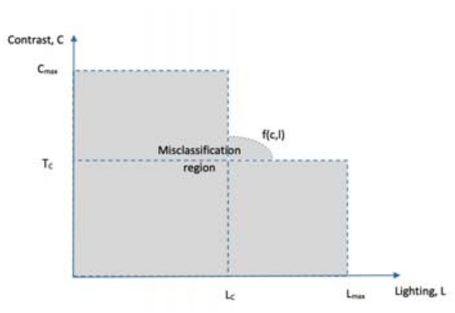


Fig. 2. Misclassification region(Conceptual)

The region of misclassification can formally be written as follows:
$\Omega = \{(c < T_C)\bigcup(l < T_L)\bigcup((l,c) < f(c,l), c > T_C, l > T_L)\}$

As soon as the misclassification events are determined, a simple fault sub-tree can be constructed (see Fig. 3).

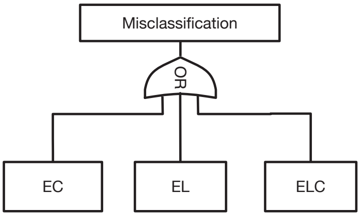


Fig. 3. A simple fault sub-tree for misclassification (with interacting threats)

Given C and L are independent random variables and their probability density functions are known, $f_C(x)$ and $f_L(y)$, the probability of misclassification $P_{\text{misclassification}}$ can be calculated:

$$P_{\text{misclassification}} = \iint_{\Omega} f_C(x) f_L(y) dx dy \qquad (2)$$

6.2. *An AV example of misclassification*

To demonstrate the influence of the perturbations and their combination, we trained a 5-layer-CNN with the German Traffic Sign Recognition Benchmark (GTSRB) dataset for the traffic sign classification Stallkamp et al. (2012). The GTSRB dataset has 43 different traffic signs in various sizes and lighting conditions and is very similar to real-life data. The prediction accuracy for clean test images is 98.97%.

We adopt the algorithm from Zhong et al. (2020) to emulate the deviation of brightness and contrast, and algorithm from Goodfellow et al. (2014) to implement the FGSM attack. Fig. 4 presents: (a) a set of misclassified images with brightness=0.8. In this case, the prediction accuracy dropped to 84.8%, (b) brightness=0.6, FGSM attack with attack strength=0.2, the prediction accuracy dropped to 18.76%.

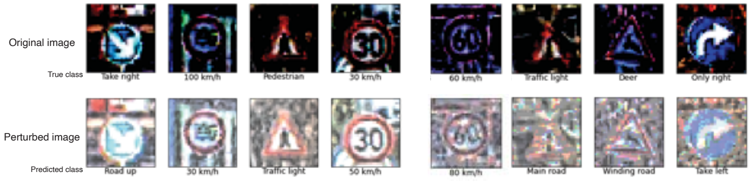
1) Brightness $X' = Clip(X + l)$, where X is the original test image, l is a constant value to be added, $X'$is the resulting new image, Clip is a function to make sure X' is in a valid pixel intensity range of [0,255] or [0,1].

2) Contrast Reduction $X' = Clip((1 - c) \cdot X + c \cdot C)$, where X is the original test image, c is the contrast level, C is a constant factor.

In this experiment, we set prediction accuracy at 90% as the acceptance level of model robustness. Instead of showing the case of low brightness/contrast, we test the influence of increasing brightness and contrast reduction due to the low brightness/contrast nature of the GTSRB dataset. Fig. 5 shows prediction accuracy curves corresponding to (a) brightness variations, and (b) contrast variations. It shows that the upper limit for brightness increase is 0.66 in Fig. 5 (a), and the upper limit for contrast reduction is 0.54 in Fig.5 (b).

Then we test the combination of brightness and contrast reduction. The brightness level is set from 0.01 to 1, and contrast reduction is from 0.01 to 1, respectively. This experiment is intended to show how the small deviation of contrast and brightness affects prediction accuracy. In Fig. 6, the values of prediction accuracy are represented as colors. The lighter the color, the higher the prediction accuracy. It shows that even brightness level and contrast reduction do not exceed their upper limits (i.e., in the correct classification region). Their combination can fall into the misclassification region (i.e., prediction accuracy is lower than 90%).

It is worth noting that contrast and lighting are just two of the challenges to the NN performance, which require a hybrid fault tree approach. In fact, almost all of the threats listed in Sections 4.1 and 4.2 have continuously varying intensities. In most cases, pairs of threats can interact to make the joint deviation worse than any single deviation alone. A particularly difficult example that was found is obscurations coupled with shadows. Some of the threats (e.g., adversarial examples) are hard for a human to understand. Methods in the field of explainable AI (XAI) can be employed to identify the influence of threats on the NN performance Zhang and Li (2020). We include more results and discussions in a technical report Taylor et al. (2021) due to the page limits.



(a) brightness=0.8

(b) brightness=0.6, FGSM attack strength=0.2

Fig. 4. Examples of misclassified traffic signs

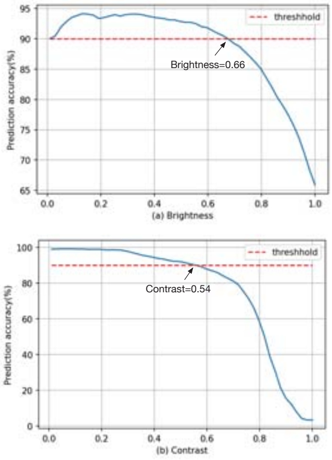
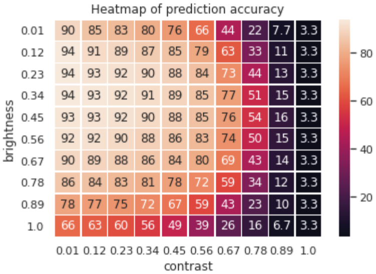


Fig. 5. Examples of prediction accuracy curves when brightness and contrast vary



	0.01	0.12	0.23	0.34	0.45	0.56	0.67	0.78	0.89	1.0
0.01	90	85	83	80	76	66	44	22	7.7	3.3
0.12	94	91	89	87	85	79	63	33	11	3.3
0.23	94	93	92	90	88	84	73	44	13	3.3
0.34	94	93	92	91	89	85	77	51	15	3.3
0.45	93	93	92	90	88	85	76	54	16	3.3
0.56	92	92	90	88	86	83	74	50	15	3.3
0.67	90	89	88	86	84	80	69	43	14	3.3
0.78	86	84	83	81	78	72	59	34	12	3.3
0.89	78	77	75	72	67	59	43	23	10	3.3
1.0	66	63	60	56	49	39	26	16	6.7	3.3

Fig. 6. Prediction accuracy matrix with small deviation of brightness and contrast in combination

7. Conclusion

From this study, it became clear that detailed hazard identification can be made for AVs, including both hardware and NN components. The procedure is:

(1) Complete the overall high-level hazard identification using an FTA approach.
(2) Identify the functional failures of the NNs which contribute to the overall FTA.
(3) Identify the challenges which can cause the NN functional failure, e.g., using the checklist in Sections 4.1 and 4.2.
(4) Determine the robustness of the NNs when challenged by perturbations of single parameters or by the combination of parameter perturbations via testing NN performance and making a heatmap as in Fig. 6.
(5) Determine the probability of the occurrence of parameter perturbations.
(6) Incorporate the contribution of NNs into the FTA using the templates given in Fig. 1 and Fig 3.

A further conclusion is that a detailed hazard assessment can be essential in determining the scope of controller component testing.

One of the key findings of the studies described here is that safety and security analysis becomes much easier when an integrated approach is taken. There are many potential cases where individual controller components (e.g., NN for image recognition) can fail due to an attack, but where accidents can be avoided by other components taking over. This is particularly an issue where there is a possibility of a crash and poor visibility conditions. In these cases, lidar and radar provide less informative but more robust detection of hazards.

Safety in AVs is not ensured by hazard detection alone. It is not safe, for example, to simply stop the vehicle when a crash potential is detected in fast-moving traffic. Policies, strategies, plans, and algorithms for safe state recovery are needed. Our next challenge, then, is to carry out hazard identification and risk assessment on these recovery plans.

References

Banerjee, S. S., S. Jha, J. Cyriac, Z. T. Kalbarczyk, and R. K. Iyer (2018). Hands off the wheel in autonomous vehicles?: A systems perspective on over a million miles of field data. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 586–597. IEEE.

Bengio, Y., G. Mesnil, Y. Dauphin, and S. Rifai (2013). Better mixing via deep representations.

In *International conference on machine learning*, pp. 552–560. PMLR.

Goodfellow, I. J., J. Shlens, and C. Szegedy (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grigorescu, S., B. Trasnea, T. Cocias, and G. Macesanu (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics 37*(3), 362–386.

Hendrycks, D. and T. Dietterich (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint*.

Henriksson, J., C. Berger, M. Borg, L. Tornberg, S. R. Sathyamoorthy, and C. Englund (2019). Performance analysis of out-of-distribution detection on various trained neural networks. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 113–120. IEEE.

ISO 26262:2011 (2011, November). Road vehicles – Functional safety. Standard, International Organization for Standardization.

ISO/PAS21448:2019 (2019, January). Road vehicles — Safety of the intended functionality. Standard, International Organization for Standardization.

Johnson, C. (2018). The increasing risks of risk assessment: On the rise of artificial intelligence and non-determinism in safety-critical systems. In *the 26th Safety-Critical Systems Symposium*, pp. 15. Safety-Critical Systems Club York, UK.

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

Kannan, H., A. Kurakin, and I. Goodfellow (2018). Adversarial logit pairing. *arXiv preprint*.

Ke, T.-W., M. Maire, and S. X. Yu (2017). Multigrid neural architectures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6665–6673.

Lecuyer, M., V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana (2019). Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE.

Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu (2019). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Moosavi-Dezfooli, S.-M., A. Fawzi, and P. Frossard (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582.

Pizer, S. M., E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing 39*(3), 355–368.

Schwall, M., T. Daniel, T. Victor, F. Favaro, and H. Hohnhold (2020). Waymo public road safety performance data. *arXiv preprint arXiv:2011.00038*.

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

Smith, L. and Y. Gal (2018). Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*.

Stallkamp, J., M. Schlipsing, J. Salmen, and C. Igel (2012). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks 32*, 323–332.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Taylor, R. (1975). Sequential effects in failure mode and fault tree analysis. *Reliability and Fault Tree Analysis*.

Taylor, R. and I. Kozin (2021a). Design for emergent safety problems in handbook of engineering systems design. In-press, Springer.

Taylor, R. and I. Kozin (2021b). Hybrid fault trees for continuous systems. Unpublished manuscript.

Taylor, R., J. Zhang, I. Kozin, and J. Li (2021). Safety And Security Analysis for Autonomous Vehicles. `https://github.com/safe-ai-tech/Reports_Papers`. [Technical report].

Varshney, K. R. (2016). Engineering safety in machine learning. In *2016 Information Theory and Applications Workshop (ITA)*, pp. 1–5. IEEE.

Xie, S., R. Girshick, P. Dollár, Z. Tu, and K. He (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.

Zhang, H., Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan (2019). Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR.

Zhang, J. and J. Li (2020). Testing and verification of neural-network-based safety-critical control software: A systematic literature review. *Information and Software Technology*, 106296.

Zhong, Z., Z. Hu, and X. Chen (2020). Quantifying dnn model robustness to the real-world threats. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 150–157. IEEE.