Could We Issue Driving Licenses to Autonomous Vehicles?

Jingyue Li[1], Jin Zhang[1], and Nektaria Kaloudi[1]

Norwegian University of Science and Technology, NO-7491, Trondheim, Norway
{jingyue.li, jin.zhang, nektaria.kaloudi}@ntnu.no

Abstract. Many companies are studying autonomous vehicles. One trend in the development of control algorithms for autonomous vehicles is the use of deep-learning approaches. The general idea is to simulate a human driver's decision-making and behavior in various scenarios without necessarily knowing why the decision is made. In this position paper, we first argue that traditional safety analysis methods need to be extended to verify deep-learning-based autonomous vehicles. Then, we propose borrowing ideas from the process of issuing driving licenses to human drivers to verify autonomous vehicles. Verification of autonomous vehicles could focus on sufficient training as well as mental and physical health checks. Based on this position, we list several challenges that need to be addressed.

Keywords: Autonomous vehicle · Artificial intelligence · Deep learning · Verification · Safety · Security.

1 Introduction

Many companies, e.g., Google [1], are developing autonomous vehicles. One key challenge of developing autonomous vehicles is ensuring safety. Several safety incidents caused by vehicle autonomy have been presented in the media, such as Uber's fatal car accident [2]. In addition to the safety incidents caused by failures of the autonomous system, security breaches of autonomous vehicles can potentially lead to safety issues; for example, a demo showed that autonomous vehicles can be hijacked and remotely controlled [3].

The Society of Automotive Engineers (SAE) has described six levels of autonomous driving [4]. A Level 0 vehicle has no autonomous capabilities and the human driver is responsible for all aspects of the driving task. For a Level 5 vehicle, the driving tasks are managed solely by the autonomous driving system. When developing autonomous vehicles targeting a high level of autonomy, one industrial trend is to use deep-learning algorithms to implement the vehicle control algorithms. The idea is to first log the information, such as images a human driver acquired during driving and the driver's corresponding driving behavior. Such logged information is used as a training dataset for deep-learning algorithms to train the autonomous vehicles to simulate what human drivers do when driving. One key characteristic of deep-learning-based autonomy is that

the decision-making part of the vehicle is almost a black box. This means that in most cases, we as human drivers must trust the decisions made by the deep-learning algorithms without knowing exactly why and how the decisions are made. As an analogy, driving an autonomous vehicle with a high level of autonomy (e.g., Level 5) is like hiring a robotic taxi driver who is driving the car for us. Usually, when we sit in a taxi, we do not always try to understand and influence how the taxi driver makes decisions and drives the car. We simply trust that the taxi driver's driving licenses indicate that he or she has sufficient training, good mental health to make proper decisions, and good physical health to sense the environment and execute the decisions. With this analogy in mind, we propose adapting safety analysis approaches for a greater focus on:

- **Training sufficiency of the autonomous vehicle**, i.e., whether the dataset used to train the deep-learning algorithms is sufficient;
- **Mental health of the autonomous vehicle**, i.e., whether there is malicious code hidden in the decision-making algorithms; and
- **Physical health of the autonomous vehicle**, i.e., whether the sensors and actuators of the autonomous vehicle work properly and whether the vehicle is resilient when the devices fail.

2 Background

Computer vision and deep learning are two major approaches to designing autonomous vehicle control algorithms. Traditional computer vision techniques can play an important role in lane detection and object detection at moderate distances, but they are unlikely to meet the robustness requirements for handling very complex and intelligent tasks such as distinguishing between different metal objects or unexpected obstacles.

2.1 Deep-Learning-Based Autonomous Vehicles

The deep-learning based approach enables vehicles to learn meaningful road features from raw input data automatically and then output driving actions. The so-called end-to-end learning approach can be applied to solve complex real-world driving tasks. When using deep-learning based approaches, the first step is to use a large number of training datasets (images and/or other sensor data) to train a deep neural network (DNN). Then a simulator is used to evaluate the performance of the trained network. After that, the deep-learning-based autonomous vehicle will be able to *"execute recognition, prediction and planning"* driving tasks in diverse conditions [12]. Nowadays, Convolutional Neural Networks (CNNs) are the most widely adopted deep-learning model for fully autonomous vehicles [5–8]. In 2016, NVIDIA introduced an automotive supercomputing platform named DRIVE PX 2 [9]. DRIVE PX 2 is being used by more than 50 companies in the automotive industry. The development flow by using NVIDIA DRIVE PX 2 includes 1) data acquisition to train the deep neural network; 2) deployment of the output of a deep neural network in a car; 3) autonomous application development; and 4) testing in-vehicle or with simulation.

2.2 Some Approaches to Analyzing Safety of Autonomous Vehicles

The safety standard of the automotive industry, ISO 26262, is being updated to address the growing complexity and autonomy of vehicles. Besides using classical safety analysis methods such as FTA (Fault Tree Analysis) and FMEA (Failure Mode and Effects Analysis), production of a safety case is explicitly mandated by ISO 26262. A safety case comprises three parts: (1) the safety goal that must be achieved; (2) the available evidence for achieving this goal; and (3) the structured argument, which establishes the systematic relationship between the evidence and the goals. One challenge of using the safety case approach is arguing that the evidence is sufficient to ensure safety of the system. The forthcoming version of ISO 26262:2018 and its extension, ISO/PAS 21448, which is also known as SOTIF (Safety of the Intended Functionality) [10], will likely provide a way to handle the development of autonomous vehicles. But SOTIF will only provide guidelines for Level 0-2 autonomous vehicles [11], which are not designed for the validation of deep-learning-based autonomous vehicles.

Along with updating the safety verification standards, some studies investigate how to verify safety of fully autonomous vehicles by treating the autonomous vehicle control algorithms as black boxes. The general idea is to use a combination of brute force road testing and testing using simulators to enumerate all potential corner cases. The proposed safety metrics of autonomous vehicles include Miles Per Disengagement (MPD) and Miles Per Intervention (MPI) [12]. Some other studies try to open the black boxes to interpret the deep neural networks and verify their internal logics [13–15].

2.3 Security Attacks Targeted at Autonomous Systems

As the development of Artificial Intelligence (AI) technologies progresses, attackers will also learn to create new smart attacks. We define a smart attack as an AI-enabled attack in which malicious actors can use AI technologies to attack "smart" components inside autonomous systems. The smart attack is usually executed via a persistent, finely targeted, combined, and multilayered exploitation of multiple security zones in a camouflaged way [16]. Examples of potential smart attacks include:

- **Training smart systems to have two behaviors**, e.g., a robot can be trained to behave normally in most cases, but behave maliciously and make an attack when a certain face is recognized [17];
- **Training systems to personalize the hack**, e.g., an attacker can train systems to generate a finely personalized vulnerability profile and then perform the hack by creating tailored exploits for such a vulnerability [18, 19];
- **Training systems to evolve themselves**, e.g., malicious code can continuously update itself with dozens of new exploits by using fuzzing techniques [17]; or
- **Distributing AI-generated content**, e.g., an attacker can automate tasks involved in surveillance, persuasion, deception, privacy violation, and social manipulation by distributing AI-generated content and targeted disinformation campaigns through social media [20].

3 Key Issues of Verifying Deep-Learning-Based Autonomous Vehicles

When driving Level 5 autonomous vehicles, the human driver will behave like a passenger of a taxi. The taxi driver is now the deep-learning based-control algorithms. As a passenger of a taxi, we usually trust that the taxi driver is sufficiently trained because we trust the taxi driver training program and the qualification a driving license implies. Normally, if a taxi driver is well-trained, sensible, and in good health, and if the hardware and software of a vehicle is functioning, safety is guaranteed. However, as mentioned in Section 2.2, most current safety analyses, certification approaches, and standards focus only on whether the vehicle's hardware and software are working as intended. The qualification of the taxi driver is defined in a separate standard by which driving licenses are issued, which is often regulated by the police and followed by driving schools. For fully autonomous vehicles, the control algorithm is an integrated part of the vehicle. We therefore argue that the safety analysis and certification approach should be extended to treat the control algorithms as a taxi driver and to test it to answer some important questions.

3.1 Has the Autonomous Vehicle Been Sufficiently Trained?

When we study in driving school, a complete training program starts with driving theories and rules. We first learn different road signs and to understand how to drive the vehicle according to those road signs and driving regulations. Afterward, we need to practice driving in different scenarios, such as in the city center, through a roundabout, on the highway, in slippery conditions, and so on. In addition, when driving assessments are carried out by driving instructors to evaluate drivers' behavior, a formal process aims at fixing the drivers' errors.

When we take the driving license test, we are supposed to show competence to drive the vehicle properly in different scenarios, including scenarios we may know in theory but have not practiced, such as giving way to emergency vehicles. When verifying the completeness of the training dataset of the autonomous vehicles, how can we learn from the driving school and find ways to train autonomous vehicles and quantify their learning completeness? To improve training sufficiency, the "error analysis" process of examining the instances in which the deep-learning algorithms erred can also help suggest good practices and develop new features. Brute force road testing is not an efficient way to assure safety. The traffic signs and regulations of countries are different. For example, a white dotted line is used in Sweden to separate lanes, but a yellow dotted line is used in Norway. If the autonomous car is trained using Swedish traffic regulations, it may become confused when it drives in Norway. Thus, tests measured by MPD or MPI [12] in one country may not be valid in another country. In addition, what happens if the traffic regulations of a certain country are updated? Should we undertake another billion miles of road testing?

3.2 Is There Any Malicious Code in the Brain of the Autonomous Vehicle?

When a driving license is issued to a driver, the driver should not have severe mental health problems. When a taxi driver is working, the driver is not supposed to be drunk. As explained in Section 2.3, successful smart attacks can gain access to the decision-making algorithms of autonomous vehicles. Malicious inputs into training datasets can cause the model to behave normally in most cases, but behave maliciously in a certain scenario. Because few smart attacks have been exploited in practice, people have not reported them in vulnerability repositories, and therefore have not studied in depth how to identify such attacks and defend against them. However, we expect such AI-based attacks will be perpetrated in the future [17]. If the attack is carried out, the consequence could be that the autonomous vehicle suddenly behaves like a drunk or mentally compromised taxi driver. When certifying deep-learning-based autonomous vehicles, we should require the vehicles to have self-checking or remote-checking mechanisms to ensure that no malicious code has ever been inserted in the control algorithms.

3.3 Are the Sensors and Actuators of the Autonomous Vehicle Reliable and Resilient to Failures?

To get a driving license, a human driver should have physical health, e.g., good eyesight and capability to operate the vehicle in normal and abnormal situations. Current safety certification standards focus sharply on the reliability of vehicle hardware and software. Analyzing failure modes and how the vehicles react to failure is a crucial part of the safety analysis. The architecture of deep-learning neural networks makes it hard to decipher how the algorithmic decisions were made, which in turn makes it hard to explain how dynamic driving behaviors are generated [21]. Thus, it can be difficult to interpret and predicate how a failure, such as wrong sensor data, will influence driving behavior. When we verify the safety of deep-learning-based control algorithms, we need to rethink how to perform failure mode and effect analysis, how to analyze interdependencies between sub-systems of a vehicle, and how to assure the resilience of the system. For resilience assurance, we need to determine where to put safety barriers and how to place them in the deep neural network to ensure that even if some vehicle hardware or software does not behave as expected, the vehicle can still sense the risk, avoid the risk before the incident, and mitigate the risk effectively when the incident happens.

3.4 Human vs. Machine

Ideally, autonomous cars should behave equally or even better than human beings. Besides the three principles mentioned above, a comparison between human and machine capabilities is needed to identify some limitations that should be considered as a further evaluation of autonomous vehicles. As shown in Table 1, a

Table 1. Human and machine superiority.

Human	**Machine**
Originality and creativity	Precise repetitive operations
Emotions and feelings	Mechanical brain
Rapid retraining	Quicker response times with minimum delay
Performing under overloaded conditions	Storing and recalling large amounts of data
Acting in high-noise environments	Sensitivity to a variety of stimuli
Risk evaluation for unexpected events	Function in a wide range of stress conditions
Using equipment beyond specified limits	Stronger and faster
Reasoning inductively	Reasoning deductively

function analysis between humans and machines during space missions identifies the differences in their superiority [22].

A human is shown to be better at "risk evaluation for unexpected events" and "rapid retraining" than a machine. For example, when the car suddenly experiences a longer breaking distance than normal, the human driver will realize that the road is slippery and will drive slower and more carefully. The "rapid retraining" competence of a human is usually not verified during driving license tests because we view it as human nature. If we want to have autonomous cars with performance superior to that of humans, and if we use the human driving license approach to verifying autonomous vehicles, we also need to consider the importance of testing human superiority in the entire evaluation process of autonomous vehicles.

4 Conclusions and Future Work

Our position is that certifying a deep-learning-based autonomous vehicle is like issuing a driving license to an AI-based taxi driver. To verify safety, we need to learn from the systematic method of training and testing human drivers. We need to guarantee that the training dataset of the autonomous vehicle covers all knowledge and skills taught in a driving school. We should have technologies to ensure that no malicious code is hidden in the autonomous vehicle either in design or in operation. The vehicle should also have highly reliable hardware and software and should be resilient in the face of expected and unexpected failures. When we make safety cases according to ISO 26262, we propose including all these aspects as safety arguments and evidence. To acquire evidence for these arguments, we will also need to combine black box testing technologies to test deep-learning algorithms with technologies to understand and interpret them.

Acknowledgements. This work is supported by the SAREPTA (Safety, autonomy, remote control and operations of industrial transport systems) project, which is financed by Norwegian Research Council with Grant No. 267860.

References

1. Google, "The Google self-driving car", https://www.google.com/selfdrivingcar/. Last accessed May 2018
2. Andrew J. Hawkins: "Uber self-driving car saw pedestrian but didn't brake before fatal crash, feds say", https://www.theverge.com/2018/5/24/17388696/uber-self-driving-crash-ntsb-report. Last accessed 24 May 2018
3. A. Greenberg: "Hackers remotely kill a Jeep on the highway", https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/. Last accessed 21 July 2015
4. SAE International, "Automated vehicles: Levels of automation", https://autoalliance.org/wp-content/uploads/2017/07/Automated-Vehicles-Levels-of-Automation.pdf. Last accessed May 2018
5. A. E. L. Sallab et al.: "Deep reinforcement learning framework for autonomous driving". Electronic Imaging, vol. 2017, no. 19, pp. 70-76, (2017)
6. M. Bojarski et al.: "End to end learning for self-driving cars". arXiv preprint arXiv:1604.07316, (2016)
7. B. Huval et al.: "An empirical evaluation of deep learning on highway driving". arXiv preprint arXiv:1504.01716, (2015)
8. A. Navarro et al.: "Development of an autonomous vehicle control strategy using a single camera and deep neural networks". SAE Technical Paper, pp. 01-0035, (2018)
9. NVIDIA Deep Learning Institute, "Deep learning for autonomous vehicles-perception", https://www.nvidia.com/en-us/deep-learning-ai/education/. Last accessed May 2018
10. G. Griessnig et al.: "Development of the 2nd Edition of the ISO 26262". European Conference on Software Process Improvement, pp. 535-546: Springer, (2017)
11. The Hansen Report on Automotive Electronics, "Standardization Efforts on Autonomous Driving Safety", http://www.hansenreport.com/. Last accessed Feb 2017
12. WAYMO, "Waymo Safety Report: On the Road to Fully Self-Driving", https://waymo.com/safety/. Last accessed May 2018
13. Y. Tian et al.: "DeepTest: Automated testing of deep-neural-network-driven autonomous cars". arXiv preprint arXiv:1708.08559, (2017)
14. X. Huang et al.: "Safety verification of deep neural networks". pp. 3-29: Springer, (2017)
15. G. Katz et al.: "Reluplex: An efficient SMT solver for verifying deep neural networks". pp. 97-117: Springer, (2017)
16. R. Koch et al.: "A revised attack taxonomy for a new generation of smart attacks". Computer and Information Science, vol. 7, no. 3, p. 18, (2014)
17. M. Brundage et al.: "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". arXiv preprint arXiv:1802.07228, (2018)
18. A. Giaretta, and N. Dragoni: "Community Targeted Spam: A Middle Ground Between General Spam and Spear Phishing". arXiv preprint arXiv:1708.07342, (2017)
19. J. Seymour, and P. Tully: "Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter". Black Hat USA, (2016)
20. Y.M. Kim: "The Stealth Media? Groups and Targets Behind Divisive Issue Campaigns on Facebook". (2018)
21. M. T. Ribeiro et al.: "Model-agnostic interpretability of machine learning". arXiv preprint arXiv:1606.05386, (2016)
22. F. Schenkelberg: "Comparing Human and Machine Capability", https://accendoreliability.com/comparing-human-and-machine-capability/. Last accessed 2018