Work-in-Progress: Crash Course
Can (Under Attack) Autonomous Driving Beat Human Drivers?

Marchiori, Francesco; Brighente, Alessandro ; Conti, Mauro

DOI
10.1109/EuroSPW61312.2024.00047

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)

2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&amp;PW) | 979-8-3503-6729-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/EUROSPW61312.2024.00047

Work-in-Progress:
Crash Course: Can (Under Attack) Autonomous Driving Beat Human Drivers?

Francesco Marchiori
University of Padua
Padua, Italy
francesco.marchiori.4@phd.unipd.it

Alessandro Brighente
University of Padua
Padua, Italy
alessandro.brighente@unipd.it

Mauro Conti
University of Padua
Padua, Italy
Delft University of Technology
Delft, Netherlands
mauro.conti@unipd.it

***Abstract*—Autonomous driving is a research direction that has gained enormous traction in the last few years thanks to advancements in Artificial Intelligence (AI). Depending on the level of independence from the human driver, several studies show that Autonomous Vehicles (AVs) can reduce the number of on-road crashes and decrease overall fuel emissions by improving efficiency. However, security research on this topic is mixed and presents some gaps. On one hand, these studies often neglect the intrinsic vulnerabilities of AI algorithms, which are known to compromise the security of these systems. On the other, the most prevalent attacks towards AI rely on unrealistic assumptions, such as access to the model parameters or the training dataset. As such, it is unclear if autonomous driving can still claim several advantages over human driving in real-world applications.**

This paper evaluates the inherent risks in autonomous driving by examining the current landscape of AVs and establishing a pragmatic threat model. Through our analysis, we develop specific claims highlighting the delicate balance between the advantages of AVs and potential security challenges in real-world scenarios. Our evaluation serves as a foundation for providing essential takeaway messages, guiding both researchers and practitioners at various stages of the automation pipeline. In doing so, we contribute valuable insights to advance the discourse on the security and viability of autonomous driving in real-world applications.

1. Introduction

Autonomous Vehicles (AVs), also known as self-driving or driverless vehicles, refer to vehicles that do not need any form of human interaction to operate. An AV is capable of sensing the environment and recognizing the route, the roadway, and possible obstacles, as well as managing the speed and acceleration of the car. This is achieved through the cooperation of signals from various sensors, such as Global Positioning System (GPS), radars, Light Detection And Rangings (LiDARs), and thermographic cameras. While the final stage of AVs are still being developed and are not ready for mass production yet, the advantages that can come with them are clear: reduction of vehicle collisions, overall reduction of fuel consumption, increased comfort for passengers, and, most importantly, safer transport [26]. Given the complexity of such signals, data-driven approaches such as Machine Learning (ML) and Deep Learning (DL) must be used to elaborate all the information and decide the actions to perform in real time. However, this increased reliance on the technological components of the vehicle and the need for communication with its surroundings poses some security concerns. Indeed, while the cameras installed on the vehicle's various parts can implement a robust computer vision system, even a single error could precipitate collisions and crashes with far-reaching consequences. Processing these signals is often handled by ML or DL models, which, however, are inherently vulnerable to adversarial attacks [8]. One example is *evasion attacks*, in which the attacker can target the computer vision system to impose misclassification at test time. This attack can be applied to street road signs, a scenario in which the consequences can be profound and pose significant risks to public safety and traffic management systems [18].

Contribution. This paper delves into the question: *can autonomous driving systems outperform human drivers when subjected to adversarial attacks?* Traditionally, the focus has been on the efficiency and safety of AVs in ideal conditions. However, the emergence of evasion attacks [8], poisoning attacks [24], and other adversarial techniques raise critical concerns about the robustness of autonomous driving algorithms. This paper explores the concrete risks associated with autonomous driving under attack. By defining a realistic threat model and conducting comparative analyses between autonomous and human drivers, we seek to shed light on the capabilities and vulnerabilities of Artificial Intelligence (AI) systems. Furthermore, we scrutinize the current state-of-the-art to highlight and evaluate the assumptions on attacks towards AVs. Our findings have implications for the technological advancement of autonomous driving and the broader discourse on integrating AI into safety-critical applications.

Our contributions can be summarized as follows.

- We evaluate the vulnerabilities of all levels of automation for AVs. We highlight possible threats in real-world scenarios by formalizing the contribution of AI vs. human drivers in each of these levels.
- We delineate a realistic threat model in the context of attacks to AVs. We identify differences between common assumptions in adversarial attacks and realistic knowledge of a possible attacker.
- We perform a literature review on attacks to AVs and discuss their requirements to be successful.

DOI 10.1109/EuroSPW61312.2024.00047

We identify four common requirements and discuss their feasibility in real-world scenarios.
- We provide a list of security requirements and suggestions for safe and secure implementation of AI systems in AVs.

Organization. The paper is organized as follows. In Section 2, we give some background on AVs. Section 3 delves into the current literature and highlights the assumptions of related works on attacks towards AVs. Section 4 proposes a realistic threat model in the current landscape based on this discussion. In Section 5, we identify the criteria for our evaluation, and in Section 6, we analyze the vulnerabilities of each level of automation. In Section 7, we distill the takeaway messages from our study, and Section 8 concludes this work.

2. Background

The Society of Automotive Engineers (SAE) defines six levels of automation [10], [21]: starting from level 0, where the vehicle only provides warning messages and momentary assistance, we reach level 5, where the human driver is not needed in any road condition. We show an overview in Table 1. We can divide SAE levels into two significant groups. In the first, the vehicle provides driver support features such as automatic emergency braking (level 0), lane centering (level 1), and both lane centering and adaptive cruise control (level 2). In the second group, the vehicle is equipped with automated driving features such as traffic jam chauffeur (level 3), local driverless taxis in geofenced areas (level 4), and driverless vehicles in all regions (level 5). Furthermore, with levels 4 and 5, pedals and steering wheels may not be installed, preventing any traditional control from the driver to the vehicle.

To achieve these capabilities, an autonomous driving system includes the sensing, perception, planning, and actuation layers [11]. The sensing layer leverages technologies such as LiDAR, cameras, radar, GPS, and Inertial Measurement Unit (IMU) to acquire information from the surrounding environment. The perception layer then leverages the collected data by fusing and interpreting them to comprehend the vehicle's surroundings. The planning layer then leverages the perceived state to design a routing plan for the vehicle. The general approach is to generate a global trajectory with intermediate waypoints and adjust it in real-time to reach the destination safely. Lastly, the actuation layer defines a concrete motion plan, including acceleration and deceleration patterns and steering wheel angles.

Most of the applications of ML in autonomous driving are related to the perception and planning layers [20], [27]. The images collected via cameras need to be processed to detect lanes, traffic signs, and traffic lights and interpret them. Furthermore, ML plays a fundamental role in detecting pedestrians and other moving objects that may impact road safety. Sensor fusion algorithms leverage ML to extract meaningful information from different sensing sources. The planning layer makes an even more extensive use of ML. Indeed, algorithms for path searching, path planning, cooperative decision-making, collision avoidance, and decision generation all leverage ML mostly with deep, recurrent, and generative structures [20]. Lastly, ML has mainly been employed to develop intrusion and anomaly detection at the previously described layers.

3. Related Works

In the literature, it is possible to find different kinds of adversarial attacks able to fool almost any ML model. The two most prevalent and effective attacks are *evasion attacks* and *poisoning attacks*. Evasion attacks consist of carefully crafted noise to apply to samples at test time to drive the model to misclassification [8]. The noise is created by accessing the model parameters and is scaled to avoid human detection. Poisoning attacks involve injecting malicious data into the training set to manipulate model behavior during testing [24]. By tampering with training samples, attackers aim to influence the model's decision boundaries, often with subtle alterations to evade detection by human observers. However, in some adversarial experimental settings, datasets are too simple and fail to accurately represent realistic scenarios in which DL models could be deployed. Other studies propose attacks that can fool the state-of-the-art with high attack success rates but require extensive knowledge of the dataset or the model parameters to be effective.

Table 2 summarizes the current state-of-the-art attacks towards AVs and the assumptions and requirements needed to succeed. Here, we focus on papers on evasion attacks that focus on AV applications or mention and test their methodology on AV tasks. We decide not to include poisoning attacks since this technique requires access to the training dataset. Indeed, in real-world scenarios, automotive companies will be required to produce their datasets and manually validate them to avoid those threats. We identify four requirements and assumptions that are commonly made in the literature.

- **Model Parameters** – When targeting models in white-box scenarios, it is common to use techniques such as Fast Gradient Sign Method (FGSM) [8] or Basic Iterative Method (BIM) [14] to craft adversarial samples. However, these techniques require access to the models' gradients as it is needed in their optimization process. As such, a real-world attacker aiming to compromise a DL system with these methods would require complete access to the models, an assumption that is often unrealistic in many current scenarios.
- **Model Output** – Other adversarial attack techniques work in supposed black-box scenarios, i.e., do not require access to models' parameters to generate adversarial examples [2], [9]. As such, the assumptions on the attacker's knowledge are far reduced, as they would require only access to the system's output. Indeed, several of these techniques query the target model in their adversarial sample generation process. In real-world scenarios, this assumption might require additional effort from the attacker, as they would need to own the target vehicle's image recognition model to be able to query it.
- **Direct Input** – The most common adversarial attacks toward image recognition systems use imperceptible perturbations computed through the

TABLE 1. SAE Levels of Automation and Involvement of AI.

Level	Automation	Example Features	AI	Driver	Example Tasks
0	-	-	○	●	-
1	Partial Assistance	Adaptive Cruise Control (ACC)	●	●	Decision making
		Lane departure warning	●	●	Detection, sensor fusion
2	Partial Automation	ACC	●	●	Decision making
		Lane keeping assistance	●	●	Detection, sensor fusion
		Driver monitoring	●	●	Biometrics analysis
		Traffic jam assistant	●	●	Traffic pattern recognition
3	Conditional Automation	Environment monitoring	●	○	Sensor fusion
		Traffic jam autopilot	●	◐	Autonomous decision making
		Driver disengagement	●	◐	Autonomous decision making
		Autonomous driving	●	◐	Lane change, navigation
4	High Automation	Navigation in geofenced areas	●	○	Path planning
		Autonomous decision making	●	○	Traffic management
		Safety overrides	●	◐	Limited safety-critical tasks
5	Full Automation	Safety and redundancy	●	○	Anomaly detection
		V2X communications	●	○	Resource optimization
		Navigation	●	○	Autonomous navigation

●: present, ○: not present, ◐: partially present.

model's gradient to cause misclassification. The adversarial noise is applied to each sample and then fed to the target model. However, applying these perturbations on physical objects, such as street signs, is particularly challenging, as it would require the complete re-print of the image. Thus, the only way to apply the adversarial noise at test time is to have control of the models' input. In this scenario, an attacker would need to modify the images captured by the cameras before they are fed to the model. While this assumption can be realistic if an attacker has gained control of the AV internal network (and thus carrying a man-in-the-middle attack), it relies on fundamental vulnerabilities of the vehicle bus.

- **Physical Implementation** – To solve the challenge posed by the *direct input* assumption, several papers have delved into the feasibility of adversarial attacks using patches to apply to the target object [4], [7]. While this requirement can be more realistic in most AV scenarios, the attacker must know the road on which the target vehicle will drive and carefully apply the patch, as environmental conditions can challenge the attack's success.

4. Threat Model

Based on the related works discussed in Section 3, we identify four key factors that differentiate the current literature from real-world attackers.

- **Limited Access to Model Architecture** – In many practical deployments, the detailed architecture of the autonomous driving model is proprietary and closely guarded by manufacturers or developers. Adversaries are unlikely to have unrestricted access to the intricate details of the model's structure, layers, and parameters.
- **Restricted Knowledge of the Training Data** – The datasets used to train autonomous driving models often consist of diverse real-world scenarios, making them extensive and complex. Realistically, adversaries would not possess exhaustive knowledge about the complete training dataset, limiting their ability to craft highly tailored attacks.
- **Constrained Sensor Data Manipulation** – Autonomous vehicles rely on sensors such as cameras, LiDAR, and radar to perceive their environment. Manipulating these sensor inputs in real-time poses a substantial challenge, as physical access to the vehicle and the ability to tamper with sensor hardware is a great barrier for potential attackers.
- **Environmental Variability** – Real-world driving conditions encompass many environments, weather conditions, and traffic scenarios. Adversaries attempting to launch successful attacks must contend with the inherent variability in the operational environment, making it challenging to devise universally effective exploits.

In the case of evasion attacks, it is worth noting that they can be transferable, i.e., they can be computed on surrogate models to attack another model. However, it has been shown that this solution is not optimal and is often outclassed by simple non-gradient-based manipulations of the input [1]. While poisoning and backdoor attacks also show a degree of transferability, they rely on access to the training dataset, which is limited [16].

5. Criteria

We now identify the criteria we use for discussing the security of each SAE level of automation. Indeed, depending on which tasks are outsourced to ML algorithms, four different properties determine the advantage of humans or AI in specific scenarios. The criteria are formulated through a synthesis of existing research in autonomous driving, machine learning, and adversarial attacks. As such, they are grounded in a thorough examination of the pertinent literature and expert consensus within the domain.

TABLE 2. Requirements for state-of-the-art evasion attacks in AV tasks.

Attack	Misclassification Task	Model Parameters	Model Output	Direct Input	Physical Implementation
Arnab et al. [3]	Semantic Segmentation	●	●	●	○
Brown et al. [4]	Road Sign	●	●	○	●
Cao et al. [5]	LiDAR	●	●	○	●
Cao et al. [6]	LiDAR	○	●	○	●
Eykholt et al. [7]	Road Sign	●	●	○	●
Kong et al. [12]	Road Sign	○	●	○	●
Kumar et al. [13]	Road Sign	○	●	●	○
Li et al. [15]	Road Sign	○	●	●	○
Ma et al. [17]	Object Tracking	●	●	○	●
Papernot et al. [19]	Road Sign	○	●	●	○
Sharma et al. [22]	Misbehavior Detection	○	●	●	○
Sitawarin et al. [23]	Road Sign	○	●	●	○
Xiang et al. [25]	LiDAR	●	●	●	○
Zhu et al. [28]	LiDAR	○	●	○	●

●: required, ○: not required.

- **Ease of Attack** – This criterion assesses the simplicity or difficulty of executing an adversarial attack on human and AI drivers. It considers factors such as the technical expertise required and the accessibility of attack methods.
- **Response Time** – Response time evaluates the speed and efficiency of human and AI drivers react to adversarial situations. It reflects the ability to make quick decisions and take appropriate actions in response to unforeseen challenges.
- **Recovery Time** – Recovery time measures how swiftly human and AI drivers can recover and resume normal operation following a successful adversarial attack. It determines the resilience and adaptability of each system to bounce back from disruptions.
- **Adaptability** – Adaptability examines how well human and AI drivers can adjust and learn from adversarial encounters. It considers the capacity to evolve and improve responses, adapting to new attack strategies.

6. Evaluation

We now discuss and evaluate the levels of AV automation. We divide our discussion into two parts. First, we focus on each SAE automation level and discuss the feasibility of attack w.r.t. our threat model (Section 6.1). Then, we focus on our identified criteria and determine the advantages of AI or human drivers for each of them (Section 6.2).

6.1. Evaluation on Threat Model

We now analyze each level of the SAE scale of automation, and by contextualizing them on the threat model, we provide insights into their advantages over human driving.

Level 1 - Partial Assistance. Despite the limited automation, AVs at this level can provide valuable assistance in specific driving tasks, such as steering or acceleration (but not simultaneously). As such, evasion attacks can occur. However, the limitations of the threat model might not make these attacks feasible, given restricted access to the model architecture or training data. Incremental safety improvements and potential fuel efficiency benefits suggest Level 1 AVs may offer advantages over human drivers in specific scenarios but still rely on their judgment.

Level 2 - Partial Automation. Similarly to the previous level, AVs in this category help to steer and accelerate but can do that simultaneously. Thus, the same takeaways from the previous level apply, with different risks involved in the interaction between the assisted actions. However, without access to the training data, it is unlikely to build hidden triggers in the model to cause harm.

Level 3 - Conditional Automation. AVs at this level automate several driving phases but still require driver attention to take over if prompted. As such, potential challenges might occur during the handover (i.e., the transfer of control from the automated system back to the human driver). Indeed, as the number of automated processes increases, the attacker's limitations on sensor data manipulations become fewer since it grants the attacker multiple vectors for causing safety issues. However, the limited access to the model architecture makes attack attempts easier to detect, and consequently, human drivers can take over.

Level 4 - High Automation. Since drivers may not need to intervene at this level, exploiting limitations beyond predefined scenarios is risky. At this stage, it becomes imperative that the attacker cannot access the model parameters or tamper with the dataset. If those measures are ensured, level 4 AVs present enhanced safety and energy-efficient operation within specified contexts.

Level 5 - Full Automation. Comprehensive safety benefits, enhanced fuel efficiency, and increased accessibility suggest level 5 AVs may outclass humans in a broader range of driving scenarios. However, this is also the most challenging scenario, as it challenges system reliability and ethical decision-making.

6.2. Evaluation on Criteria

Instead of focusing on each SAE automation level, we now highlight each criterion's impact at the varying number of automated components in AVs. A summary of our analysis is shown in Table 3.

Ease of Attack. Assuming the attacker's knowledge is still limited according to our system model, attacking AVs becomes more feasible as the level of automation increases. Indeed, at level 1 or 2, malicious actors are limited to the restricted tasks assigned to AI algorithms. As such, attacks need to be specifically targeted towards specific models. Given the increased number of automated components at higher automation levels, there is a higher chance of disrupting the systems with generic and black-box adversarial attacks. Thus, while it might still be challenging to cause targeted misclassification, performing a Denial of Service (DoS) attack becomes easier.

Response Time. In unconstrained situations where the attacker has unlimited knowledge of the system, AVs would become more vulnerable as the level of automation increases since attacks cannot be detected. As such, they would rely on human take-over (available only in levels 1, 2, and 3) to promptly respond. However, given the restricted assumptions allowed by our realistic threat model, attack attempts are not as successful as they would be. Therefore, there is the possibility of implementing Anomaly Detection Systems (ADSs) to anticipate the presence of an attacker and thus react accordingly. Once an attack is detected, human take-over can be prompted if present, or other safety measures can be employed to prevent its effect.

Recovery Time. If an attack is successful, an increased level of automation might be more effective for mitigating its effect. The restricted automated capabilities of AVs in the first levels prevent them from having sufficient aptitude and adaptability to the environment. Therefore, it might be more challenging to resume normal operations, as external conditions might have drastically changed. Instead, at high levels of automation, an AV is expected to operate in any environment, as human drivers are often not present in the system model. They would thus be able to recover from the attack and mitigate it more promptly.

Adaptability. At lower levels of automation (1 and 2), the adaptability of AVs is limited, relying heavily on human intervention for decision-making in complex scenarios. In these situations, the adaptability of human drivers surpasses that of the automated systems. However, as automation levels progress, especially level 3 and beyond, the AV's ability to adapt to various attack scenarios can increase. Indeed, while there may be challenges during the handover phase in level 3, the overall adaptability rises, leveraging automated processes. At higher levels (4 and 5), where driver intervention may not be needed, AVs can exhibit advanced adaptability to diverse adversarial situations, continuously learning and improving responses over time. This also relates to the response time considerations, as the constrained threat model challenges the attack effectiveness and thus allows the system to not only prevent them but also learn from them through techniques such as adversarial training.

7. Takeaways

In this section, we summarize the main takeaway messages that we determine from our discussion.

TABLE 3. CONTRIBUTION OF AI IN AVS FOR SAFETY UNDER DIFFERENT CRITERIA FOR ALL SAE AUTOMATION LEVELS.

Level	Ease of Attack	Response Time	Recovery Time	Adaptability
1	●	◐	○	○
2	●	◐	○	○
3	◐	●	◐	◐
4	○	●	●	●
5	○	●	●	●

●: increased safety.
◐: unclear.
○: no improvement or decreased safety.

Takeaway 1: *Closed model architectures mitigate several adversarial threats across all levels of AV automation.*

As demonstrated by [1], access to the model architecture is the most critical aspect to consider when defending against adversarial attacks. Although adhering to the "security by obscurity" paradigm, prevalent in the automotive industry, restricting access to these resources is the best solution until effective defenses against ML attacks are developed.

Takeaway 2: *A well-defined Operational Design Domain (ODD) is a cornerstone of AV security while progressing towards full automation.*

As seen while increasing the level of automation in Section 6, the definition of the ODD is fundamental for ensuring safety. Setting boundaries on the automated tasks and formalizing the human intervention can mitigate risks while climbing towards the complete automation of AVs.

Takeaway 3: *Securing autonomous driving demands collective standards and innovation, considering the risks of a realistic threat model while prioritizing advantages over human driving.*

Securing a trustworthy future demands collaboration among industries, academia, and regulators in the dynamic landscape of autonomous driving. Establishing unified standards for security, transparency, and ethics is essential. As such, the balance between innovation and resilience defines the path toward a secure autonomous landscape.

8. Conclusions

The literature surrounding AVs highlights their crucial role in enhancing road safety by reducing collisions. With precise environmental sensing, these vehicles offer the potential to revolutionize safety. Additionally, studies suggest AVs could contribute to a more sustainable transportation landscape by lowering fuel emissions.

Contribution. In this paper, we evaluated the risks of autonomous driving and established a pragmatic threat model highlighting the delicate balance between AV advantages and security challenges. We explored whether autonomous driving systems can outperform human drivers under adversarial attacks, particularly considering evasion attacks. Our vulnerability evaluation comprised all SAE automation levels in light of the realistic threat model.

Furthermore, we reviewed the AV attack literature and suggested security measures for AI implementation in AVs. Our contributions aim to advance discourse and technological development while ensuring the safe integration of AI into safety-critical operations.

Future Works. In the ongoing pursuit of securing autonomous driving, future research should refine threat models and innovate defense mechanisms. A crucial direction is assessing the adaptability of AI systems to emerging adversarial strategies and improving the resilience of both human and AI drivers against evolving threats. Considering the diverse levels of SAE automation in AVs, developing targeted countermeasures becomes imperative for a more robust autonomous driving landscape. Furthermore, an essential future work direction involves evaluating the criteria identified on a practical testbed. This testbed could simulate real-world conditions, allowing for the empirical validation of the criteria's efficacy in assessing the security of autonomous driving systems against adversarial challenges. Addressing feasibility and practicality remains central to advancing the field.

References

[1] Marco Alecci, Mauro Conti, Francesco Marchiori, Luca Martinelli, and Luca Pajola. Your attack is too dumb: Formalizing attacker scenarios for adversarial transferability. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 315–329, 2023.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020.

[3] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 888–897, 2018.

[4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[5] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2267–2281, 2019.

[6] Yulong Cao, Chaowei Xiao, Dawei Yang, Jing Fang, Ruigang Yang, Mingyan Liu, and Bo Li. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*, 2019.

[7] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[9] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

[10] ISO/SAE PAS 22736:2021 Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard, International Organization for Standardization, August 2021.

[11] Seulbae Kim, Major Liu, Junghwan" John" Rhee, Yuseok Jeon, Yonghwi Kwon, and Chung Hwan Kim. Drivefuzz: Discovering autonomous driving bugs through driving quality-guided fuzzing. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1753–1767, 2022.

[12] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14254–14263, 2020.

[13] K Naveen Kumar, C Vishnu, Reshmi Mitra, and C Krishna Mohan. Black-box adversarial attacks in autonomous vehicle technology. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2020.

[14] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[15] Yujie Li, Xing Xu, Jinhui Xiao, Siyuan Li, and Heng Tao Shen. Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. *IEEE Internet of Things Journal*, 8(8):6337–6347, 2020.

[16] Yiyong Liu, Michael Backes, and Xiao Zhang. Transferable availability poisoning attacks. *arXiv preprint arXiv:2310.05141*, 2023.

[17] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. Wip: Towards the practicality of the adversarial attack on object tracking in autonomous driving. In *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*, 2023.

[18] Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, and Yuval Weisglass. Fooling a real car with adversarial traffic signs. *arXiv preprint arXiv:1907.00374*, 2019.

[19] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[20] Zi Peng, Jinqiu Yang, Tse-Hsun Chen, and Lei Ma. A first look at the integration of machine learning models in complex autonomous driving systems: a case study on apollo. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1240–1250, 2020.

[21] SAE J3016 Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Standard, Societey of Automotive Engineers, April 2021.

[22] Prinkle Sharma, David Austin, and Hong Liu. Attacks on machine learning: Adversarial examples in connected and autonomous vehicles. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–7. IEEE, 2019.

[23] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018.

[24] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.

[25] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019.

[26] J Yang and Joseph F Coughlin. In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers. *International Journal of Automotive Technology*, 15:333–340, 2014.

[27] Jingyuan Zhao, Wenyi Zhao, Bo Deng, Zhenghong Wang, Feng Zhang, Wenxiang Zheng, Wanke Cao, Jinrui Nan, Yubo Lian, and Andrew F Burke. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, page 122836, 2023.

[28] Yi Zhu, Chenglin Miao, Foad Hajiaghajani, Mengdi Huai, Lu Su, and Chunming Qiao. Adversarial attacks against lidar semantic segmentation in autonomous driving. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 329–342, 2021.