Transparency of Autonomous Driving Algorithms: Finding an Optimal Balance Between Developer Security and Customer Safety

Dolgopolyi Roman

The American College of Greece

r.dolgopolyi@acg.edu

Abstract. This paper examines the tension between secrecy of autonomous vehicles (AVs) algorithms and the need for transparency to safeguard consumer interests. Observing AV developments from early government-funded prototypes to breakthroughs in deep neural networks, it shows that increased algorithmic complexity has limited the ability of the data processing's assessment from a developer and a user perspectives. Ethical theories, applied in this paper, particularly Act Utilitarianism and Ethical Egoism, provided theoretical foundation for transparency claims of the public. Utilized Legal frameworks, including the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in California, emphasized existing gap between legal requirements and characteristics of deployed algorithms worldwide. The paper concludes by emphasizing that a new approach on AV algorithms transparency has to be adopted. One that would preserve the technical secrecy essential for security of intellectual property and cybersecurity posture, while guaranteeing required transparency to ensure meaningful consumer safety and trust.

Keywords: Autonomous vehicles, cybersecurity, technological transparency, big data, artificial intelligence, neural networks.

Table of Contents

1 Introduction

AV (autonomous vehicle) is characterized by being equipped with technologies that partly or fully replace the need for a human operator. Already more than 50% of produced vehicles in major East Asian markets incorporate autonomous driving technologies, and projections indicate similar or higher adoption rates in other major markets, such as the US, and West and North Europe [1].

Despite this increase in production and demand, the general public has limited awareness of specific technologies and algorithms that power these vehicles. Surveys reveal that substantial number of people primarily rely on marketing claims rather than objective knowledge about AV's technologies [2]. This state of affairs leads to dangerous technological over-reliance and overestimation of safety. Moreover, major AV developers rarely release important details on how their vehicles make critical decisions, especially in complex and dangerous real-world driving scenarios that involve life-and-death dilemmas [3].

From the developer's perspective, such secrecy is understandable. To guard proprietary technology from malicious hacking or theft, companies refrain from disclosing their algorithms' or decision-making logic [4]. Yet, withholding this knowledge leaves consumers with undermined personal safety. Consumers become unaware of the precise risk-benefit trade-offs that they make while deciding to acquire or rent an AV.

To address this conflict of interests, the paper explores the balance between algorithmic transparency in autonomous vehicles, focusing on both the developer's security and customer safety. Subsequent sections outline the development of autonomous driving technology to better understand the evolution of related security dangers and highlight the industry's shift towards complex artificial intelligence algorithms that are in the center of most public safety concerns of today. Next, the identified problematic areas will be examined through ethical theories and legal frameworks to offer an unbiased assessment of the justification behind the industry's current approach to the discussed issues.

2 Autonomous vehicles of the past (1960–2000s)

2.1 First appearance and evolution of the technology

The concept of AVs evolved alongside advances in computer science. During the 1960s, first in its kind projects such as Stanford's "Lunar Vehicle Remote Control" and NASA's "Roving Vehicle Motion Control" showcased never seen before autonomy features like line tracking and surface identification [5], [6]. Since the available hardware at that time was computationally insufficient for real-world applications in everyday driving scenarios, these projects focused exclusively on highly controlled or predictable environments such as research labs or obstacle-free surroundings.

Until the mid-1980s, AV research was mostly government-driven, with agencies such as NASA and DARPA supporting projects for strategic benefits [7]. As sensing hardware (machine-vision cameras, radar, and infrared sensors) improved, possibilities for real-world, civilian applications expanded. The turning point came in 1987 with the PROMETHEUS project (Programme for a European Traffic of Highest Efficiency and Unprecedented Safety) [8]. This initiative attracted major automakers like Mercedes-Benz, BMW, and Fiat, which recognized AV technology's commercial potential. PROMETHEUS culminated in a self-driving vehicle capable of intercity travel with minimal human intervention, effectively resolving the gap between research prototypes and consumer-ready applications.

Building on breakthroughs from such initiatives, auto manufacturers began incorporating computer systems with their partial autonomy features into commercial cars by the 1990s and early 2000s [9]. Capabilities like adaptive cruise control, lane-keeping assistance, and early collision-avoidance systems were sold as safety features rather than self-driving technology. Consumers generally viewed these incremental enhancements with optimism [10]. Nonetheless, as vehicles became more computerized, cybersecurity threats started to be a frequent discussion [11].

2.2 Related security and privacy issues

Since vehicles started to rely on computer systems for brakes, engines, and steering, black-hat hackers gained a possibility to identify and exploit potential security and data breaches [12]. Conferences like ESCAR (Embedded Security in Cars) appeared in the early 2000s to address these threats for the first time [13]. At such conferences, many automakers faced criticism for relying only on a "security through obscurity" stance, limiting public information to safeguard the systems [14]. This contradicted well established cybersecurity principles like Kerckhoffs's principle, which states that systems should remain robust even if their internal structure is known [15].

Though experts raised concerns, manufacturers continued to classify information regarding proprietary algorithms for intellectual property reasons while integrating better cybersecurity practices such as cryptography and secure software development framework (SSDF) [16]. In most instances, customers were unconcerned by the implied secrecy since the autonomous features were easily understood and predictable in terms of logic and expected behavior (e.g., adaptive cruise control - maintains a distance of 4 meters to the next vehicle; lane tracking - adjusts the steering angle to keep the vehicle in the center of a line). This combination of user-friendly communication and consistent performance effectively maintained consumer trust without a need for disclosure of the algorithms' internal architecture or logic [17].

3 Autonomous vehicles of today (2010–2020s)

3.1 New range of capabilities

While automakers and researchers were incorporating more advanced features with higher levels of autonomy, confusion about the varying degrees of driver assistance started to grow. For this reason, the Society of Automotive Engineers (SAE) published the J3016 standard in 2014. This classification outlines six autonomy levels, from 0 (no automation) to 5 (full automation) [18]. Most commercially available models at the time fell under Levels 0–2, but SAE included higher levels to anticipate rapid technological progress.

Table 1. SAE J3016 Levels of Driving Automation

Level	*Name*	*Driver Role*
0	No Automation	Full control
1	Driver Assistance	Driver must remain fully engaged
2	Partial Automation	Driver monitors and may intervene
3	Conditional Automation	Must be ready to take control
4	High Automation	No intervention in defined scenarios
5	Full Automation	No human intervention required

That progress indeed happened and was mainly driven by breakthroughs in artificial intelligence, characterized by advances in deep neural networks (DNNs) [19]. DNNs allowed for improvement in computer vision tasks by learning from data rather than relying solely on human programming [20]. This approach proved to be ideal for unpredictable road conditions that AVs confront. Consequentially, many researchers and companies envisioned possibilities for self-driving cars surpassing the limitations of earlier rule-based algorithms [21].

Tesla's Model S, introduced with self-driving features in 2014, was among the first widely accessible Level 2 vehicles employing neural networks for tasks like lane changes and highway merges. Its system architecture illustrated the real-world potential of DNN approach for driving [22]. Using that, the car's internal algorithm acquired its rules from large-scale data inputs (geolocation, camera images, biometrics and behavioral patterns of drivers and passengers, and etc.), refining its performance over time until a point that was considered as optimum [23].

Following Tesla's lead, companies like Uber, Waymo, and Nvidia adopted similar data-driven methods. Nvidia's specialized AI platforms significantly advanced AVs' real-time decision-making, enabling vehicles to navigate complex pedestrian-rich environments [24]. Waymo launched driverless taxi services in cities such as Phoenix and Arizona, pushing towards final Level 5 of autonomy [25]. However, despite these impressive achievements, the growing reliance on DNN created a significant cybersecurity threat, the "black-box" system characterized by hardly assessable internal logic [26].

3.2 Related security and privacy issues

DNNs often mask how decisions are reached, offering visibility only into inputs and outputs. Therefore, even developers may struggle to describe if a vehicle stops at a red light because it truly understands traffic signals or because it associates any red color's source with an instant stop response [27]. Industry leaders overcame this difficulty with extensive simulation-based testing. By exposing AVs' algorithms to vast numbers of scenarios inside virtual environments, stakeholders can verify safety outcomes without delving deeply into the algorithm [24].

While this approach addresses engineering risks, it leaves customers uncertain about how AVs will behave in novel situations. The previous reliance on "security through obscurity" is inadequate here. Customers have no more clear guidelines to rely on to understand a detailed picture of how the data collected real-time from their driving experience (geolocation, camera images, biometrics, behavioral patterns, etc.) is being processed by an AV's algorithm [23, 27]. Moreover, customers lack a clear reference standard for the evaluation of this algorithm's outputs—namely, the AV's driving behavior—and thus cannot readily determine which actions are normal versus malfunctioning [27].

With growing public distrust of AV technology, which has risen from 54% to 66% in the last three years, the number of protests and acts of activism requesting more transparency and better governance of the technology has climbed as well [28, 29]. Examples of these protests range from highly violent, with cases where AVs are intentionally damaged and turned into non-repairable vehicles, to less violent that intend to showcase security flaws without causing significant damage. The "Week of Cone" organized by SSR (Safe Street Rebels) in 2024 in California, is an example of a less violent protest, with the main tactic being to showcase that AVs algorithms cannot handle unpredictable situations on roads, such as the appearance of a traffic cone in minimal distance to a camera sensor [30].

Given the preceding discussion, It becomes increasingly clear that preserving secrecy can no longer address security threats and safety issues around modern AV technology. Therefore, ethical, and legal frameworks, which would be introduced in the next section, should be explored to find the balance between proprietary algorithms' transparency and consumer safety.

4 Theoretical Framework

4.1 Ethical Theories

Act Utilitarianism

. To be able to assess the transparency-security tradeoff from the perspective of how it affects large and diverse groups of people, we would need to apply Act Utilitarianism. This ethical theory was developed by Jeremy Bentham in the late 18th century with the purpose of making the assessment of an act's morality straightforward and logical. Therefore, the utilitarian logic constitutes that an act is morally permissible if the consequences that result from it produce the greatest amount of good for the greatest number of people affected by it [31].

Ethical Egoism

. To inspect how decisions regarding the secrecy of AV algorithms might be evaluated if guided by self-interest, the Ethical Egoism theory is particularly useful. Shaped by Ayn Rand in the 20th century, Ethical Egoism maintains that an act is morally right if it servs self-interests of the agent itself, without consideration of its impact on others. Accordingly, it judges actions primarily by how they benefit the individual or entity in the act, rather than by broader social outcomes or universal duties [32].

4.2 Legal Frameworks

European Legislation - General Data Protection Regulation (GDPR)
. The General Data Protection Regulation (GDPR), introduced in 2018, is currently the main legal framework governing how personal data is collected, processed, stored, and distributed in the European Union [33]. It automatically applies to any entity that works with data within the EU's borders or to any entity globally if that entity works with the EU's citizens' data. This framework is an important legal instrument for the assessment of the discussed AV's algorithms, since the collection and processing of data is the central part of these algorithms' operations.

US Legislation - California Consumer Privacy Act (CCPA)
. Across the United States, regulations regarding the data privacy significantly vary depending on the particular state. The California Consumer Privacy Act (CCPA), which is active from 2018 and only under California state's jurisdiction, is one of the most comprehensive and up to date (amended in 2023) examples of legislation frameworks that focuses on safeguarding personal information, disclosing data practices, and data ownership's rights [34]. When applied to practices of AVs' development and deployment, it may provide important insights how most modern attempts of data-focused regulations address introduced issues.

5 Analysis

5.1 Ethical Theories

Act Utilitarianism
. As previously stated, the current approach of "security through obscurity" in the AV industry caused a significant level of public distrust and civil disobedience [28, 29, 30]. Conversely, there is no known public campaign that would have requested the preservation of the current state of AV algorithms' transparency. Therefore, by applying Act Utilitarianism that evaluates an act as morally permissible based on its causation of greatest amount of good for the greatest amount of people, we may conclude that the act should be considered as unethical, since the developers have failed to establish the transparency level that was requested by the public [31].

Ethical Egoism
. Under Ethical Egoism, developers' decisions to keep AV algorithms classified appear morally justifiable from a self-interests perspective. Secrecy protects intellectual property and preserves a competitive advantage [4]. However, these gains are likely to be short-lived. As consumer mistrust grows, AV developer companies risk losing their

market value. As an illustration, Tesla's stock price reportedly declined by 36% over the past three years, while the number of people who are afraid of AV technologies increased by more than 10% [28]. Thus, while Ethical Egoism supports secrecy of AV algorithms and evaluates it as ethical in the short term, it may fail to justify the developers' decisions if potential long-term consequences that prevent the satisfaction of self-interest in financial and developmental growth are taken into consideration [32].

5.2 Legal Frameworks

European Legislation - General Data Protection Regulation (GDPR)

. The GDPR states that any personal data processing must be both minimal and transparent, as specified by Article 5(1)(c), which indicates that personal data "shall be adequate, relevant and limited to what is necessary." Recital 39 also highlights that "the principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand" [33]. However, these requirements appear conflicting with current practices in AV development, where much of the decision-making logic is under the curtain, and "black-box" algorithms leave consumers and even developers unclear about how specifically collected data informs critical driving decisions [23, 27].

US Legislation - California Consumer Privacy Act (CCPA)

. The CCPA in Section 7002 requires that "the business's collection, use, retention, and/or sharing of a consumer's personal information shall be reasonably necessary and proportionate to achieve the purpose(s) for which it was collected or processed" [34]. Yet the reliance on highly effective but highly unassessable DNN algorithms in AVs poses a direct confrontation to that principle, since it is unclear whether all of the parts of large-scale real-time data collection and processing are necessary for efficient automated decision-making [23, 27]. Furthermore, Section 7003 states that "disclosures and communications to consumers shall be easy to read and understandable to consumers" [34]. However, as we already concluded while applying the GDPR framework, transparency requirements, because of the "black-box" nature of DNN algorithms, could not be met at the current stage of the technology development [27].

6 Conclusion

Throughout the research, the issues of the trade-off between AV developers' security and customers safety, based on the transparency of the algorithms, were illuminated from different ethical and legal perspectives. Thus, it was concluded that the Act Utilitarian Ethics would consider the current state of the algorithms' transparency as unethical, since it raises negative reactions and public protests. The application of the Ethical Egoism showcased that conclusions could be two-sided based on the short- or long-term evaluation of keeping the algorithms classified. Short-term outcomes, under the Ethical Egoism, prove the secrecy of the algorithms as ethical since they serve self-interests of the developers to keep a competitive advantage. However, the long-term consequences of the classified approach are proving it to be unethical since the developers are likely to lose profitability with time, as public distrust in the technology increases. The application of two legal frameworks (GDPR, CPPA) demonstrated that contradictions exist between the current transparency state of the algorithms and legal requirements. Both legislations emphasized that there are two main problematic areas. The first is the absence of clear proof from the developers' side that every collected and processed data type is necessary for efficient operation of the algorithms. The second is the developers' inability to clearly describe processing logic of used DNNs algorithms, because of their "black-box" nature.

Therefore, it is clear that a new approach on security and transparency of DNN-based AV algorithms should be implied. It has to meet two requirements to resolve the described conflict of interests. Firstly, the approach has to protect the developers from intellectual property theft as well as from over-exposure of critical information leading to security breaches. Secondly, public concerns and legal requirements should be addressed in the new approach by a guarantee of customers' ability to independently assess AVs' safety in terms of how collected data affects the final decisions of the algorithms.

Funding: This research received no external funding.
Clinical trial number: not applicable.

References

[1] P. Bansal and K. M. Kockelman, “Forecasting Americans’ long-term adoption of connected and autonomous vehicle technologies,” *Transportation Research Part A: Policy and Practice*, vol. 95, pp. 49–63, Jan. 2017, doi: https://doi.org/10.1016/j.tra.2016.10.013.

[2] L. Dixon, “Autonowashing: The Greenwashing of Vehicle Automation,” *Transportation Research Interdisciplinary Perspectives*, vol. 5, p. 100113, May 2020, doi: https://doi.org/10.1016/j.trip.2020.100113.

[3] T. Schneider *et al.*, “Don’t fail me! The Level 5 Autonomous Driving Information Dilemma regarding Transparency and User Experience,” *RADAR*, Mar. 2023, doi: https://doi.org/10.1145/3581641.3584085.

[4] M. J. Ryan, "Secret Algorithms, IP Rights, and the Public Interest," *Nevada Law Journal*, vol. 21, pp. 61–92, 2020. [Online]. Available: https://scholars.law.unlv.edu/nlj/vol21/iss1/3

[5] Stanford University Department of Mechanical Engineering, *Lunar Vehicle Remote Control: A Study by Stanford University Mechanical Engineering Design Division—Outtakes* [Film], USA: Stanford (Calif.), 1966.

[6] B. P. Miller, T. M. Corry, D. E. Johnson, R. J. Johnston, and J. E. Lingerfelt, "Roving Vehicle Motion Control: First Quarterly Report Covering the Period 1 March 1967 Through 31 May 1967," AC Electronics – Defense Research Laboratories, General Motors Corporation, Santa Barbara, CA, USA, Tech. Rep. TR67-34, Jun. 1967. [Online]. Available: https://ntrs.nasa.gov/api/citations/19670026750/downloads/19670026750.pdf

[7] A. Roland and P. Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*. Cambridge, MA, USA: MIT Press, 2002.

[8] E. D. Dickmanns, "The development of machine vision for road vehicles in the last decade," *Intelligent Vehicle Symposium, 2002. IEEE*, Versailles, France, 2002, pp. 268-281 vol.1, doi: 10.1109/IVS.2002.1187962.

[9] R. Bishop, "Intelligent vehicle applications worldwide," in *IEEE Intelligent Systems and their Applications*, vol. 15, no. 1, pp. 78-81, Jan.-Feb. 2000, doi: 10.1109/5254.820333.

[10] A. Vahidi and A. Eskandarian, "Research advances in intelligent collision avoidance and adaptive cruise control," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 3, pp. 143-153, Sept. 2003, doi: 10.1109/TITS.2003.821292.

[11] Checkoway, S., McCoy, D., Kantor, B., Anderson, D., Shacham, H., Savage, S., Koscher, K., Czeskis, A., Roesner, F., & Kohno, T, "Comprehensive Experimental Analyses of Automotive Attack Surfaces" in *USENIX Security Symposium,* 2011, doi: 10.5555/2028067.2028073

[12] Kim, S., Shrestha, R. (2020). Introduction to Automotive Cybersecurity . In: *Automotive Cyber Security. Springer, Singapore.* https://doi.org/10.1007/978-981-15-8053-6_1

[13] “History,” *Escar.info*, 2024. https://escar.info/escar-europe/history [Accessed: Apr. 1, 2025].

[14] S. Bittl, “Attack Potential and Efficient Security Enhancement of Automotive Bus Networks Using Short MACs with Rapid Key Change,” *Lecture notes in computer science*, pp. 113–125, Jan. 2014, doi: https://doi.org/10.1007/978-3-319-06644-8_11.

[15] F. A. P. Petitcolas, “Kerckhoffs’ Principle,” *Encyclopedia of Cryptography and Security*, pp. 675–675, 2011, doi: https://doi.org/10.1007/978-1-4419-5906-5_487.

[16] M. Wolf, A. Weimerskirch, and T. Wollinger, “State of the Art: Embedding Security in Vehicles,” *EURASIP Journal on Embedded Systems*, vol. 2007, pp. 1–16, 2007, doi: https://doi.org/10.1155/2007/74706.

[17] J. R. Sayer, Mary Lynn Mefford, and P. S. Fancher, “Consumer Acceptance of Adaptive Cruise Control following Experience with a Prototype System,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 39, no. 17, pp. 1092–1096, Oct. 1995, doi: https://doi.org/10.1177/154193129503901706.

[18] SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE Standard J3016_202104, Revised Apr. 2021, Issued Jan. 2014, doi: https://doi.org/10.4271/J3016_202104

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[20] A. Karpathy, *Connecting images and natural language*, Ph.D. dissertation, Stanford Univ., Stanford, CA, 2016.

[21] M. Bojarski et al., "End to End Learning for Self-Driving Cars," NVIDIA Corporation, 2016, doi: https://doi.org/10.48550/arXiv.1604.07316

[22] B. Bauchwitz and M. Cummings, “Evaluating the Reliability of Tesla Model 3 Driver Assist Functions,” 2020. Available: https://www.roadsafety.unc.edu/wp-content/uploads/2020/11/R27_interim-deliverable_DMS_experiment_1_summary_CSCRS_report.pdf

[23] M. Harris, "The Radical Scope of Tesla's Data Hoard: Every Tesla is providing reams of sensitive data about its driver's life," in *IEEE Spectrum*, vol. 59, no. 10, pp. 40-45, October 2022, doi: 10.1109/MSPEC.2022.9915627

[24] NVIDIA Corporation, "NVIDIA Omniverse for Automotive Design and Engineering," Dell Technologies, White Paper, 2022. [Online]. Available: https://www.delltechnologies.com/asset/en-

ca/products/workstations/industry-market/nvidia-omniverse-for-automotive-design-and-engineering.pdf [Accessed: Apr. 2, 2025].

[25] Waymo, "Scaling Waymo One safely across four cities this year," Waymo Blog, Mar. 2024. [Online]. Available: https://waymo.com/blog/2024/03/scaling-waymo-one-safely-across-four-cities-this-year [Accessed: Apr. 2, 2025].

[26] M. Tegmark and S. Omohundro, “Provably safe systems: the only path to controllable AGI,” *arXiv.org*, 2023. https://arxiv.org/abs/2309.01933

[27] F. Utesch, A. Brandies, P. Pekezou Fouopi, and C. Schießl, “Towards behaviour based testing to understand the black box of autonomous cars,” *European Transport Research Review*, vol. 12, no. 1, Jul. 2020, doi: https://doi.org/10.1186/s12544-020-00438-2.

[28] L. Stewart, M. Musa, and N. Croce, “Look no hands: self-driving vehicles’ public trust problem,” *World Economic Forum*, Aug. 12, 2019. https://www.weforum.org/stories/2019/08/self-driving-vehicles-public-trust/.

[29] S. Nordhoff, “Resistance towards autonomous vehicles (AVs),” *Transportation Research Interdisciplinary Perspectives*, vol. 26, pp. 101117–101117, Jul. 2024, doi: https://doi.org/10.1016/j.trip.2024.101117.

[30] S. Hind, “Resisting Decisions: Coneheads in California,” *Driving Decisions*, pp. 225–264, 2024, doi: https://doi.org/10.1007/978-981-97-1749-1_8.

[31] J. Bentham, *An Introduction to the Principles of Morals and Legislation*, Oxford: Clarendon Press, 1789.

[32] A. Rand, *The Virtue of Selfishness: A New Concept of Egoism*, New York, NY, USA: Signet, 1964.

[33] European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, Official Journal of the European Union, L 119, pp. 1–88, May 4, 2016.

[34] State of California, *California Consumer Privacy Act of 2018 (CCPA), California Civil Code*, Sacramento, CA, USA: State of California, 2018.