Driving on Trust: The Moral Responsibility of Tesla in the March 2018 Autopilot Crash

STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Adam Khan

May 1, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR

Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

On March 23rd, 2018, a tragic incident involving a Tesla vehicle operating under its Autopilot system resulted in a fatality when the car collided with a highway divider at high speed without any driver intervention. This event has sparked significant debate regarding the ethical implications of Tesla's decision to develop, market, and deploy an autonomous driving system that is still technically in beta testing. Critics argue that Tesla's marketing of the Autopilot system, which suggests a level of autonomy and reliability that the technology does not fully achieve, has misled consumers, and exposed them to significant risks. This criticism is juxtaposed against defenders who assert that the onus of responsibility lies with the drivers, who are warned of the system's limitations in the vehicle's owner manual.

The discourse surrounding this incident and Tesla's Autopilot system reflects broader concerns about the ethical responsibilities of corporations in the age of emerging technologies. While the technological community has extensively debated the technical and safety aspects of autonomous driving systems, there is less discussion on the moral responsibilities of companies like Tesla in marketing such technologies. This gap in the discourse overlooks a crucial aspect of technological advancement: the ethical considerations in the portrayal and deployment of new innovations to consumers.

To determine the moral responsibility of the March 2018 Autopilot system incident, I will apply Ibo van de Poel's definition of passive responsibility from his book *Ethics, Technology, and Engineering: An Introduction*. This framework utilizes four conditions to assign moral responsibility: wrong-doing, causal contribution, foreseeability, and freedom of action (Van de Poel & Royakkers, 2011). I claim that Tesla is morally responsible for the fatality involving its Autopilot system to a significant extent. This responsibility arises from (1) the company's

overstatement of the system's capabilities, contributing to wrongful expectations of safety (wrong-doing), and (2) its failure to address known technological limitations, which directly contributed to the incident (causal contribution). These aspects were foreseeable by Tesla, and despite facing competitive pressures, the company had the freedom to act differently (foreseeability and freedom of action).

Literature Review

While literature on the Tesla Autopilot crash of March 2018 is limited, analyses on other crashes involving Tesla autopilot and ethics analyses on autonomous vehicles can be insightful.

In "Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016", the authors investigate the circumstances surrounding a fatal Tesla crash in May 2016. They argue that while the preliminary investigation by the National Highway Traffic Safety Administration concluded that the driver of the Tesla was at fault, the reason behind that driver error was actually "designer error", implicating the design of Tesla's Autopilot (Banks et al., 2018). The authors analyze the crash using the Perceptual Cycle Model, a cognitive theory proposed that suggests perception is a cyclic process involving an active interplay between cognitive schemas (expectations) and sensory input to make sense of the world. They claim that within the realm of autonomous driving systems, human error should be viewed as a symptom of safety breakdowns rather than being viewed as a cause. Lastly, the authors state that the original design of the Autopilot system set the driver up to fail because of the requirement for drivers to monitor a highly reliable system continually. This article is helpful for understanding the cognitive and perceptual dynamics

between human operators and autonomous vehicle systems, particularly in highlighting how design choices can influence user interaction and potentially lead to critical incidents.

In addressing the ethical implications of autonomous vehicle technology, it is crucial to consider the responsibility manufacturers bear in ensuring the safety and reliability of their products. As Alexander Hevelke and Julian Nida-Rümelin (2015) articulate, manufacturers are "ultimately responsible for the final product," which includes not just the vehicle itself but the sophisticated systems guiding it. If a system's design flaw or decision tends to precipitate accidents under certain conditions, it stands to reason that manufacturers were aware or should have been aware of such risks, raising questions about their ethical responsibility in selling these vehicles. However, the imposition of full liability on manufacturers could stifle innovation and development in autonomous vehicle technology due to the prohibitive cost of potential lawsuits. This delicate balance between promoting safety and encouraging technological advancement presents a complex ethical dilemma: how to structure liability in a way that fosters continuous improvement without unduly hampering progress (Hevelke & Nida-Rümelin, 2015).

The intersection between Banks et al. (2018) and Hevelke and Nida-Rümelin (2015) lies in the recognition that the responsibility for autonomous vehicle safety extends beyond the driver to the designers and manufacturers. Both sources converge on the idea that manufacturers must navigate the ethical responsibility of system design and user safety while fostering an environment conducive to technological advancements.

This paper extends the scholarly discourse by addressing a critical gap in the literature: the moral obligations of manufacturers like Tesla in the context of real-world incidents. While previous research has explored the cognitive aspects of autonomous system interaction and the ethical considerations of liability, this paper provides a focused analysis of the March 2018 crash

by applying Van de Poel's criteria of moral responsibility. By doing so, it highlights the specific ethical failings that led to this incident and argues for a model of responsibility that holds manufacturers accountable for not only the design and safety of their products but also for the clarity and honesty of their communication regarding technological capabilities.

Conceptual Framework

My analysis of the March 2018 Autopilot system incident draws on Ibo van de Poel's definition of passive moral responsibility. According to his book, the definition of passive responsibility is "Backward looking responsibility, relevant after something undesirable occurred; specific forms are accountability, blameworthiness, and liability." (Ibo van de Poel & Lamber Royakkers, 2011). Accountability is the sense of being held to account for one's actions. Blameworthiness is the sense of being a 'target of blame' for the consequences of one's actions. In order for someone to be blameworthy, and, therefore, morally responsible, the following conditions must apply:

- Wrong-doing
- Casual contribution
- Foreseeability
- Freedom of action

Figure 1: Van de Poel's criteria of responsibility.

Wrong-doing is defined as any action or inaction that breaches ethical, moral, legal, or societal standards, causing harm or injustice. The standard may be moral, legal, or organizational. Causal contribution is the involvement through action or inaction that is necessary for a particular outcome to occur, underscoring the role of both acts and omissions in the chain of events leading to that outcome. Foreseeability is defined as the ability to predict or anticipate the consequences of one's actions, particularly the potential for harm or damage, based

on the knowledge and understanding of the risks involved at the time. It implies that individuals or parties can only be held accountable for outcomes that were reasonably predictable given the information and circumstances available to them prior to the event. Freedom of action is defined as the ability of an individual to make choices and take actions based on their own will, without being subjected to external compulsion or coercion. It underlines the principle that individuals can only be held fully accountable for actions undertaken by their own free will. Each of these conditions offers valuable insights into different aspects of Tesla's responsibility in the incident.

My analysis has been structured to correlate directly with individual pieces of evidence, linking each to relevant criteria of moral responsibility as articulated by Van de Poel. This approach ensures a focused examination of Tesla's actions in relation to the Autopilot system. For each distinct piece of evidence, I will determine whether it demonstrates wrong-doing by Tesla, such as violating ethical or safety norms. Concurrently, I will analyze Tesla's causal contribution, scrutinizing how the company's decisions and the Autopilot system's capabilities may have influenced the incident's outcome. The foreseeability of the accident will be evaluated through the lens of each evidence piece, considering whether Tesla could have anticipated the risks associated with Autopilot's known limitations. Lastly, I will explore Tesla's freedom of action, assessing how competitive pressures may have shaped Tesla's ethical choices. This segmented analysis allows for a nuanced assessment. If it is found that two or more of these criteria—wrong-doing, causal contribution, foreseeability, and freedom of action—are met within the pieces of evidence, it would support the conclusion that Tesla bears moral responsibility for the March 2018 Autopilot incident.

Analysis

I will argue that Tesla violated all four of van de Poel's criteria of responsibility and is therefore morally responsible for the March 2018 incident. This will be done through the lens of 5 pieces of evidence.

Misleading Autopilot Claims

In a 3-hour-long investor presentation, Elon Musk, CEO of Tesla, stated: "We'll have over a million Tesla cars on the road … at a reliability level that we would consider that no one needs to pay attention … [you] could go to sleep" (*Tesla Autonomy Day* 2019). Elon Musk's ambitious claim about Tesla's Autopilot technology constitutes wrong-doing by creating misleading expectations about the system's capabilities. He suggests a level of autonomy that does not align with the current technological capabilities or regulatory approvals. The highest level of self-driving capability in a Tesla is SAE Level 2 (Teoh, 2020). SAE Level 2 automation involves systems that can control both steering and acceleration/deceleration simultaneously under certain conditions, but the driver must remain engaged and monitor the environment at all times. This level of automation requires the driver to be ready to take control of the vehicle at any moment (Teoh, 2020). It is important to note the weight of Musk's words, given his influence as CEO and figurehead of Tesla. This discrepancy between Musk's claims and the real-world functionality of Tesla Autopilot leads consumers to overestimate the technology's readiness and safety, posing risks to public safety. Musk's statement overstates the capabilities of Tesla Autopilot, leading to wrong-doing by misguiding consumers and stakeholders, thus breaching ethical standards for honest and responsible communication. By actively promoting the advanced capabilities of their Tesla Autopilot technology and setting ambitious goals for its

deployment, Tesla has contributed causally to setting a high level of expectation among consumers and stakeholders regarding the operational capabilities and safety of their autonomous driving technology. This contribution is not merely promotional but shapes how the technology is perceived and used by the public. Should Tesla's Autopilot technology not meet the lofty expectations set by such statements, particularly in terms of safety and reliability, the company's failure to adequately manage and temper expectations could be seen as a failure to act responsibly.

Autopilot's Value Hyperbole

The second piece of evidence is a quote from Elon Musk, CEO of Tesla. He stated the technology is "the difference between Tesla being worth a lot of money and being worth basically zero," (*Elon Musk on life, the universe and everything: Interview part 2* 2022).

The emphasis on the Autopilot system as a make-or-break factor for Tesla's valuation underscores the potential for misleading communications, which is central to the issue of wrong-doing. This statement is crucial as it directly showcases why Tesla, led by its CEO, may have significantly overstated the capabilities and future potential of the Autopilot system, leading to unrealistic expectations among consumers and investors regarding the current functionalities and near-term advancements of the technology. Such a portrayal could mislead stakeholders about the safety, reliability, and operational readiness of the Autopilot system, setting a premise for wrong-doing by fostering overconfidence in an underdeveloped technology. Furthermore, the freedom of action is evident in Musk's decision to make such a statement, indicating the leadership of Tesla's ability to choose their narrative around the Autopilot system.

Promised Full Autonomy Fallacy

"All you will need to do is get in and tell your car where to go." (*Autopilot*, 2016). This claim from Tesla's website paints a picture of full autonomy, suggesting a level of technological sophistication and safety that is not yet achievable. This portrayal potentially constitutes wrong-doing as it misleads consumers into believing the technology is more advanced and reliable in autonomous operation than it currently is, setting unrealistic expectations about the vehicle's operational capabilities. The claim directly implicates Tesla in making a causal contribution to the potential consequences associated with the use of its Autopilot technology. The causal contribution of Tesla in this context can be analyzed through the lens of action.

Tesla's promotion of Autopilot through such definitive statements acts as a direct causal contribution to user behavior and expectations. By suggesting that the cars are capable of complete autonomous driving, Tesla encourages users to rely heavily on the technology. This could lead to over-reliance and misuse by drivers who interpret the technology's capabilities as more advanced than they currently are, based on Tesla's promotional materials.

As I have shown, Tesla's portrayal of its Autopilot system, suggesting near or full autonomy, potentially misleads consumers about its current capabilities, setting unrealistic expectations about operational safety and the need for driver engagement. Advocates for Tesla's approach argue that the deployment of Autopilot and its incremental updates through real-world usage is an effective strategy for technological improvement. User M3FanOZ on the 'TeslaMotors' subreddit stated, "Tesla just does things differently. Waiting for the software to be perfect could delay cars for months or years. OTA updates provide a way of adding features that were not even thought of when the car was built. This is the future, the world needs to get used to it" (M3FanOZ, 2018). Advocates contend that this method allows for continuous learning and

adaptation, leveraging vast amounts of data to enhance safety features and system performance over time.

However, this perspective overlooks the critical importance of clear communication about the current limitations of Autopilot and the essential role of the driver in monitoring the driving environment. The work of Amudha Kamaraj et al. (2023) explored how varying levels of trust in vehicle automation influence responses to automation errors, shedding light on the importance of trust in the interaction between drivers and automated systems. The study found that automation-induced complacency and overreliance undermine drivers' responses to unexpected errors (Kamaraj et al., 2023). The study identified that contributory factors were drivers not being adequately engaged, highlighting the gap between perceived and actual system capabilities. Furthermore, ethical considerations demand that companies not only pursue technological advancements but also prioritize user understanding and safety in their deployment strategy. Thus, while continuous improvement is crucial, it should not come at the expense of clear and honest communication about the technology's present state and its safe use.

Overambitious Summon Feature Forecast

In 2016, Musk claimed, "I think within two years you'll be able to summon your car from across the country" (Frankel, 2016). Musk's prediction intersects with the criterion of foreseeability, which requires an assessment of whether the consequences of one's actions—or in this case, statements—were reasonably predictable. In this context, the reason to scrutinize Musk's statement is to evaluate whether it was reasonable for Musk to foresee the technological challenges and regulatory hurdles that would make his two-year timeline for cross-country car summoning ambitious, if not unfeasible. The evidence of Musk making such a specific

technological forecast suggests that, as a leading figure in the automotive and technology industries, Tesla should have been aware of the complexities involved in achieving such a feat.

Under the criterion of freedom of action, Musk's statement can be examined for the degree of autonomy and discretion he had in making such a forecast. The reason this analysis is pertinent is that it explores the extent to which Musk, as CEO, was free to make forward-looking statements about Tesla's technological advancements without undue pressure or compulsion. The evidence of Musk's proactive declaration suggests that he acted with a significant degree of freedom, not constrained by external forces but rather motivated by a vision for the future of autonomous vehicles.

Irresponsible Autopilot Demonstration

During a 2018 '60 Minutes' interview, Musk demonstrated Tesla's Autopilot without physically touching the steering wheel, claiming, "Not doing anything. No hands. No feet." (Stahl, 2018). By engaging in this demonstration, Musk conveyed a message that could encourage unsafe usage of the Autopilot system among Tesla drivers. The evidence of Musk not only driving hands-free but also acknowledging that he was not actively controlling the vehicle serves as a stark illustration of the discrepancy between responsible technology demonstration and the portrayal of an autonomous driving capability that is not yet ready for such hands-off use. This act of showcasing advanced technological features without emphasizing the current limitations and safety requirements can be seen as a violation of the ethical obligation to promote safe and informed usage of emerging technologies.

The evidence of Musk's hands-free driving on national television, coupled with his acknowledgment of not actively controlling the vehicle, may encourage similar behavior among

Tesla owners, despite the technology not being fully autonomous. This is a causal contribution to possible misunderstandings and unsafe practices.

Regarding foreseeability, the evidence provided by the interview suggests that Musk, given his expertise and position, should have been able to foresee the potential for his demonstration to be misinterpreted as an endorsement of hands-free driving, despite the technological and regulatory constraints that necessitate driver engagement.

In the context of freedom of action, Musk's demonstration of the Autopilot system without hands on the wheel can be analyzed for the degree of autonomy and choice involved in deciding to showcase the technology in this manner. The evidence of Musk's actions during the interview indicates that he had significant freedom in choosing how to present the Autopilot system to the public. This freedom, however, comes with a responsibility to prioritize safety and accurate representation of the technology's capabilities.

The evidence presented underscores a clear violation of Van de Poel's four criteria of moral responsibility by Tesla.

Conclusion

The March 2018 incident involving Tesla's Autopilot system raises profound questions about moral responsibility in the realm of autonomous technology. Through the lens of Van de Poel's criteria for passive responsibility, Tesla's role in this tragedy cannot be overlooked. The company's promotion of the Autopilot system crossed the line into wrong-doing by creating a perception of safety and reliability that was not fully supported by the technology's actual capabilities. Additionally, Tesla's apparent inaction in the face of known limitations of the Autopilot system represents a clear causal contribution to the fatal incident. Tesla could foresee the potential misuse of its system given its current developmental stage, and it still had the

freedom to implement more stringent safety measures or marketing practices. Despite the pressures of a competitive industry, ethical obligations must be paramount, especially when public safety is at stake. Tesla's actions, or lack thereof, in both marketing and developing the Autopilot system, have demonstrated a significant level of moral responsibility for the unfortunate loss of life.

References

Autopilot. Tesla. (2016, November 18). https://www.tesla.com/autopilot

Banks, V. A., Plant, K. L., & Stanton, N. A. (2018, October 1). Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. SAFETY SCIENCE, 108, 278 - 285.

Frankel, T. C. (2016, January 11). *Elon Musk says Tesla's autopilot is already 'probably' better ...* The Washington Post. https://www.washingtonpost.com/news/the-switch/wp/2016/01/11/elon-musk-says-teslas-autopilot-is-already-probably-better-than-human-drivers/

Hevelke, A., Nida-Rümelin, J. Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Sci Eng Ethics* **21**, 619–630 (2014, June 11). https://doi.org/10.1007/s11948-014-9565-5

Ibo van de Poel & Lamber Royakkers. (2011). *Ethics, Technology, and Engineering: An Introduction.* Wiley-Blackwell.

Kamaraj, A. V., Lee, J., Parker, J., Domeyer, J. E., Liu, S.-Y., & Lee, J. D. (2023, October 19). Bimodal Trust: High and Low Trust in Vehicle Automation Influence Response to Automation Errors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*(1), 1144-1149. https://doi.org/10.1177/21695067231196244

M3FanOZ. (2018, March 21). Tesla just does things differently… [Comment on the online forum post "OTA updates actually hendering Tesla image?"]. Reddit. https://www.reddit.com/r/teslamotors/comments/86158s/comment/dw1kcse/

Stahl, L. (2018, December 9). Tesla CEO Elon Musk: The 60 minutes interview. CBS News. https://www.cbsnews.com/news/tesla-ceo-elon-musk-the-2018-60-minutes-interview/

Teoh, E. R. (2020, February). What's in a name? drivers' perceptions of the use of five sae level 2 driving automation systems. Journal of Safety Research, 72, 145-151. https://doi.org/10.1016/j.jsr.2019.11.005

Tesla. (2019, April 22). Tesla Autonomy Day. YouTube. https://www.youtube.com/watch?v=Ucp0TTmvqOE

Tesla Owners Silicon Valley. (2022, June 14). *Elon Musk on life, the universe and everything: Interview part 2*. YouTube. https://www.youtube.com/watch?v=iHmSrK238vI&t=2333s