Regulation of Black Box Models:
A Case Study of Autonomous Vehicles

An STS Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Skylor Matsuda

Fall Semester 2021

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

Student

________________Skylor Matsuda__________________Date __________

Advisor

____Date ____________

Richard D. Jacques, Ph.D., Department of Engineering and Society

Introduction

Black box models pose a serious threat to the integrity of our social systems, whether in education or infrastructure, their reputation can be tarnished by their unpredictability. They are too complex to understand entirely by an individual- with the possibility of millions of variables being altered, it becomes a near impossible task to draw causation and thereby rectify the system. These networks serve to approximate mathematical functions, and with almost no exception, are inherently flawed in some form. In education, especially if gone unnoticed, the biases picked up by a machine will be transformed into biases in the individuals who learn from them. In our infrastructure, mathematical models will be forced to quantify the value of human life, leaving room for a torrent of moral discussion. It is paramount that we as a society, use these tools without malicious intent, but rather use them to further humanity as a whole.

Imagine a world where in place of a human, there is a black box model behind every automated wheel. Organizations such as the Congressional Research Service (2020) acknowledge that these models, in their maturity, have the ability to far outclass the driving ability that appears on the road today. These experimental models in their current form must be tested live on the road as they derive their intelligence from experience just as humans do, and not every scenario can be preprogrammed. The result in society is not a technological question, but rather a philosophical one, where each stratum of society is asked to weigh the costs of rational fear and sacrifice against potential long-term societal and socioeconomic benefits.

In theory, autonomous vehicles offer solutions to a conglomerate of modern issues. However, in practice, they spawn a host of controversial dilemmas. Early-stage technologies are appropriately subject to intense scrutiny and regulatory petitioning. A plethora of socioeconomic groups in the United States have lobbied all levels of government over the regulation of the development of self-driving cars. Incidents involving autonomous vehicles have provoked liability questions and concerns as to where to point the blame. A byproduct of these issues is the identification of potential scapegoats, these being entities such as the corporation and its cogs, the inattentive individual behind the wheel, and the failed governmental check. The arrival of automotive automation uproots some current laws and precedents designed for an antiquated era in which a tangible human was expected to be behind the wheel.

Autonomous vehicles are praised by technologists as a key twenty-first century innovation. On an individual consumer level, they provide convenience and robust economic incentives by saving time and enabling families to reduce their required number of cars (Pettigrew, 2018). In a macro environment, they offer a series of promising possibilities: decreased fatality rates due to traffic and human error, a reduction in harmful emissions through reduced traffic and clean energy sources, and fewer hours wasted on already dilapidated, overloaded infrastructure. These arguments are further reinforced by some of autonomous vehicles' practical outcomes such as the elimination of drunk and distracted driving or increased accessibility for the disabled (Pettigrew, 2018). These vehicles appeal to a multitude of heterogeneous social groups that vary in their opinions and psychology regarding topics such as environmentalism, individualism, utility, and safety.

Literature Review

‘Autonomous vehicles (AVs) derive their roots from a variety of sources, but the most influential institutions with regard to the genesis of AVs have been governmental entities. Driven by national security and institutional agendas, the United States government established numerous programs to accelerate the development of autonomous technologies. A particularly notable fruitful push for technological innovation was the Stanford Cart, which was developed during the apex of the Cold War between the Soviet Union and the United States.   The Stanford Cart is considered to be the world’s first self-driving wheeled vehicle - a lunar rover designed to overcome latency engineering challenges through self-driving (Kubota, 2019).

In response to a period of stagnation of autonomous driving technology, the Defence Advanced Research Project Agency (DARPA), sponsored the “DARPA Challenges” in 2004 . This initiative funded by the research arm of the United States Department of Defense subsequently reinvigorated the development of autonomous technologies through economic incentives and propaganda measures, as the underlying technologies appealed to futurists, industrialists/entrepreneurs, and militarists alike (Anderson et al., 2014). Notably, the U.S. government has repeatedly leveraged cultural phenomena in media and institutions such as Hollywood to sway public opinion. Past instances of governmental intervention include Hollywood films such as Transformers, Iron Man, and The Terminator, all of which contain highly technological and militaristic themes and were recipients of U.S. Department of Defense allocations (Underhill, 2013).

Initial reactions to AVs in Western cultures (for simplicity, Western Europe and the United States) have been relatively ambivalent. While they acknowledge the potential of AVs, these cultures and societies typically emphasize individualism; As a result, technologies that conflict with Western doctrine can be and are met with stiff resistance. The first documented case of a pedestrian fatality involving an autonomous vehicle occurred on March 18, 2019. Republican Governor Ducey of Arizona, in his letter to Uber ordering the suspension of AV testing (2019), wrote:

> "...As governor, my top priority is public safety. Improving public safety has always been the emphasis of Arizona's approach to autonomous vehicle testing, and my expectation is that public safety is also the top priority for all who operate this technology in the state of Arizona.
>
> The incident that took place on March 18 is an unquestionable failure to comply with this expectation. While the incident is currently under investigation by the National Transportation Safety Board and the National Highway Traffic Safety Administration, Arizona must take action now..." (paras. 2-3).

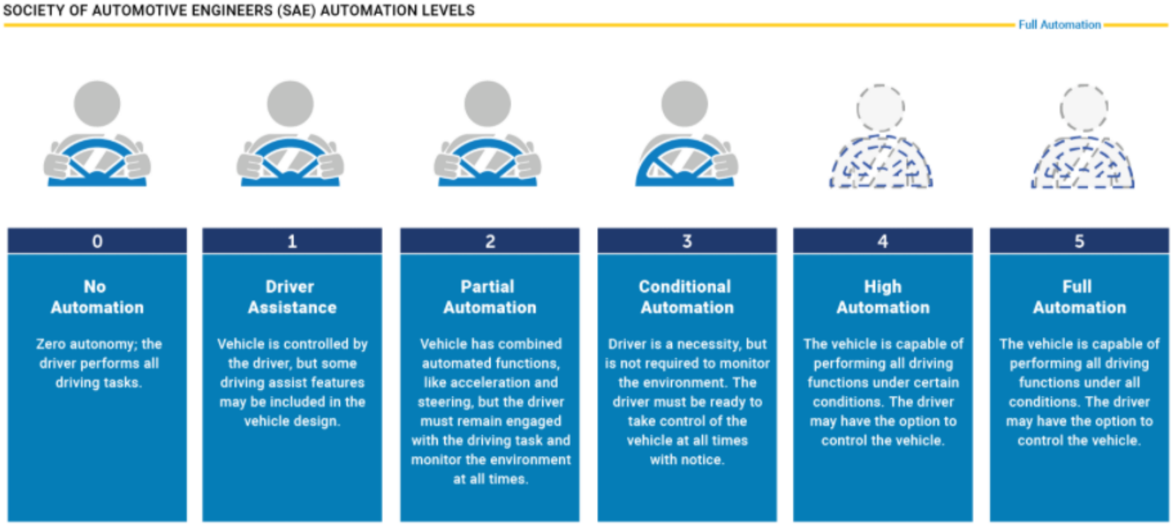
Ironically, Governor Ducey had previously been a proponent of AVs. In 2015, via an executive order entitled the "Self-Driving Vehicle Testing and Piloting in the State of Arizona; Self-Driving Vehicle Oversight Committee", he established optimistic goals and groundwork for AV development within the state of Arizona (Ducey, 2015). In the first clause of his order, he effectively wrote a blank cheque:

> (1) The Department of Transportation, Department of Public Safety, and all other agencies of the State of Arizona with pertinent regulatory jurisdiction shall undertake any necessary steps to support the testing and operation of self-driving vehicles on public roads within Arizona (para. 1).

His positional reversal occurred in response to the public uproar following the demonization of AVs by the media, and the combination of political interest groups and public backlash put Uber in a position where it was forced to temporarily shutter its research and development programs in AV technology (Congressional Research Service, 2020). In the following analysis of the incident, the Congressional Research Service notes that "pace of autonomous vehicle commercialization may have slowed due to the 2018 death in Arizona of a pedestrian… highlighted the challenges of duplicating human decision making," and that the National Transportation Safety Board concluded that a core cause of the fatality was "'inadequate safety culture' at Uber" (Congressional Research Service, 2020). To this day, some of the top search engine results on Google (incognito) for "uber self driving car" include "...accident," "...death," and "...kills pedestrian," demonstrating the damaging impact not only on company image and legacy, but more critically on the public's confidence in AV technology as a whole.

Societal pressure on regulatory authorities prompted a governmental investigation into vehicle sensor systems and dashcams. It was found that both the software and safety driver were at fault, the former suffering from experimental braking systems and the latter indicted on a count of negligent homicide. Fortunately, AVs provide no shortage of evidence to put before a judge or jury. However, further implications of this practice call into question democratic ideals such as privacy and morality, which must necessarily be addressed in any calls for the adoption of this technology.

Although AV technology is in its infancy, the impact of the level of sophistication influences the construction of the overarching system. As per the National Highway Traffic Safety Administration (NHTSA), autonomy levels are defined as follows (*The Evolution of Automated Safety Technologies,* 2020):


SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

0	1	2	3	4	5
No Automation	**Driver Assistance**	**Partial Automation**	**Conditional Automation**	**High Automation**	**Full Automation**
Zero autonomy; the driver performs all driving tasks.	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

These guidelines are vague and include flexible verbiage such as “some circumstances” and “other circumstances.” The guidelines’ interpretative flexibility is both a blessing and a curse; it establishes expectations whilst other branches of government catch up, but in doing so yields all the negative byproducts of an accelerated, nascent industry. While autonomous driving systems are sophisticated and designed to be robust, an assortment of various complex sensors and software suites intended to gain insight on the road are all subject to failure. By synthesizing a fully autonomous system and marketing it as such, manufacturers expose themselves to an extensive framework of product liability law (Villasenor, 2014). While not all-encompassing, litigation efforts in civil courts involving AVs typically fall into the categories of strict liability, negligence, and misrepresentation.

Strict liability in AVs is the premise that manufacturers own the burden for harm caused by their products. Negligence is the lack of due diligence on the part of a manufacturer in ensuring the safety of their products. Misrepresentation of a product by the manufacturer in determining product capabilities, notably AV autonomy level, is commonplace due to conflicting interests on the part of the individual and manufacturer (Grossman, 2018).

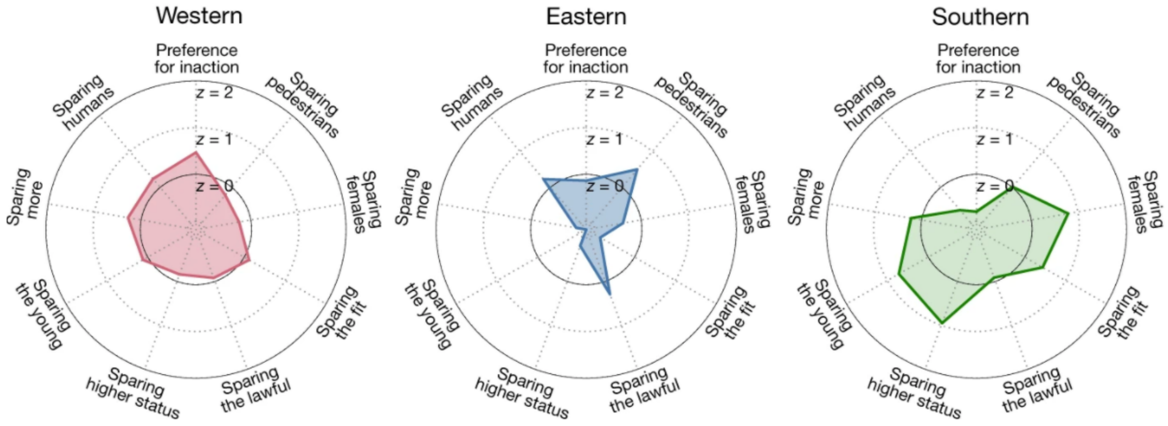
Despite the complexity of the legal system, Andrew Garza argues in his publication in the *New England Law Review* titled "'Look Ma, No Hands!': Wrinkles and Wrecks in the Age of Autonomous Vehicles," that "despite catastrophization of critics, increased manufacturer liability will need not be a dire concern" (Garza, 2012). Garza emphasizes the improvement of all aspects of autonomous vehicles over predecessors, which he claims "will lead to a net cost of insurance and litigation." Garza further notes that "Despite their historic reluctance to incorporate safety devices, manufacturers will ultimately benefit by the accompanying reduction in injuries and fatalities," drawing parallels to the development of seat belt laws and airbags, and their general acceptance by society due to their matured development cycle.

In an attempt to gauge societal expectations of an AVs response to a moral question, Edmond Awad et al. (2018) in their acclaimed research paper titled "The Moral Machine experiment," created a global challenge called "The Moral Machine" where humans input responses to theoretical moral dilemmas derived from unavoidable AV accidents. Awad et al. found multiple "cross-cultural variation, and uncovered three major clusters of countries," identifying these clusters as Western, Eastern, and Southern. These researchers also note dominant religious affiliations, introducing yet another set of stakeholders, in particular regions where "cluster 2 (Eastern) consists mostly of countries of Islamic and Confucian cultures" and "cluster 1 (Western) has large percentages of Protestant, Catholic, and Orthodox countries in Europe."

They further extend their analysis by highlighting the "systematic differences between individualistic and collectivistic cultures," and that individualistic cultures tend show a stronger preference for saving a greater number of individuals, whereas individuals that comprise collectivist cultures typically place emphasis on elders and thus are spared at a higher rate in these societies. Although this study fails to recognize numerous biases such as ethnicity in the underlying models, this discrepancy in cultural preferences highlights the importance of adjusting black box models and their regulation to fit the will of each society.

> "[Ethical] work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo," (Awad E., et al. 2018)

Figure 1



(Awad E., et al. 2018)

Kallioinen et al. (2019) suggest that: “[P]eople considered it more morally acceptable for self-driving cars to endanger fewer lives, even at the expense of the occupants' lives,” but ironically “preferred to purchase cars that would protect occupants,” highlighting yet another clash of interest between stakeholders. Luzuriaga et al. (2019) furthers this notion that in previous studies, researchers found empirical evidence in surveys of individual preference for AVs to be self-sacrificial, but when individuals are placed in a simulated AV driving environment, the participants, on average, prioritized the passengers.

Methodology

The Social Construction of Technology (SCOT) and Artifact frameworks was utilized in the previous analysis. SCOT analysis allows for the evaluation of human action and influence on AV advancement, while Artifact analysis was leveraged to unveil politics behind the construction of AV systems.

In order to promote AV adoption, this research aims to gather data concerning the safety metrics of autonomous vehicles. It aims to draw a cross-comparison between the safety yield of autonomous vehicles and the yield of the implementation of seat belt laws and similar legislation in an effort to contextualize the impact of a transition to AV technology on its primary concern: public safety. In doing so, it clarifies a topic riddled with confusion and misinformation. Supporting datasets and documentation are in the process of becoming available for public use and the impact on safety can be modeled and adjusted for confluence, but the end objective is to create a set of interpretable artifacts that can be evaluated by their effect on individual opinion. In addition to these quantitative datasets, policy, and archival studies were used to gain additional insights and to stratify against sources of bias such as the regional implementation of legislation.

Results and Discussion

The efficacy of autonomous driving technologies is highly dependent on the underlying datasets and proportionality of the market penetration of AVs. Research into the infrastructural and macroeconomic impact of autonomous vehicles, conducted by Bernhard Friedrich (2016), concluded that a sizable portion of the benefits associated with AVs are reduced or eliminated by low market penetration. In AV simulated environments, homogeneous traffic flow allows for a magnitude increase of up to 40 and 80 percent in cities and interstates, respectively. An automotive fleet composed solely of AVs provides increased traffic stability, which mitigates the effects of high-volume, inconsistent, and rush hour traffic. When non-autonomous elements are introduced, the efficacy of the researchers' construct drops. For example, when a single human driver is introduced into an autonomous driving column, the lack of "column hegemony" leads to lower velocity, a severe reduction in capacity, and an increased probability of an accident (Friedrich, 2016). Analogously, a human driver is to automation as a virus is to vaccination.

Simone Pettigrew et al. (2018) argues that the Australians are more “likely to be receptive to autonomous vehicles when provided with information relating to their public health benefits.” They appeal to public health as an approach to AV adoption in society, but in their questionnaire, they inherently add biases as the very nature of their survey is to introduce additional positive information in relation to AVs. Regardless, Pettigrew contends that the introduction of health benefits aside from reduced crashes, such as stress reduction, pollution reduction, and increased mobility for the elderly/disabled prompted awareness and increased saliency in the rate of adoption.

Pettigrew and Bernhard Friedrich’s research provokes societal and cultural questions with regard to the infusion of autonomous technology into society. A powerful moral case that advocates for safety and the right to life, constructed from possible safety “improvements” of technological innovation, can also be manifested by opposition parties using congruent logic, but with a varied time horizon. The result becomes a matter of controversy where individuals are asked to rationalize unproven, experimental, yet life-saving technologies with temporary, inflated risk. The controversy over concepts as innate as safety and well-being underscores the complexity of assimilation and the need for resolution.

Conclusion and Recommendations

AV systems perform better with more experience. If safety were the only priority, AV data ought to be nationalized, but this fully contradicts economic principles in the West. If individuals choose to use AVs in the first place, they should be as safe as possible and, by extension, take a larger proportion of market share on the roads that are shared by all branches of society. However, AV technology is met with intense public skepticism, scrutiny, and backlash. To alleviate these concerns, this research aims to help build the basis for garnering the support of society to integrate hardware into all coming vehicles, while the software can be deployed later. It aims to draw congruences to the seat belt, a tangible safety item that has shown proven effectiveness. Seat belt laws have been left mostly to the states; however, federal regulation mandates the very existence of the seat belt in automotive vehicles. This research aims to establish that the safety that an AV brings is not a novelty, and that capability should be built into every car, just like the seat belt.

References

Anderson, J. M., N. D., Stanley, K., Sorensen, P., Samaras, C., & Oluwatola, O. A. (2014). Autonomous Vehicle Technology: A Guide for Policymakers. Retrieved from https://www.jstor.org/stable/10.7249/j.ctt5hhwgz.11?seq=1#metadata_info_tab_contents.

Awad, E., Dsouza, S., Kim, R. et al. (2018). The Moral Machine experiment. Nature 563, 59–64. Retrieved from https://www.nature.com/articles/s41586-018-0637-6.

Congressional Research Service. (2020, February 11). Issues in Autonomous Vehicle Testing and Deployment. Retrieved from https://fas.org/sgp/crs/misc/R45985.pdf.

Ducey, D. A. (2019, March 26). Ducey Gov Arizona Uber Letter. Retrieved from https://www.documentcloud.org/documents/4424723-Ducey-Gov-Arizona-Uber-Letter.htMl.

Friedrich, B. (2016). The Effect of Autonomous Vehicles on Traffic. *Autonomous Driving,* 317-334. doi:10.1007/978-3-662-48847-8_16.

Front. Psychol., (2019, 01 November). Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02415/full**.**

Garza, A. P. (2012). Look ma, no hands: Wrinkles and wrecks in the age of autonomous vehicles. New England Law Review, 46(3), 581-616.

Grossman, M. D. (2018, April 5). The New Norm: Strict Liability For Autonomous Automobile Accidents. Retrieved from https://medium.com/@marcdgrossman/the-new-norm-strict-liability-for-autonomous-automobile-accidents-38bbd0dc0a43#:~:text=In%202017%2C%20a%20GM%20Cruise,before%20recording%20its%20first%20accident.&text=Hence%2C%20law%20advocates%20agree%20the,malfunctions%20in%20self%2Ddriving%20mode.

Kubota, Taylor. (2019, January 16). Stanford's robotics legacy. Retrieved from https://news.stanford.edu/2019/01/16/stanfords-robotics-legacy/.

Lee, Timothy B. (2020, September 15). "Safety driver in 2018 Uber crash is charged with negligent homicide." *ars Technica*. Retrieved from https://arstechnica.com/cars/2020/09/arizona-prosecutes-uber-safety-driver-but-not-uber-for-fatal-2018-crash/ .

Luzuriaga, Miguel et al. (2019, March 20) Hurting Others vs. Hurting Myself, a Dilemma for our Autonomous Vehicle. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3345141&download=yes.

Pettigrew, Simone et al. (2018, July 4) The health benefits of autonomous vehicles: public awareness and receptivity in Australia. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/1753-6405.12805.

The Evolution of Automated Safety Technologies. (2020, June 15). Retrieved from https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

Underhill, Stephen. (2013, February 4). "Complete List of Commercial Films Produced with Assistance from the Pentagon." Retrieved from https://www.academia.edu/4460251/Complete_List_of_Commercial_Films_Produced_with_Assistance_from_the_Pentagon.

Villasenor, J. (2014, April 24). Products Liability and Driverless Cars: Issues and Guiding Principles for Legislation. Retrieved from https://www.brookings.edu/research/products-liability-and-driverless-cars-issues-and-guiding-principles-for-legislation/