The Introduction of Autonomous Vehicles into Society: Encoding Morals into the Machine
(STS Paper)

Katie Kleeman
Spring 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.

Signature _____ Date _____

Approved _____ Date _____
Travis Elliot, Department of Engineering and Society

STS Thesis

Introduction

As autonomous vehicle technology improves, it is important to analyze how this could influence, and be influenced by, society. Almost every major car manufacturer is attempting to develop autonomous cars meant for passenger transport, and they all make bold claims about improvements in safety and increased productivity for those who buy into the craze. However, many experts are asking if they will really be able to deliver on those promises. This issue is divided into two parts: first, how the autonomous vehicles interact with and influence the world around them; second, how society will influence the development and introduction of autonomous vehicles. The first aspect is best explained through the lens of the actor-network theory (ANT) framework. A key tenet of this framework is the principle of generalized symmetry - stating that human and non-human actors have equal agency within their network. Using this as a base, I will analyze how autonomous vehicles operate within their network and impact various stakeholders. Next, I will use the Social Construction of Technology (SCOT) to analyze how societal views of autonomous technology will impact the development and adoption of these vehicles in order to address the second aspect. I will be defining the different stakeholders involved, and how their various perspectives could potentially affect the success or failure of autonomous vehicles.

Background and Context

Autonomous vehicles, colloquially known as self-driving cars, have the potential to be transformative to the transportation sector and automotive industry. Over forty companies, ranging from existing car brands to technology giants, are developing their own version of an

autonomous vehicle — and with good reason ("Autonomous Vehicles & Car Companies l CB Insights", 2020). By 2025, the market for self-driving cars is expected to reach $42 billion while maintaining a compound annual growth rate of approximately 21% through the year 2030. In addition to the immense business potential, autonomous vehicles also have humanitarian reasons for their growth in popularity. The World Health Organization (WHO) approximates that traffic accidents account for 1.35 million deaths globally each year (Snell, 2019). In just the United States alone, that number can reach over 40,000 traffic-related deaths. The U.S. National Highway Traffic Safety Administration reports that 94% of those deaths are caused by human error, meaning that the promise of well-developed autonomous vehicles could save thousands of lives. The question then becomes - how effectively can these companies deliver on their promises for increased safety? The international Society of Automotive Engineers defined six levels of autonomy for vehicles ranging from 0 (no automation) to 5 (full autonomy) (“The Rise of Autonomous Vehicles.”, 2020). The highest level of autonomy that has been reached (and is available to the public) is a three, meaning it may not reap the full benefits of a fully autonomous car. Regardless of this, it is clear that the introduction of self-driving cars will be disruptive, creating new markets and business models, as well as changing the workforce, leading to a variety of ethical and societal implications.

Actor-Network Theory

Actor-network theory (ANT) views “society” as an ongoing achievement, rather than a stagnant concept, and provides the tools to analyze the way in which society redesigns itself (Callon, 2001). It perceives scientific knowledge to be the effect of relationships and connections between objects, animals, ideologies, humans, social rules, and any other thing that affects

technological development - these are the actors. Evolving from the idea of constructivism in the 1980s, ANT's originating authors, Bruno Latour, Michel Callon, and John Law, rejected the idea of social determination of scientific knowledge (Detel, 2001). The primary tenet of ANT is that the aforementioned actors interact with each other in a heterogeneous network. Moreso, the principle of generalized symmetry states that within that network human and non-human actors are to be given equal amounts of agency (L, 2007). The network operates via translations, when actors transform other actors, and by passing tokens (or "quasi-objects") between the actors. Tokens are created when the interactions within the network are successful ("Actor–network theory", 2020). When the actor-network is running smoothly, societal order is produced, but this order breaks down when certain actors are removed or tokens are failed to be transmitted (L, 2007). Following the principle of generalized symmetry, I plan to acknowledge autonomous vehicles as having equal agency as all other actors in their network. Other non-human and ideological actors include morals and ethics, safety, social rules, and influential institutions. While acknowledging the most prominent actors involved in this network, I will investigate how they might translate each other, pass tokens, or generally interact to contribute to the development of autonomous technology.

ANT Analysis

Autonomous vehicles are complex technological innovations that, if implemented at a large scale, will have a great impact on everything around them. From the perspective of actor-network theory and the aforementioned property of generalized symmetry, these vehicles could play a huge role in the development of evolution of our society. Society is defined through many elements - social and non-social, human and non-human - that interact with each other in a

way that creates a functioning system. Actor-network theory can then be used to map how technology and other artifacts participate in everyday lives. These material objects are more than just tools, but rather are inscribed with our practices and culture (Kien, 2016). Therefore, it is crucial to analyze how autonomous vehicles will serve as actors in their network - how they will be programmed to make decisions, how morality will be tokenized as it is passed from the engineer to the technology, and what translation could look like with regards to the vehicles.

The Agency of Autonomous Vehicles

In 1942, science-fiction writer Isaac Asimov published a short story "Runaround" and, although he had implied them in previous works, it was the first time he explicitly stated the Three Laws of Robotics that set a precedent for relationships with artificial intelligence. The rules were as follows: "1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; 2) A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law; and 3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws" (Asimov, 1942). These rules, while easy to understand, are hard to properly implement, especially in complex cases such as autonomous vehicles. Sentient technology has to make decisions on its own, in a way that impacts the network in which it operates. Moreso, in a future where self-driving cars occupy the majority of the road, small changes in operation will multiply, resulting in large changes in the aggregate. Therefore the decisions on the programming of that technology will make big differences in the future, especially as power translates from the developers to the vehicles themselves (Himmelreich, 2020). Understanding this, and accepting that most decisions will fall into grey areas, one can see how important context comes into play

when making judgements or applying Asimov's laws. This is not only computationally strenuous, but also often impossible for the human engineer to know ahead of time (Roff, 2019).

The Algorithms Behind the Autonomy

The discussion from the last paragraph highlights the idea that there are skills that humans are inherently good at, yet are incredibly hard to teach because they were built in by evolution - this is known as Moravec's Paradox (Himmelreich, 2020). Decision making, when not guided by objective reasoning, is very difficult to program into a computer. In the discussion on autonomous technology, questions akin to the famous Trolley Problem often arise - given no other options, should the car hit a pedestrian or swerve into a wall, injuring or killing its passengers? The two approaches to this problem are rule-based and logic-based decision making. In rule-based decision making, engineers hard-code explicit rules into the algorithm. This is beneficial as the autonomous behavior will be more predictable and their choices have a clear explanation. However some experts argue that logic is the ideal choice for encoding ethics, as it more closely resembles the way humans make decisions. This would involve having the robots learn from the past through machine learning, making them more adaptable and useful (Deng, 2015). This involves making decisions to prioritize certain objectives over others, and the tradeoffs involved in that. The way this is done is through probabilistic modeling, such as partially observable Markov decision process which models an actor's decision process given incomplete information about the circumstances, and reinforcement learning. In order to do this effectively, however, one must identify the values we want to realize through the programming of self-driving cars, and create functions that maximize them (Roff, 2019).

Morality as a Token

Morality, similarly to the autonomous vehicles themselves, are a non-human yet vital aspect of this network. The concepts defined as morality are created through interactions of the societal actor-network, and subsequently passed between actors within the network - in this case the engineers and the machine learning (Kien, 2016). With regards to mathematically defining these value functions concerning morality in autonomous systems, expectations dictate that it should be "universally appropriate" and "representative of rational human judgment" ("The Rise of Autonomous Vehicles.", 2020). However, this may be even more difficult than previously thought. An international survey of 2.3 million people revealed that a universal ethical code does not exist, but rather the local cultural norms influence these tough decisions. For example, in the case of choosing between hitting someone illegally crossing the road or swerving and killing the passengers, people from countries with strong governments more often opted to hit the jaywalker than people from nations with weaker governments. Moreover, given the choice between hitting a homeless person or an executive, people from countries with a large economic disparity choose the lower-status person compared to nations with small wealth gaps. This further extends to self-driving cars as in surveys many people say they would rather autonomous technology protect a pedestrian over the passengers, but that they wouldn't buy a car that was programmed this way (Maxmen, 2018). It is clear that the morality issue is not an easy one to solve, but given that autonomous cars will be in near-complete control of its passengers, it is an important discussion to be had.

Translation and the Role of the Autonomous Vehicle

Translation occurs when actors transform other actors in the network, and there is no question that autonomous technology is set to be transformative. As Grant Kien states it, "translation is the mechanism by which the social and natural worlds progressively take form -

the result is a situation in which certain entities control others" (Kien, 2016). This leads to a question of power, in which the controlling actor, the autonomous vehicle, has power over others - the passengers and other drivers. This concept of power, defined through the lens of the ANT framework, posits that power is the effect of associating actors together. In this case, it is the association of autonomous technology with morality and ethics that gives autonomous vehicles the power (Kien, 2016). This is translation in our case - the recombination and alignment of human actors, sentient technology, morals, and mathematical programming to enact and promote self-driving cars in society (Grabher, 2009).

Social Construction of Technology

Drawing on actor-network theory and inspired by the principle of symmetry from Sociology of Scientific Knowledge, the Social Construction of Technology framework holds that the success of an innovation is not because it "works" better than failed innovations, but rather is due to the social context that promotes (or fails to promote) it ("Social construction of technology (SCOT)"). To that point, SCOT aims to investigate and analyze the metrics used to judge technology, which groups or stakeholders defined them, who is included or excluded from that group, and why they defined it in a particular way. It proposes that human action shapes technology, and its uses need to be understood in the context of society. The leading scholars on this topic, Wiebe Bijker and Trevor Pinch, began this work as a response to technological determinism, the idea that a society's technology determines its social structure and values. Disagreeing with that viewpoint, they adapted the Empirical Programme of Relativism (EPOR), a method that analyzes how scientific findings are socially constructed, and produced the SCOT framework (Klett, 2018). A few key tenets to this concept are relative social groups, interpretive

flexibility, and stabilization. Relative social groups are defined as "all members of a certain social group who share the same set of meanings, attached to a specific artifact" ("Social construction of technology (SCOT)"). These different groups may attach different meanings and morals to the problem at hand, which leads to the next tenet. Interpretive flexibility states that there is no ground truth "best" artifact, technology, or methodology, but rather each group defines what they mean by "best" based on the interpretation of the problem they are trying to solve. Finally, stabilization occurs when one social group dominates the others, and therefore their ideals and design prevails while the others are forgotten ("Social construction of technology (SCOT)"). By defining some of the important social groups involved in this issue and analyzing the ways in which they interpret and understand autonomous vehicles, I will be able to assess what stabilization could look like for autonomous technology in the future.

SCOT Analysis

With a technology as newsworthy and disruptive as autonomous vehicles, the impact will be far-reaching. The core ideology of SCOT is that technology succeeds because it has been socially supported, rather than that it has inherent "goodness". More specifically, it succeeds because it aligns with the metrics defined by a dominant social group, who then supports that technology. This process is also known as stabilization, but the emergence of a single dominant group is not always instantaneous. With the prospect of autonomous vehicles on the horizon, there are multiple sets of stakeholders, each with their own metrics, values, motivations, and interpretations of the problem. The four main relative social groups I will discuss are the general public, the developers of this technology, the government and lawmakers, and transportation industry workers and employees. That is not to say that there are no other relevant stakeholders,

but focusing on these main four will allow me to dive deeper into their interpretations and understand how it will affect the development and adoption of autonomous vehicles.

The General Public

One of the largest, yet most unpredictable, relative social groups is the general public. This group will be the consumers of the products and services allowed by this technology, but could also be affected by it, even as a non-consumer. What makes them so unpredictable is the diversity of thought, even within this group, and - even more so - the fear factor. It is a human tendency to be scared of the unknown and uncontrollable. This can be seen throughout history, including the autonomization of the elevator. When operators were removed from elevators, people were far too scared to use them. It required creative advertisements, the addition of a soothing voice, and big red stop buttons to dispel the fear (Snell, 2019). When it comes to self-driving cars, they are predicted to cause fewer accidents proportionally compared to human drivers, but due to the novelty of the technology, those accidents will receive more attention. The increased negative coverage in the media could lead to an availability bias, where the public interprets the technology as less safe, even when the opposite is true (Maxmen, 2018). Another source of fear is data privacy and hacking concerns — as the Internet of Things (IoT) continues to expand, people are becoming more aware of the amount of personal data being collected. In addition to personal privacy concerns, a computer-run car becomes susceptible to cyberattacks, putting both the passenger and those around the car in danger (“The Rise of Autonomous Vehicles.”, 2020). However, many of these sources of fear come from consumers not understanding the technology and a lack of control. Building trust and understanding could help reduce this fear. As Michael Fisher, a computer scientist at the University of Liverpool, UK, says, “People are going to be scared of robots if they're not sure what it's doing, but if we can

analyse and prove the reasons for their actions, we are more likely to surmount that trust issue" (Deng, 2015). Once the public reduces the fear factor, there are actually a few aspects of autonomous technology that could be beneficial to them, namely the potential democratization of access to transportation. Autonomous cars will have lower operating costs per mile, increase access to reliable transportation, and give the general public more disposable income by removing the sunk cost of car ownership - thus leading to economic growth in urban areas ("The Rise of Autonomous Vehicles.", 2020). Therefore, according to the SCOT principle of interpretive flexibility, autonomous vehicles could be a success in the eyes of the public - and this is an incredibly important group. Consumers decide what technology is advanced by voting with our money, leading developers to follow the standard set by consumers (Snell, 2019).

Developers of Autonomous Vehicles

Advanced technology, such as autonomy, is developed in leading technology companies and developers, rather than governmental research. This fact is supported by comparing the research and development budget of the top five U.S. defence contractors to that of any major technology company - developers like Google, Uber, or Amazon spend over twice as much (Snell, 2019). This means that technological developers have the biggest say in how autonomous technology is developed and implemented, but they are still required to answer to other relative social groups. As mentioned above, consumers - the public - use their money to support companies that develop products beneficial for them. Therefore, any successful developer needs to be aware of and include the wants and needs of the public in order to be successful. After all, while the public interprets a successful autonomous vehicle as one that improves safety and reduces their cost, developers define success by equating it to profit (and occasionally technological advancement). In addition to the public, developers are also subject to standards

set by the law, and in the case of a technology as new as autonomy, these standards can be unclear. Without human drivers in control, when an accident does occur there is ambiguity to how the legal system would apply blame. Does the passenger or owner of the vehicle claim responsibility, or does negligence fall to the developers and executives who made and sold the vehicle? This question becomes important for companies to consider as they develop this new technology ("The Rise of Autonomous Vehicles.", 2020). While this problem might seem to be the responsibility of lawmakers to solve, as we can see from other technological advancements, regulation is not able to stay up to date with fast growing technology. Therefore, developers will have to make decisions guided by supply and demand in order to be successful (Snell, 2019).

Government and Lawmakers

While the speed of autonomous vehicle development has made it difficult to regulate, there has been some progress by the government and lawmakers. In 2017, the Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution (SELF DRIVE) act was proposed to the U.S. congress. The goal of the legislation was "transferring jurisdiction over autonomous vehicle testing from American states to the federal government" ("The Rise of Autonomous Vehicles.", 2020). This is an important factor for developers, because it would mean all companies, regardless of location, would be subject to the same metrics. In addition, the legislation also assigns "full control over autonomous vehicle design and construction" to the National Highway Traffic Safety Administration, requiring it to define a series of safety guidelines within two years ("The Rise of Autonomous Vehicles", 2020). This bill has been passed by the House as of 2020 and is awaiting Senate approval. In 2018, California defined a set of rules to qualify for permits for a level four autonomous vehicle. These qualifications defne that the vehicle must "be able to resist cyberattacks, operate only within specific regions of a

city, and come equipped with a form of two-way communication, as adjudicated by the California Department of Motor Vehicles" ("The Rise of Autonomous Vehicles", 2020). Outside the United States, Germany has led the European Union in autonomous vehicle policy, which is unsurprising due to its thriving automotive industry. Analyzing this relative social group, one can see that safety is the main priority among governments around the world. For this group, a successful introduction of autonomous vehicles would rely on clear policy that removes ambiguity and emphasizes safety, by holding various companies to federally defined standards.

Transportation Industry

Finally, transportation industry workers are set to face major changes as a result of autonomous vehicles, and potential job loss. If autonomous vehicles are able to be mass produced and implemented, it could give rise to brand new business and ownership models. In the United Kingdom (as of 2017), there were 281,000 licensed vehicles (taxis and private hire) in operation. The drivers of these vehicles are almost entirely from the lower middle class, utilizing the steady transportation job as a vehicle for economic mobility. The idea of taxi and rideshare companies utilizing autonomous vehicles to reduce costs and increase efficiency means a massive loss of jobs for the middle class. This could cause massive problems for urban populations if there is no plan for these displaced workers ("The Rise of Autonomous Vehicles", 2020).

Stabilization

In looking at these various social groups, and their interpretations of and impact on autonomous vehicles, one can understand the differences in their perspectives on the topic. Regardless of the engineering success of autonomous vehicles (getting them to operate correctly), how do these various groups define success and what will that mean for the future of

autonomous vehicles? We can see that for the public, success means both physical safety and data security, as well as lower costs and increased efficiency. For developers of the vehicles, they need to make a profit from these cars, and therefore need to appeal to the public. Lawmakers need to ensure a standardization of the metrics, as well as the safety of the public. Industry workers want to ensure that they are not completely displaced. One common thread through each definition of success is the public. They are the largest group, as well as the group with the most influence, from the way they influence developers through supply and demand, to the way they influence lawmakers through the election process in the United States. From this, one could hypothesize that in the stabilization process the public is who comes out on top, but only time will tell.

Conclusion

Autonomous vehicles are on the rise, and are set to have a transformative impact on society. As free-acting agents, they will interact with other drivers on the road, pedestrians and bicyclists, and their riders. A key component to determining how this translation will occur is focusing on the tokenization of morality, and defining the values we want embedded in this artificial intelligence, as well as how those values can be mathematically programmed. However, understanding the role of autonomous vehicles does not help forecast if this novel and exciting technology will be successful. The success will be determined by the metrics set by the public, the developers, lawmakers, transportation workers, and other stakeholders. Stabilization has yet to happen, and therefore the future of autonomous vehicles has not yet fully formed.

References

Actor–network theory. (2020, October 09). Retrieved October 18, 2020, from https://en.wikipedia.org/wiki/Actor–network_theory

Asimov, I. (1942) "Runaround." *Astounding Science Fiction*, Street & Smith.

Autonomous Vehicles & Car Companies l CB Insights. (2020, December 16). Retrieved February 23, 2021, from https://www.cbinsights.com/research/autonomous-driverless-vehicles-corporations-list/#:~:text=Using CB Insights' investment, acquisition,technology brands and telecommunications companies.

Callon, M. (2001). Actor Network Theory. *International Encyclopedia of the Social & Behavioral Sciences,* 62-66. doi:10.1016/b0-08-043076-7/03168-5

Deng, B. (2015, July 1). Machine ethics: The robot's dilemma. Retrieved October 18, 2020, from https://www.nature.com/news/machine-ethics-the-robot-s-dilemma-1.17881

Detel, W. (2001). Social Constructivism. *International Encyclopedia of the Social & Behavioral Sciences,* 14264-14267. doi:10.1016/b0-08-043076-7/01086-x

Grabher, G. (2009). Networks, International Encyclopedia of Human Geography (2nd ed., pp. 373-380). Elsevier Ltd.

Himmelreich, J. (2020, April 16). The everyday ethical challenges of self-driving cars. Retrieved October 18, 2020, from https://theconversation.com/the-everyday-ethical-challenges-of-self-driving-cars-92710

Kien, Grant. "Actor-Network Theory." Serious Science, 2 June 2016, Retrieved March 3, 2020, from serious-science.org/actor-network-theory-5973.

Klett, J. (2018, July 20). SCOT. Retrieved October 18, 2020, from https://stsinfrastructures.org/content/scot

L, D. (2007, March 23). Actor-Network Theory (ANT). Retrieved October 18, 2020, from https://www.learning-theories.com/actor-network-theory-ant.html

Maxmen, A. (2018, October 24). Self-driving car dilemmas reveal that moral choices are not universal. Retrieved October 18, 2020, from https://www.nature.com/articles/d41586-018-07135-0

The Rise of Autonomous Vehicles. *The Rise of Autonomous Vehicles | Digital Watch*, 10 Dec. 2020, https://dig.watch/trends/rise-autonomous-vehicles

Roff, H. M. (2019, October 25). The folly of trolleys: Ethical challenges and autonomous vehicles. Retrieved October 18, 2020, from https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/

Snell, R. (2019, April 2). The Rise Of Autonomous Vehicles And Why Ethics Matter. Retrieved from https://www.digitalistmag.com/improving-lives/2019/04/02/rise-of-autonomous-vehicles-why-ethics-matter-06197534/

Social construction of technology (SCOT). (n.d.). Retrieved October 18, 2020, from https://web.archive.org/web/20180410205247/http://www.stswiki.org/index.php?title=Social_construction_of_technology_(SCOT)