**Analyzing Wariness of**

**Autonomous Vehicles**

**Through the Vehicle**

**Decision-Making Process**

(Technical Paper)

**Analyzing Wariness of**

**Autonomous Vehicles**

**Through the Vehicle**

**Decision-Making Process**

(STS Paper)

A Thesis Prospectus Submitted to the

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

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

Quinn Dawkins
Fall 2018

Technical Project Team Members
Jihyeong Lee

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_____
Quinlan Dawkins

Approved_____ Date _____
Hongning Wang, Department of Computer Science

Approved_____ Date _____
Kent Wayland, Department of Engineering and Society

## General Research Problem: The Risks of Algorithmic Authority

*How can we ensure that the machine learning algorithms we use in a growing number of systems are not imposing poorly understood risks?*

The increase in machine learning (ML) applications has come from faster computing resources and improved algorithms. With these improvements, there has been an increase in the complexity of ML algorithms. The best example of this increase in complexity are deep neural networks (DNNs), which are used to approximate solutions to highly complex problems. With increasingly complex problems and models, recent studies have shown strange behavior of DNNs in image recognition tasks when imperceptible changes are made to an image. Because machine learning algorithms rely on being able to infer behavior from data, when data is either tampered with or new to the model, behavior can be unpredictable. Tampering with the input to a data driven algorithm is referred to as an adversarial attack and poses a potential risk when relying on machine learning models to drive important systems. Additionally, adversarial examples do not have to come from tampered data and can occur in more natural data (Zhao, Dua, & Singh, 2018).

As machine learning is applied to more problems, the number of ways this risk can manifest grows. This can be through unexpected behavior when these systems are put in new environments, or through security vulnerabilities from bad actors. There is both a technical approach and social approach to limiting this risk. My research aims to understand both angles. The technical part is investigating adversarial examples, one method by which machine learning algorithms can exhibit unpredictable behavior. The social approach involves understanding how

various social groups are interacting to ensure safety in ML systems. The example I am approaching it from is autonomous vehicles, where error has a significant cost.

## Reinforcement Learning and Adversarial Attacks

Machine learning is a technique of building a program that can infer the relationships between an input and a desired output. This is often done by taking a large data set then feeding it into a learning algorithm that will build an approximate mathematical representation of the data. The subset of machine learning that my research will focus on is the field of reinforcement learning.

Reinforcement learning poses machine learning in the following framework: the model, referred to as the "agent" attempts to take actions in order to maximize the "reward", received from interacting with an "environment". Reinforcement learning is designed to solve problems where rewards are delayed and cumulative, making it effective for problems that require planning for the future. (Arulkumaran, Deisenroth, Brundage, & Bharath, 2017). An example of a reinforcement learning program is AlphaGo, an AI algorithm that beat professional Go players for the first time by a computer in 2016 (Silver et al., 2016). In this example, the environment is the game board, in which the agent AlphaGo can place pieces on the board in order to win the game, which is its final reward. In this game, we can see that the agent does not receive any sort of reward until the game is over.

In general, creating a reinforcement learning algorithm can be framed as attempting to select the best action for the agent, given the state of the current environment, in order to

maximize its expected cumulative reward. Mathematically, the problem can be modeled as a finding a Markov Decision Process, which provides an action given the current state that satisfies the Bellman Optimality Equations (Arulkumaran et al., 2017). The details of these mathematical formulations are beyond the scope of this proposal, but one key conclusion is that the computational complexity required in order to solve these equations is very high. Therefore, there is a need to find approximate solutions, which is commonly performed by models known as neural networks (Arulkumaran et al., 2017).

The neural network is a popular ML algorithm that is roughly modeled after the human brain - information from an input flowing between multiple layers of artificial neurons. Thanks to their complexity, they are powerful enough to approximate the Bellman Optimality Equations for reinforcement learning (Arulkumaran et al., 2017). But despite this power, neural networks are far from perfect – they can be tricked into giving the wrong output. Adversarial machine learning entails ways to "trick" an ML algorithm. In particular, neural networks are one of the most common subjects of these attacks. There are several methods with different effects, ranging from preventing a model from learning the correct relationships or forcing the model to behave undesirably (Liu et al. 2018). One common method is by performing modifications to an input so that they look normal to a human being, but force the model to be incorrect. As shown in Figure 1, Eykholt et al. (2018) present adversarial attacks that could be used to fool image recognition neural networks used to identify road signs.

The two main methods of categorizing adversarial attacks are referred to as white box and black box attacks (Liu et al., 2018). In a white box setting, the adversary is given complete access to the model, whereas in a black box setting, the adversary's access is more restricted. While white

box attacks describe a specific setting, the degree of restriction in a black box attack can vary (Liu et al., 2018).

For reinforcement learning, attacks can take on a slightly different form than that of regular adversarial attacks. This is because attacks on reinforcement learning need to influence the agent to take incorrect actions over a long period of time, which means that the adversary may need to attack the agent multiple times. This is a costly requirement. Two types of white box attacks that address this issue are enchanting attacks and strategically timed attacks (Lin, 2017). Both types examine the model and try to determine a short sequence of adversarial attacks against the agent. In contrast, the attack proposed by Eykholt et al. (2018) in Figure 1 is based on supervised learning, which is a subset of machine learning that's simply based on predicting a label given an input, and thus the adversary only needs to attack once.


Figure 1. Examples of adversarial attacks on image recognition neural networks. The stop signs are interpreted as 45 MPH speed limit signs and the right turn signs are classified as stop signs despite relatively small changes to the appearance of the signs (Image source: Eykholt et al, 2018)

Our research will focus on developing a novel approach to adversarial attacks under a chosen threat model. For developing attacks, we are interested in making attacks more feasible in reality. Possible approaches include limiting the type of states or inputs the adversary is allowed to perturb and an attack where the goal is to prevent the agent from reaching certain states. Examples of each could be where an adversary can only change the appearance of a street sign itself and none of the surroundings, and an adversary trying to hide certain information respectively. So far, we have conducted a survey of adversarial attacks against reinforcement learning, part of which is presented in this proposal. Moving into the next semester we will set up various novel ideas in a breadth of reinforcement learning problems, primarily Atari games.

This should take us into April at which point we can migrate to selecting and refining the most successful approaches.

## Determining Safety Standards for Autonomous Vehicles

*Given current safety standards, who are the different relevant groups in safety standards setting for autonomous vehicles, what are the main challenges, and how does the standards setting process compare against other standards setting processes?*

The rise of autonomous vehicles (AVs) has become a major discussion in both technological and political spheres. As the technology has developed safety concerns become more prevalent. This makes sense in the setting of AVs where error can mean bodily harm, and as a result, proper safety standards are paramount. Understanding how these standards are being set requires understanding the set of stakeholders involved in standards setting and how they are interacting. Furthermore, understanding the process from an established point of reference provides insight into how standards setting should proceed in the context of the challenges facing AVs. This leads into the above question. An answer to this question would provide insight into the future direction of safety standards for AVs. This provides an example of a social approach to protecting against potential risks of algorithmic authority.

### Background

A report from RAND from last year suggests that there is no industry-wide standard definition of safety with regards to AVs (Fraade-Blanar, Blumenthal, Anderson, & Kalra, 2018).

This does not mean work has not been done in this space, however. That same report, as well as another from University of Michigan, make independent attampts to define frameworks for measuring safety of an autonomous vehicle (Peng, & McCarthy, 2019). Furthermore, the U.S. Department of Transportation published a report in late 2018 that describes how communities can prepare for the future of transportation, particularly concerning with AVs ("Preparing for the Future of Transportation: Automated Vehicle 3.0", 2018). Finally, we have seen states begin implementing legislation pertaining to AVs in 2017 and 2018, with twenty-nine states and Washington D.C. passing some sort of legislation ("Autonomous Vehicles | Self-Driving Vehicles Enacted Legislation", 2019).

Looking at the current state of safety standards, in California, 52 companies hold permits to test self-driving cars on roads (Hancock, Nourbakhsh, & Stewart, 2019). Unfortunately, it's hard to gather significant statistical information on the testing that happens on these vehicles. AV field tests still underway have not had statistics published and DMVs do not publish separate crash statistics themselves, making it difficult to build a large enough data set for statistical analysis in an area where total volume of testing is dwarfed by daily human hours driven (Wang, & Li, 2019). A report from this year gives statistics on 113 crashes to try and understand the current safety of these systems (Wang, & Li, 2019). Another problem is that safety testing has minimal standardization outside of the permission and crash reporting process due to the wide variation in degree of automation across autonomous vehicles (Hancock et al., 2019).

Keeping in mind that substantial road-testing statistics can be hard to gather, determing an initial set of relevant social groups, the information they have access to, and the role they play

is important to conducting this research. The first are the companies and engineers manufacturing the vehicles today. They have to work with government to earn permits and determine the degree of automation of the vehicle, among other decisions. For similar reasons experts in machine learning algorithms utilized by AVs are key for providing insight into how autonomous systems functions as they are the ones who understand and design the algorithms. Think tanks and research groups are already joined into the conversation as well. Finally, state governments, specifically Departments of Motor Vehicles, are already involved as is the Department of Transportation through the implementation of initial legislation.

With an initial understanding of the challenges facing standards setting for AVs and initial set of involved social groups, we can formulate a framework for analyzing safety standards for AVs. First, I need to understand the basic process for standards setting in related contexts. This includes understanding how gathering information and implementation works, the degree of input from stakeholders, as well as how long the process takes. Additionally, understanding pitfalls in standards setting is important as well. Next, understanding challenges unique to AVs in determining safety levels is necessary for putting AVs in the context of standards setting. For example, the lack of proper statistics and the varying degrees of automation across the industry are examples of such challenges. Then, I can then frame AVs in the more general problem of standards setting based on the stakeholders and relatedness to other problems. This includes picking problems where the challenges align with that of AVs. This will provide a comparative framework for understanding how the process of standards setting is working for AVs.

**Conducting the Research**

Evidence collection and analysis will work in multiple phases that correspond to the framework I laid out. First, I will collect past safety standards and research information for a few major established sets of safety standards. Specifically, Federal Motor Vehicle Safety Standards and Regulations (FMVSS), standards set by the Occupational Safety and Health Administration (OSHA), and Consumer Product Safety Commission (CPSC) standards. Additionally, I will collect any available STS research on safety standards to aid in analyzing these standards sets on the above framework. I will then collect further industry and academic information regarding the challenges facing self-driving cars. This will partly stem from limited information provided directly by leading companies in the space such as Tesla and Uber, but primarily come from academic documents found in conferences and technical surveys. Finally, collecting documents provided by both the identified stakeholders and any unidentified stakeholders revealed during the prior document collection phases will be necessary for constructing the final comparison.

To properly analyze the collected documents, I will start with the standards documents. I will analyze these documents in an Actor Network Theory framework to identify relevant actors and technologies so I can draw comparisons at the end. Any STS research I collect may not fit into this analysis cleanly and I will instead draw out common themes using qualitative data analysis. To analyze the academic information on the challenges facing AVs I will use qualitative data analysis to draw out common problems that are identified. Finally, I will use the understanding of each standards setting problems to construct parallels regarding actors, challenges, and the relevant technologies to match effective and ineffective solutions based on

pitfalls highlighted in any STS research. This will provide the final comparative framework for analyzing standards setting for AVs.

I recognize that the scope of this framework could be extremely wide. I laid out the general framework, document collection process, and analysis process for understanding standards setting for AVs, but instead I will be restricting the scope of how I analyze AVs. I will restrict myself to analyzing self-driving cars and specifically the part of the system that directly pertains to automation. This will restrict the amount of document collection and analysis needed at each step and better focus this research in the context of the overarching problem. Put together, this research will help better understand how standards setting for AVs should act to limit the risk imposed by ML systems.

## Conclusion

Risks imposed by the use of machine learning algorithms has both a technical and social dimension. Adversarial attacks are one way to cause unpredictable behavior in ML algorithms, but a better understanding of how a society should utilize these systems is directly tied to this risk as well. Understanding both how these vulnerabilities arise and how to defend against them aid in preventing poor decision making by the algorithm. Reinforcement learning is a framework where adversarial attacks are less studied so this research is ever more relevant. Autonomous vehicles are an example of a modern application of ML with significant potential risks in the event of failure, so comparing how safety standards are built for AVs provides an example of a social approach.

## References

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6), 26-38.

National Highway Traffic Safety Administration., U.S. Department of Transportation., (2017). Automated Driving Systems: Voluntary Safety Self-Assessments; Public Workshop. Retrieved from Federal Register website: https://www.federalregister.gov/documents/2017/10/17/2017-22058/automated-driving-systems-voluntary-safety-self-assessments-public-workshop

National Conference on State Legislatures., Autonomous Vehicles | Self-Driving Vehicles Enacted Legislation. (n.d.). Retrieved from http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ..., Song, D. (2018). Robust Physical-World Attacks on Deep Learning Models. *Conference on Computer*

*Vision and Pattern Recognition.*

Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). Measuring Automated Vehicle Safety [RR-2662]. Retrieved from https://www.rand.org/pubs/research_reports/RR2662.html

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences of the United States of America*, 116(16), 7684–7691. https://doi.org/10.1073/pnas.1805770115

Lin, Y. C., Hong, Z. W., Liao, Y. H., Shih, M. L., Liu, M. Y., & Sun, M. (2017). Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. *International Joint Conference on Artificial Intelligence.*

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). A Survey on Security Threats and Defensive Techniques of Machine Learning: A data driven view. *IEEE Access*, 6, 12103-12117.

Peng, H., & McCarthy, R. L., (2019). A Concept to Assess the Safety Performance of Highly Automated Vehicles [Whitepaper]. 15. Retrieved from https://mcity.umich.edu/wp-content/uploads/2019/01/mcity-whitepaper-ABC-test.pdf

Preparing for the Future of Transportation: Automated Vehicle 3.0 [Text]. (2018, September 27). Retrieved from US Department of Transportation website: https://www.transportation.gov/av/3

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587), 484.

Wang, S., & Li, Z. (2019). Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. *PLoS ONE*, 14(3). https://doi.org/10.1371/journal.pone.0214550

Zhao, Z., Dua, D., & Singh, S. (2018). Generating Natural Adversarial Examples. *International Conference on Learning Representations*