Mitigating Harmful Machine Learning Dependencies

A 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

Gregory Victor Vavoso

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 _____
Gregory Victor Vavoso

Approved _____ Date _____
Sean Ferguson, Department of Engineering and Society

Mitigating Harmful Machine Learning Dependencies

Machine learning systems are increasingly becoming a part of everyday human life. Each year, we become more dependent on machine learning models to advise us. For example, many individuals are highly dependent on navigation apps to get from point A to point B. As of 2018, nearly 77% of smartphone users regularly depend on navigation apps to travel (Panko, 2018). This drastic increase in dependence may be cause for concern (Anderson & Rainie, 2018). This is because machine learning models are designed to react to small changes in human behavior. However, most models are only trained on past data. Models begin to fall apart when global human behavior takes a sudden turn.

The COVID-19 pandemic has caused a sudden change in human behavior, especially traveling tendencies. Evidence has shown that many software applications have been affected by this spontaneous change in human behavior. Specifically, many artificial intelligence and machine learning applications have been unable to handle this sudden change in human behavior. For example, Amazon's previous top searches were replaced with COVID-19 related products such as face masks, hand soap, and cleaning wipes, causing previous models to perform in unexpected ways (Heaven, 2020). While Amazon shopping suggestions may not seem like a dire cause for concern, there are machine learning models that are life critical, such as traffic models for self-driving cars. For example, the Google Maps team had to completely overhaul their traffic models, as they were unable to rely on old data that did not represent the new pandemic behavior (Quach, 2020).

When it comes to machine learning models amidst the pandemic, humanity was fortunate to have the ability to swiftly adjust parameters and models to return to accurate prediction

making. However, as machine learning model complexity arises and machine learning systems begin to enter more life critical applications, this fault in machine learning models could be far less forgiving.

Understanding Machine Learning Dependence

Dependence on a technology is not always a cause for concern. However, when a human becomes heavily dependent on a technology and loses their ability to act as a backup mechanism for this technology, there becomes cause for concern. This is why the concept of scripting in a normative setting will be crucial when engineers consider the ethical implications of their design (Verbeek, 2006). The concept of a "script", devised by Madeleine Akrich and Bruno Latour, describes the ways in which technologies "prescribe" actions between two actors. For example, a speed bump prescribes that the driver slows down. Technologies have numerous scripts and all must be considered (Verbeek, 2006). Thus, this paper will discuss the ways in which machine learning systems prescribe harmful dependencies upon human actors and how to script these technologies to prevent catastrophic failure.

Existing Dependencies

As technologies develop and aid humans in their day-to-day tasks, dependence occurs on these technologies. The usage of GPS devices and applications in car navigation is a prime example of a dependence. Up until the early 2000s, taxi drivers would have immense knowledge of city layouts, so much so that the hippocampus portion of their brain increased in volume proportional to their taxi driving experience (Maguire, et al., 2000). Now, Uber and Lyft drivers must avoid deviating from their assigned route, otherwise the app sends out a notification of the deviance to a safety representative. Thus, it is rare to see an Uber or Lyft driver not heavily depending on their smartphone to get to the passenger's destination.

This dependence is not the direct cause for concern, however. The loss of human autonomy in driving ability as a result of this dependence is where issues arise. For example, a heavily dependent Uber driver may not have the ability to successfully deliver their passenger to their destination if their navigation were to fail. Without being governed by the directions provided by this system, the driver has lost their ability to fulfill their job independently.

One study recognized a potential overdependence in the medical field. They realized that failures in computerized provider order entry (CPOE) and other systems were being overly depended on. They studied five different hospitals and held a conference of medical experts and acknowledged three major concerns: 1) Lack of backups lead to major inefficiencies in the case of CPOE failure, 2) the users had false expectations of the accuracy of their entered data, and 3) some clinicians were simply unable to function without the assistance of CPOE (Campbell, Sittig, Guappone, Dykstra, & Ash, 2007). This study provided valuable insight into recognizing overdependence, as well as some methods of mitigating it. Despite not formally meeting the definition of a machine learning system, this study on CPOE overdependence will help to understand the shifting actor-networks discussed in later sections.

Dependence on machine learning networks is another cause for concern, especially in the case of deep learning systems. Expressed quite simply, “In order for deep learning and similar AI algorithms to serve the purposes that we want them to serve by design, they necessarily tend to become more epistemically opaque to us, thus stymieing interpretability, communicability, and transparence” (Long, 2020). Deep learning systems can reach the point of complexity where it is infeasible for humans to attempt to understand them. As these deep learning systems internalize extreme amounts of data, they become so sophisticated that the creators of these black-boxed systems rely on empirical testing rather than underlying theory. When the creator of a black-

boxed system loses understanding of the theory behind their creation, it becomes nearly impossible to troubleshoot in the case of failure (Bathaee, 2018).

Modes of Failure

There are many predictable and unpredictable ways that complex machine learning systems can fail. In the case of this research, a primary and a secondary cause of machine learning failure will be considered: sudden shifts in human behavior and deliberate attacks of machine learning systems (Thomas, Norton, Jones, Hopper, & Ward, 2011). The similarity between these two cases is simple: one or more of the inputs to the model is no longer reliable (Steinhardt & Toner, 2020).

Sudden Shifts in Human Behavior

As previously mentioned, one of the most significant and modern instances of sudden shifts in human behavior occurred in 2019 with the COVID-19 pandemic. These machine learning systems that drive navigation systems, shopping recommendations, and other day-to-day applications are based on years of training data. These systems only perform when the behaviors of humans are within their bounds of the training data. However, when traffic drops by upwards of 47.5 percent in some areas amidst a pandemic, for example, machine learning systems not only fail, but can act unpredictably and dangerously (Elejalde-Ruiz, 2020).

Fortunately, the Google Maps team understood the theory and intricacies behind their machine learning models. Within a matter of days, they were able to diagnose the issue and return their navigation systems back to an acceptable state (Quach, 2020). However, as machine learning models become more and more complex, the transparency of the innerworkings of models become less and less clear. Neural networks may one day be as complex as the human brain. Attempting to understand and troubleshoot the innerworkings of such a complexity to

solve a sudden failure could be futile.

Deliberate Attacks

One of the first notable attacks on a machine learning system occurred in 2013 when Syrian hackers were able to compromise and post a tweet on the Associated Press Twitter that claimed two explosions occurred in the white house, injuring president Obama (Fisher, 2013). The Tweet was crafted in such a way that Twitter's algorithm rated it as a credible tweet. Nearly a minute later, the Dow Jones dropped nearly 150 points. This also exemplifies the compounding effect one system can have on another.

Another example of deliberate attacks causing catastrophic failure occurred in a lab setting. A computer vision research team was testing the ability of a self-driving machine learning system to read stop signs. In one case, they strategically placed small black stickers on the stop sign. To human eyes, they were hardly noticeable. However, in the case of the self-driving system, it interpreted it as a 45-mile-per-hour speed-limit sign (Hancock & Nourbakhsh, 2018). As Hancock concludes, "[machine learning systems] will succeed in ways that are not human, and they will also fail in ways that have nothing to do with how we fail" (Hancock & Nourbakhsh, 2018). If we do not even know the general realm of possible failures for these systems, how can we place our trust and dependence on them?

A Survey of Crowdsensing Human-Machine Networks (HMN)

In order to conduct this analysis, an actor-network theory (ANT) approach will be taken to analyze the human-machine networks (HMNs) involved in this pandemic (Tsvetkova, et al., 2017). The concepts of scripts will also be used in conjunction with ANT to assess how changes in HMNs change the actions required of humans and machines. Tsvetkova et al. (2017)

introduces and explores the various types of human-machine networks and will later be used for classification of different network types. The two overarching actors in consideration are humans and machines. The human users will be the primary stakeholders, specifically the humans receiving critical information from machine learning systems. For simplicity of analysis, humans can be an individual person or groups of individuals. In most cases, machines will refer to a single machine learning system, such as Google Maps.

Human to machine interaction will be classified as either passive contribution or active contribution. For example, Google Maps users providing location data to Google would be a passive contribution. Passive contributions can be seen as a "background process" that humans are not necessarily aware of. Machine to human contributions will mainly be defined as active contributions, such as Google Maps giving a user specific direction.

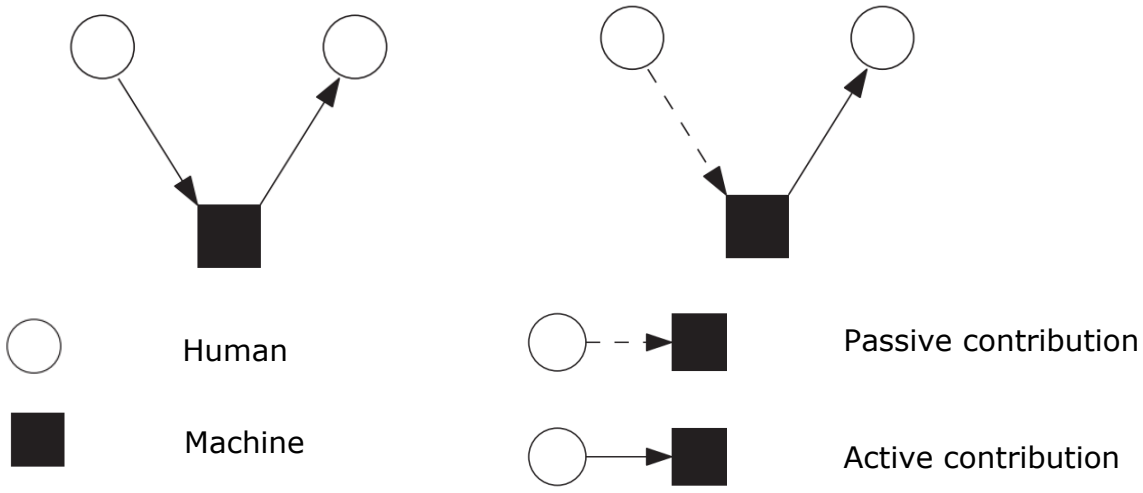

Figure 1. Two variants of crowdsensing human-machine network

Thus, the specific analysis will be centered around "crowdsensing" human-machine networks (Tsvetkova, et al., 2017). This crowdsensing form of HMN is shown in Figure 1. Human contribution in a crowdsensing network can be either passive or active, as shown by the

two networks in Figure 1. The left network shows active contribution from humans, while the right network shows passive contribution from humans. Machine to human contributions tend to be active contributions in crowdsensing networks.

In order to ground analysis in a real-world and life critical application, traffic patterns and self-driving artificial intelligence will be used as the primary crowdsensing network in consideration (Kellner, 2019). However, other applications will be mentioned to show the broad scope of this issue. The crowdsensing network will act as the core for a more general analysis. However, there are more actors to discuss for these self-driving networks (e.g. machine learning model designers, pedestrians, and policymakers). These actors specific to self-driving networks will be considered in the coming sections.

Existing Self-Driving Networks

Self-driving vehicles are becoming a reality, and fully autonomous vehicles are almost certainly becoming a reality in the coming decades. The process towards full automation is not entirely new. Prior to 2000, antilock brakes and cruise control were the beginning of driver-aid. From 2000 to 2010, new features were added such as blind spot detection, lane departure warning, and forward collision warnings. The next decade saw even more advances such as automatic emergency braking, lane-keeping assist, and even some basic forms of highway autopilot (Automated Vehicles for Safety, 2020). With all of these features came increased human dependence. Some individuals have begun to rely on their rearview systems for parking. With these increases in dependence comes decreases in human autonomy. Individuals lose practice with using mirrors as a backup because of their heavy dependence on rearview cameras. This becomes a concern when life-critical automation systems experience failure.

In terms of the pre-2000 human machine networks, some of the most notable actors were drivers, cars, pedestrians, and pedestrians. The drivers were the main force of this network, with almost total control over their vehicle. The vehicle designers had a minimal say in how the vehicle performed under certain situations, such as activated the anti-lock braking system. If the driver were to hit a pedestrian, it is almost entirely the fault of the driver in any circumstance. Drivers must respect other drivers, but ultimately one driver will only effect a few nearby surrounding neighbors on the highway or road. Drivers, passengers, and pedestrians all have a strong dependency on the driver to be vigilant and safe with their driving skills, as their lives depend on it. While the driver does depend on a vehicle manufacturer to design a structurally sound vehicle, it is also the responsibility of the driver to maintain and inspect their vehicle regularly. Therefore, there is minimal dependence on the vehicle designer in this network. The dependence is placed heavily placed on the driver.

As of 2020, drivers are still the main dependency in this actor-network. More and more self-driving cars are on the road, but most of the populace has yet to own, or even drive in such a vehicle (Hancock & Nourbakhsh, 2018). All nodes in this actor-network still lead back to the driver as the main bearer of this dependence. Simple safety features like emergency auto-braking can fail in cars now, begging the question, who was responsible for the failure? This modern-day situation is beginning to show the shift in nodal connections in this actor network: the change from driver dependence to manufacturer dependence.

A new actor has appeared in the network in the last few decades: GPS and navigation apps. This is where the concept of a crowdsensing HMN becomes a major factor in the analysis. Drivers are increasingly becoming dependent on navigation systems, such as Google Maps or Apple Maps (Thomas, Norton, Jones, Hopper, & Ward, 2011). This creates a strong nodal

connection between the driver and the technology company. The tech company relies on massive amounts of driving data to feed their models and suggest changes in routes. Failure of this new node is not yet life critical, but it could leave drivers down wrong roads and into potentially dangerous traffic.

Failure of most of these systems, whether deliberate or not, are likely to be backed up by the driver. A failure in Google Maps simply means the driver either figures out the path to their destination or simply pulls over. The worst-case failure is not life critical. For the most part, the life-critical failures all remain within the node of the driver. A catastrophic failure from the driver to another node is likely to be salvageable. This is likely to drastically change in the next decade with drastic increases in autonomous driving systems.

Predicted Futures of Self-Driving Networks

By 2025, the National Highway Traffic Safety Administration (NHTSA) fully expects "autopilot" capabilities to be rolled out to most Americans (Automated Vehicles for Safety, 2020). The rampant competition among Tesla, Google, MIT, BMW, and other major car companies is a testament to how much these companies, as well as the world, desire autonomous driving technology. The fact that 94% of crashes are due to human error alone is almost enough to justify fully autonomous vehicles (Automated Vehicles for Safety, 2020). However, this "want-it-now" mentality can lead designers to accelerate quickly and create potentially dangerous scripts associated with this technology.

In the near future, humanity is likely to see a strong symbiosis between driver and vehicle. That weak dependency will begin to shift such that vehicle will back up driver decisions,

while driver will back up vehicle decisions. This period of symbiosis between vehicle and driver can be a period of major concern. Vigilance must now be shared. Prior to the 21st century, nearly 100% vigilance was expected of the driver. Now, however, 100% vigilance is expected to be shared between vehicle and driver, thus strengthening this nodal connection. In many cases, vigilance must be shared between driver and vehicle for purposes of redundancy.

Additionally, as humans lose their autonomy to vehicles, they lose their ability to act as reliable backup systems in the case of failure. Now, humans constantly practice driving. However, as humans slowly take their hands off the wheel in the coming years, driving practice decreases and humans become an unreliable backup mechanism. This is why the in between period where dependence is shared between vehicle and driver is the period of most concern, especially towards the period of heavy vehicle dependence.

Synthesizing Commonalities Among Overly Dependent Networks

Commonalities Among Dependent HMNs

The aforementioned CPOE dependence shows the perfect example of a case where the nodal connection between human and machine was uneven and unstable. The computerized systems had a relatively high probability of failing with often no backup or highly inefficient backup. It is essential to have a functioning and relatively efficient backup to a computerized system in place. No machine learning system, even self-driving cars, can expect to be 100 percent free of risk.

The second aspect the researchers discovered on the CPOE system was an overestimation of the system's ability to make decisions about patients. Assumptions were made about the technology that simply were not true. Computerized systems have not reached the point of being able to judge and diagnose patients. Understandable metrics must be created such that practitioners understand exactly what the capabilities of the system are. Not only that, but the creators of the system must be experts in the system and have the ability to convey its abilities to the end-user. This draws many parallels to the black-box of deep learning algorithms, where transparency is not always there.

Another important situation to discuss is the failure of multiple navigation systems amidst the pandemic. When sudden behavioral changes occurred and traffic volume dropped by upwards of 50%, navigation systems temporarily failed (Quach, 2020). This was a full system failure and a sufficient backup was not present, as seen by the multiple day downtime. However, due to the transparency of the machine learning algorithm, the system was understandable and fixable. However, a black-boxed system without backup could have led to far worse consequences.

The Harmful Dependencies

As machine learning systems begin to provide more and more benefits to individuals, humans will continue to depend on them for day-to-day activities. If a car can offer an individual the ability to take their hands of the wheel and relax while experiencing increased safety, consumers will jump to it. Almost any technology goes through a shift where humans offer autonomy to a machine. The question becomes, what novel aspects of machine learning systems prescribe the most dangerous forms of human-machine dependence?

The first and commonly recurring dependence is the black-box scenario. When a machine learning system becomes unknown to the consumer, they create false expectations on its ability. Just as clinicians overestimated the ability of their IT systems, individuals may overestimate the ability of their vehicle in certain driving situations. The connection between human and machine learning system must be as transparent as possible. The individual should have a clear metric for judging the ability of a machine learning system. However, machine learning brings a novel issue to the table: the developers of a system may not fully understand what they have designed, especially in the case of deep learning systems, where complexity is immense. If designers of a black-boxed machine learning system are unable to understand the limitations of their predictions, there is no way for any of the other actors in the human-machine network to understand its capabilities. Thus, the first precaution becomes clear: systems which entirely rely on a black-boxed design should be avoided in high-stakes single decision systems (Steinhardt & Toner, 2020).

The second dangerous dependency arises when a backup system is not immediately available. In the case of Google Map's route prediction system, there was no immediate backup available. Users dealt with misdirection for days on end. The situation of backup becomes increasingly complex as machines take the brunt of responsibility and dependence in this network. As dependence shifts towards self-driving vehicles, humans become out of practice in driving. If drivers are the backup mechanism in this case, they are likely to be unpracticed and unreliable, especially in the later evolution of self-driving vehicles.

The third and final major precaution involves detection of failure. Designers of these systems must "script" their technologies such that failures are detectible. The programmers and engineers who develop these machine learning systems must anticipate and actively inscribe detection mechanisms (Verbeek, 2006). Clear operating thresholds should be established for every output to ensure the system is operating under proper conditions. The longer the system continues to run under failure, the more damage and potential loss of life incurred.

Conclusion: The Fundamental Set of Precautions

Dependence on machine-learning systems is necessary and inevitable given the momentum of current systems. As we shift our dependence away from us and towards these machine learning systems, three fundamental precautions should be kept in mind: 1) The ability and innerworkings of a machine learning system must be entirely transparent to both the designer and the consumer, 2) an immediate backup system must be in place that is robust enough to guarantee full system function or guarantee a successful emergency "landing", and 3) detection of failure must be caught before machine learning outputs go awry as to assure a smooth transition from failure to backup situation. With these three precautions in mind, a smoother and

safer transition to machine learning dependence is possible, thus a more comfortable and quicker movement towards better lives.

References

Anderson, J., & Rainie, L. (2018). Artificial Intelligence and the Future of Humans. *Pew Research Center*.

Automated Vehicles for Safety. (2020). Washington, DC, United States: National Highway Traffic Safety Administration.

Bathaee, Y. (2018, Spring). THE ARTIFICIAL INTELLIGENCE BLACK BOX AND THE. Cambridge, Massachusetts, United States.

Campbell, E., Sittig, D., Guappone, K., Dykstra, R., & Ash, J. (2007). Overdependence on Technology: An Unintended Adverse Consequence of Computerized Provider Order Entry. National Center for Biotechnology Information.

COVIDWISE. (2020). Retrieved from Virginia Department of Health: https://www.vdh.virginia.gov/covidwise/

Elejalde-Ruiz, A. (2020, May 4). If you get sick with COVID-19, is your employer liable? As businesses prepare to reopen, worker safety is a priority. Chicago, Illinois, United States.

Fisher, M. (2013, April 23). Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism? Washington D.C., United States: The Washington Post.

FreeRTOS. (2020). Retrieved from FreeRTOS: https://www.freertos.org/

Hancock, P., & Nourbakhsh, I. S. (2018, April 16). On the future of transportation in an era of automated and autonomous vehicles. Pittsburg, Pennsylvania, United States: Proceedings of the National Academy of Sciences of the United States of America.

Heaven, W. D. (2020). Our weird behavior during the pandemic is messing with AI models. *MIT Technology Review*.

Hosanagar, K., & Cronk, I. (2018, October). Why We Don't Trust Driverless Cars - Even When We Should. Cambridge, Massachusetts: Harvard Business Review.

Kellner, L. (2019). Machine Learning Algorithms Help Predict Traffic Headaches. *Berkely Lab*.

Long, B. (2020). The Ethics of Deep Learning AI and the Epistemic Opacity Dilemma. APA Online.

Maguire, E., Gadian, D., Johnsrude, I., Good, C., Ashburner, J., Frackwiak, R., & Frith, C. (2000). Navigation-related structural change in the hippocampi of taxi drivers. Montreal, Canada: Proceedings of the National Academy of Sciences of the United States of America.

Morrissey, J. (2020, June 16). Fighting the Coronavirus With Innovative Tech. *The New York Times*.

Panko, R. (2018). The Popularity of Google Maps: Trends in Navigation Apps in 2018. Wasington, D.C., United States.

Quach, K. (2020). Google declares Maps COVID-19-ready after retraining it on pandemic traffic – or the lack of it in some areas. *The Register*. Retrieved from https://www.theregister.com/2020/09/03/google_maps_covid/

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition. MIT Press.

Steinhardt, J., & Toner, H. (2020, June 8). Why Robustness is Key to Deploying AI. Brookings Institution.

Thomas, M., Norton, J., Jones, A., Hopper, A., & Ward, N. (2011, March). Global Navigation Space Systems: reliance and vulnerabilities. London, United Kingdom.

Tsvetkova, M. (2017). Understanding Human-Machine Networks: A Cross-Disciplinary. *ACM Journals, 50*.

Tsvetkova, M., Yasseri, T., Meyer, E. T., Pickering, J. B., Engen, V., Walland, P., . . . Bravos, G. (2017). Understanding Human-Machine Networks: A Cross-Disciplinary. *ACM Journals, 50*.

Verbeek, P.-P. (2006, May). Materializing Morality: Design Ethics and Technological Mediation. Enschede, Netherlands.