Safety Issues of Black Box Models in Autonomous Systems

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

Shenghui Chen
Fall, 2020

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 Shenghui Chen Date 2020/10/28
Shenghui Chen

Approved Richard Jacques Date 02 Nov 20
Richard Jacques, Department of Engineering and Society

Safety Issues of Black Box Models in Autonomous Systems

Introduction

In recent years, dramatic success in machine learning has led to a torrent of Artificial Intelligence (AI) applications. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. Notable examples include autonomous driving, drones, medical assistive technologies. However, as these applications show great potential for improving the quality of life and more convenience for mankind, one also has to recognize the risks it brings.

These advances in machine learning, especially after the success of ImageNet and AlphaGo, have led to a widespread belief that most accurate models for any given data science problem must be inherently uninterpretable and complicated – a black box. This belief stems from the historical use of machine learning in society: its modern techniques were born and bred for low-stakes decisions such as online advertising and web search where individual decisions do not deeply affect human lives (Rudin and Radin, 2019).

However, with the great progress in artificial intelligence and related domains, it is expected that such powerful technologies are to be used in building more important and safety-critical systems such as autonomous cars, robotic assistants, and personalized medicine. By relying on sophisticated machine learning models trained on massive datasets thanks to scalable, high-performance infrastructures, we risk to create and use decision systems that we do not really understand (Guidotti et al., 2018). This brings risk in various ways: wrong decisions made could compromise the safety of humans, possible biases in historical training data can lead to

unfair decisions. All of which cast doubts on whether these autonomous systems can bring society the welfare and benefits they are designed to deliver.

A series of sporadic accidents in these new applications seem to validate this concern. For example, the first recorded pedestrian fatality involving a self-driving car happened on March 18, 2018, when an autonomous car operated by Uber during real-world testing with a human emergency driver behind the wheel, struck and killed Elaine Herzberg (Schmelzer, 2019). It is reported that the fatal accident came as a result of the automated Uber's not having "the capability to classify an object as a pedestrian unless that object was near a crosswalk," according to one of the documents released by the National Traffic Safety Board. The documents also said the safety driver was working alone, streaming a television show, and didn't keep her eyes on the road (McCausland, 2019). This tragic incident revealed the fatal outcome an immature system can bring to our society and caused a new wave of scrutiny on verifying safety in these autonomous systems. This example also shows that the danger of black-box models does not only lie within itself, but is also closely related to how society, organizations, and the people using the systems deal with a system that is not fully understood. Moreover, how should different parties make an effort to prevent similar instances from happening in the future? While these systems are not so ingrained in our daily lives, how should we tailor the ship in the right direction, and not let the black boxes form a world that is too complicated to tell?

STS Framework: Technological momentum

A fitting STS theory here is the theory of technological momentum proposed by the historian of technology Thomas P. Hughes. This is a theory about the relationship between technology and society over time. In its essence, Hughes's theory is a synthesis of two models,

technological determinism and social determinism, for how technology and society interact. Technological determinism claims that society itself is modified by the introduction of new technology in an irreversible and irreparable way, and technology, under this model, self-propagates as well—there is no turning back once the adoption has taken place, and the very existence of the technology means that it will continue to exist in the future. On the other hand, social determinism claims that society itself controls how technology is used and developed. Technological momentum tries to unify the two models by adding time. In Hughes's theory, when technology is young, called ***phase one***, deliberate control over its use and scope is possible and enacted by society. However, as the technology matures, and becomes increasingly enmeshed in the society where it was created, it enters ***phase two***. In *phase two*, its own deterministic force takes hold, achieving technological momentum in the process. According to Hughes this inertia, which is particularly the case for large technological systems with their technological and social components, makes them difficult to influence and steer as they start to go more on their own way, assuming deterministic traits in the process (Vermaas et al., 2011). In other words, Hughes says that the relationship between technology and society always starts with a social determinism model, but evolves into a form of technological determinism over time and as its use becomes more prevalent and important.

Applying Hughes' theory to the case of autonomous systems with black-box models, I believe we are still in the middle of *phase one*, and many people have already realized the importance of addressing the existing problems, and are contributing to a safer, more reliable, and more ethical future. However, we have to realize that there are also opposite forces at play that, regardless of motivation, failed to put resources into this pursuit and are wasting the time we left in *phase one*. For example, the companies that profit from holding the intellectual

property of COMPAS, a recidivism risk prediction tool that is unnecessarily black-box (which will be discussed in more depth in following sections). There are already some signs suggesting our society will transition into *phase two* in the near future, and if we wait until then, it is possible that society needs to pay a much higher price in correcting the same problems.

Two types of black box models

With the broad use of complex decision-making systems in different domains, black-box models seem like a loaded term these days. An algorithm, a financial model, or a transistor, can all be called a black-box model. Therefore, it is necessary to properly define the terms discussed in this paper. The standard definition of black-box models commonly used in science, computing, and engineering, is a device, system, or object which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings. Its implementation is opaque or "black" (Kenton, 2020). From my observation, there are two types of black-box models, the first is a ***technically black-box model***, and the other is a ***socially black-box model***.

The *technically black-box model* is the ones that even its *developers* do not know or understand its inner workings. A common example is the various deep neural network models used in computer vision or some other big data analytic tasks. They may have an unexpected performance after the training of the model occurs, but even their developers cannot accurately explain why one works but the other does not. This is certainly unideal because it compromises the reliability and trustworthiness of the system. Given the nature of the task, the duty of unpacking these black boxes naturally lies with the scientists in artificial intelligence and other related fields.

The *socially black-box model*, however, describes a system where its internal logic is unknown to its *users*. This could be a good thing. Sometimes the developers of a system deliberately make a component black-box to its users in order to decrease the complexity of the task and increase accessibility. For instance, the designers of Graphical User Interfaces use graphical icons to allow users to interact with electronic devices in more effortless ways and hide the text-based command lines behind the black box. Although students in CS majors are still required to take Operating System classes to learn the inner logic behind it, no one really complains about the ease of using computers and phones by this encapsulation. However, this could also lead to undesirable consequences. A noticeable example is the crashes of Boeing 737 max, which I will explain in more depth in Case Study 2. It seems like the line distinguishing good socially black-box models from bad ones is whether being black-box is intentionally designed by the developers.

Given the wide range of application domains that black box models can be applied for, I will introduce the problems, in particular the safety issues, an analysis of possible causes in two case studies, which will provide more context and substance in the discussion.

Case study 1: Autonomous cars fooled by phantom images

Figure 1 In the Ben-Gurion University of the Negev Research, Telsa considers the phantom image (left) as a real person and (right) Mobileye 630 PRO autonomous vehicle system considers the image projected on a tree as a real road sign. Credit: (Nassi, et al., 2020)

In a new research paper, "Phantom of the ADAS," Nassi et al. demonstrated that autopilots and advanced driving-assistance systems (ADASs) in semi-autonomous or fully autonomous cars register depthless projections of objects (phantoms) as real objects. They show how attackers can exploit this perceptual challenge to manipulate the vehicle and potentially harm the driver or passengers without any special expertise by using a commercial drone and inexpensive image projector.

While fully and semi-autonomous cars are already being deployed around the world, vehicular communication systems that connect the car with other cars, pedestrians, and surrounding infrastructure are lagging. According to the researchers, the lack of such systems creates a "validation gap," which prevents autonomous vehicles from validating their virtual perception with a third party, relying only on internal sensors.

The researchers showed that the car industry has not considered this type of attack by demonstrating the attack on today's most advanced ADAS and autopilot technologies: Mobileye

630 PRO and the Tesla Model X, HW 2.5; the experiments showed that when presented with various phantoms, a car's ADAS or autopilot considers the phantoms as real objects, causing these systems to trigger the brakes, steer into the lane of oncoming traffic, and issue notifications about fake road signs.

To mitigate this attack, the researchers are developing a neural network model that analyzes a detected object's context, surface, and reflected light, which is capable of detecting phantoms with high accuracy. (Nassi et al, 2020)

While authors of this work attributed the risk to the lack of a validation infrastructure, I would also argue that the deficiency in explanation ability of the system also comes into play. If some real-time explanation of the object detection process is presented to the operator, he or she may be able to intervene in the decision process and add human judgment to it.

This case study brings out the first problem common in technically black-box systems, that users are not given an option to intervene early when the autonomous system made an erroneous decision. Why would people allow such loopholes to exist and continue to exist in such safety-critical systems such as driverless cars? Two common beliefs are presumed here. By the nature of models such as DNNs, people usually have the consensus that there are too many variables and parameters in the model for any users to process. Also, it is believed that there is an inherent trade-off between interpretability and accuracy.

However, Rudin and Radin have argued against the second belief in their article "*Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition*", which was published in *Harvard Data Science Review* in 2019, "The belief that accuracy must be sacrificed for interpretability is inaccurate. It has allowed companies to market and sell proprietary or complicated black box models for high-stakes decisions when very simple

interpretable models exist for the same tasks. As such, it allows the model creators to profit without considering harmful consequences to the affected individuals. Few questions these models because their designers claim the models need to be complicated in order to be accurate." They are quite blunt in stating that "Being asked to choose an accurate machine or an understandable human is a false dichotomy", and this is important because "understanding it as such helps us to diagnose the problems that have resulted from the use of black box models for high-stakes decisions throughout society. These problems exist in finance, but also in healthcare, criminal justice, and beyond." Rudin and Radin then go on to support their claims by giving evidence that this assumption is wrong. They show that in criminal justice systems, complicated black-box models for predicting future arrest are not any more accurate than very simple predictive models based on age and criminal history, demonstrated repeated in different works (Angelino et al., 2018; Tollenaar & van der Heijden, 2013; Zeng, Ustun, & Rudin, 2016). For instance, an interpretable machine learning model for predicting rearrest created in work by Angelino et al., considers only a few rules about someone's age and criminal history, but they work as accurately as the widely used (and proprietary) black-box model called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), used in Broward County, Florida (Angelino et al., 2018). Similar cases are found in several healthcare domains and across many other high-stake machine learning applications where life-altering decisions are being made. In all these examples, we do not see evidence of a benefit from using black box models instead of simpler interpretable ones, on the contrary, the black box models can mask a myriad of possible serious mistakes, e.g. systematic errors in the dataset, data leakage.

Besides carefully considering whether a technically black-box model is needed in constructing the next critical system, we should also be careful not to build a socially black box

model unintentionally. This leads the discussion to the next case study: the Boeing 737 Max disaster.

Case study 2: Boeing 737 Max Crashes

On October 29, 2018, a 737 MAX 8 operating Lion Air Flight 610 crashed after take-off from Jakarta, killing all 189 on-board. About a half year later, another 737 MAX 8 operating Ethiopian Airlines Flight 302 crashed shortly after take-off from Addis Ababa airport, killing all 157 on-board. Ethiopian Airlines immediately grounded its remaining MAX fleet. On March 11, the Civil Aviation Administration of China ordered the first nationwide grounding, followed by most other aviation authorities in quick succession. on March 13 after receiving evidence of accident similarities, the Federal Aviation Administration (FAA) finally grounded 737 Max, marking an effective worldwide grounding for this aircraft.

What caused this tragedy? Software executive and pilot Gregory Travis gave his analysis in an article on IEEE Spectrum called "How the Boeing 737 Max disaster looks to a software developer". Since the 737 first appeared in 1967, market and technological forces pushed the 737 into ever-larger versions with increasing electronic and mechanical complexity. The most effective way to make an engine use less fuel per unit of power produced is to make it larger, but this causes a problem: The original 737 had smaller engines, which easily cleared the ground

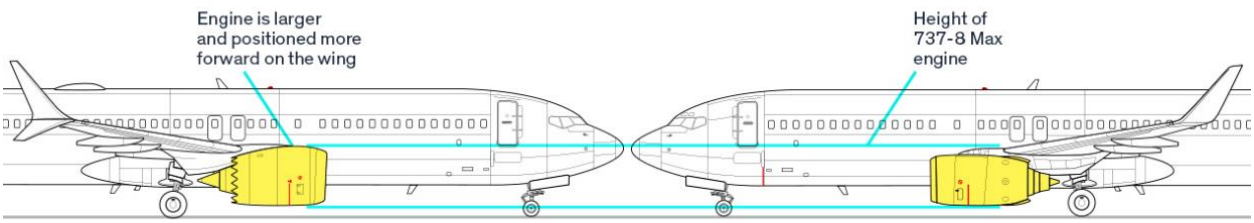

Figure 2 By substituting a larger engine, Boeing changed the intrinsic aerodynamic nature of the 737 airliner. Credit: Norebbo.com

beneath the wings. As the 737 grew and was fitted with bigger engines, the clearance between the engines and the ground started to get a little tight.

The solution was to extend the engine up and well in front of the wing. However, doing so also meant that the centerline of the engine's thrust changed. Now, when the pilots applied power to the engine, the aircraft would have a significant propensity to "pitch up," or raise its nose. Apparently the 737 Max pitched up a bit too much for comfort on power application as

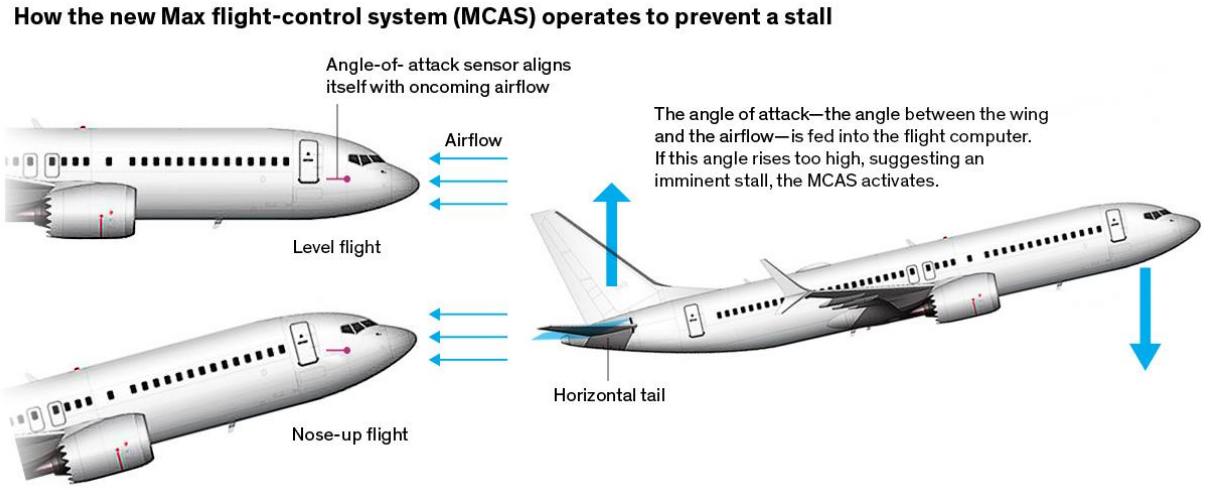

Figure 3 How MCAS works. Credit: Norebbo.com

well as at already-high angles of attack. It violated that most ancient of aviation canons and probably violated the certification criteria of the U.S. Federal Aviation Administration. But instead of going back to the drawing board and getting the airframe hardware right (more on that below), Boeing relied on something called the "Maneuvering Characteristics Augmentation System," or MCAS. It pushes the nose of the plane down when the system thinks the plane might exceed its angle-of-attack limits; it does so to avoid an aerodynamic stall. In short, Boeing tried to fix a hardware problem with the software.

Now, this is an understandable move from a commercial perspective because the major selling point of the 737 Max is that it is just a 737, and any pilot who has flown other 737s can

fly a 737 Max without expensive training, without recertification, without another type of rating. This suits the airlines' interest in having one "standard" airplane because all their pilots can fly and makes both pilots and airplanes fungible, maximizing flexibility, and minimizing costs.

However, this is where the problem occurs. MCAS makes its judgment based on the sensor input of the angle-of-attack, which is the angle between the wing and the airflow, but this is a sensor that goes haywire all the time given the extreme environments it has to go through. Hence, basing the decision on this unreliable indicator is not a wise call. What's even worse, the current implementation of MCAS denies pilots the sovereignty to take control when they decide the sensor is off.

In a way, this MCAS system should not be a technically black-box model because the developers of the software have full access to the code and the software contains no complex structures like a neural network, albeit it is possible that the software engineers employed to write this piece of code do not really have experiences in the aviation industry. Also, the software community has a different culture in face of risk and iterative improvement since they can always fix a bug in a new patch release. However, the system becomes a socially black-box one, and I think not out of the intention of the engineers. It is reported that pilots have little special training to adapt to this new feature when they are told Max is similar to old 737s, and many pilots did not receive any alerts to look into the inner workings of the MCAS system, eventually turning it into a black box. Thus, the Boeing 737 is an unwanted social black box, but how is it allowed to exist?

The answer is an economic, political, social, and technical mistake. It is discussed previously that the original motivation for this change in 737 is more out of a business point-of-

view than out of technical necessity because economically this is the move to maximize profit for airlines and manufacturers. Also, socially the difference in culture between software and the aviation industry can be attributed as one of the sources of error. But how does the FAA play a part in this picture? If the design has this big of a flaw, how have they failed their responsibility as regulators? The answer is closely linked with the concept of “Designated Engineering Representative", or DER. In the old days, the FAA had armies of aviation engineers in its employ. Those FAA employees worked side by side with the airplane manufacturers to determine that an airplane was safe and could be certified as airworthy. As airplanes became more complex and the gulf between what the FAA could pay and what an aircraft manufacturer could pay grew larger, more and more of those engineers migrated from the public to the private sector. Soon the FAA had no in-house ability to determine if a particular airplane’s design and manufacture were safe. Thus was born the concept of the DER, who are people in the employ of the airplane manufacturers, the engine manufacturers, and the software developers who certify to the FAA that it’s all good. Now, this is not quite as sinister a conflict of interest as it sounds. It is in nobody’s interest that airplanes crash. However, this policy did allow a faulty system to pass the test.

So, for some technically or socially black box system like in the two cases, who is at stake, and who should bear the blame when something goes wrong? This brings the next section of the discussion on stakeholders and their corresponding responsibilities.

Stakeholders and their responsibilities

Although the people affected in various autonomous systems with black-box models are distributed broadly, I will divide them into three broad categories, each representing a distinct type of force shaping society and technology, like ***academia***, ***the public***, and ***industry***. I will each give an account of their role in the process and the efforts made recently.

Academia

Ever since the birth of existing technically black box models such as DNNs and CNNs, motions to improve transparency, interpretability and prevent automated discrimination is present in the field. As these technologies gain rapid progress and traction in real-world applications, we are seeing a surge in the amount of work in this effort. The upward trend can be seen in Figure 4, the total number of citations with the keyword “Explainable AI” in the web of science. It is also shown in the treemap that most of the work is in the domains of computer science, telecommunications, robotics, and automation control systems, etc.

Within these new work, two nascent approaches that hold promise for increasing model transparency are local-interpretable-model-agnostic explanations (LIME) and attention techniques. LIME attempts to identify which parts of input data a trained model relies on most to

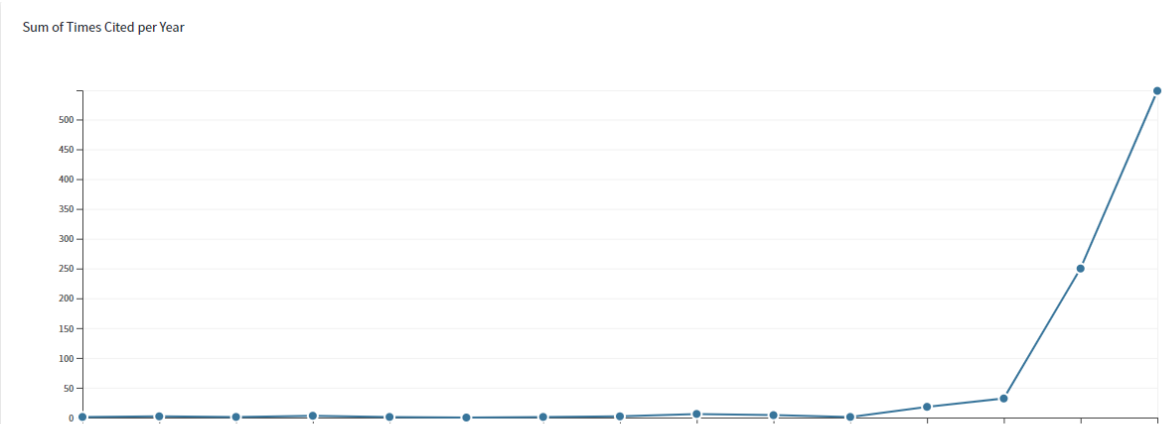

Figure 4 Sum of times cited per year of work with keyword "Explainable AI".

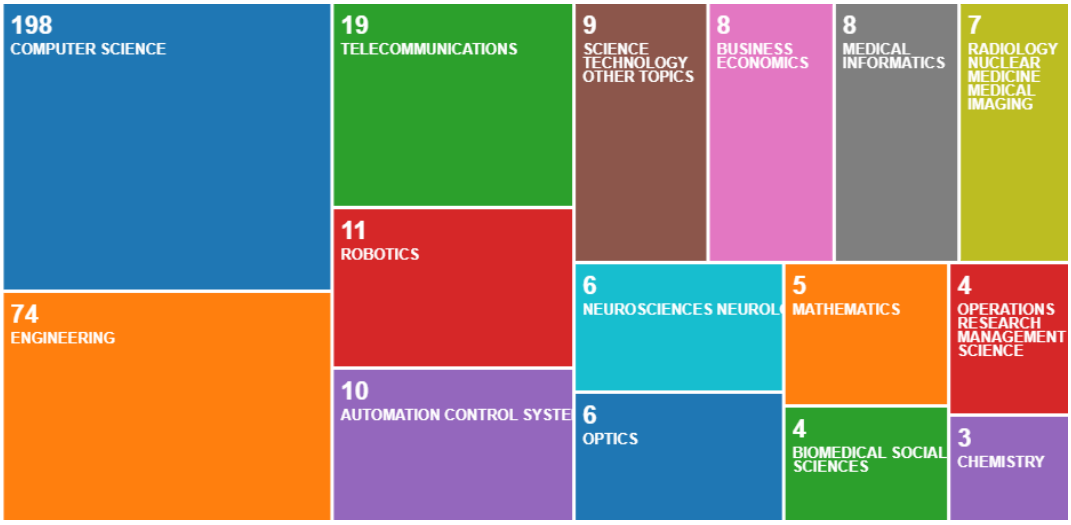

Figure 5 Tree map of research areas with the most work with keyword "Explainable AI"

make predictions in developing a proxy interpretable model (Guestrin et al., 2016). This technique considers certain segments of data at a time and observes the resulting changes in prediction to fine-tune the proxy model and develop a more refined interpretation (for example, by excluding eyes rather than, say, noses to test which are more important for facial recognition). Attention techniques (Wang et al., 2016) visualize those pieces of input data that a model considers most as it makes a particular decision (such as focusing on a mouth to determine if an image depicts a human being (Chui et al., 2018).

The public

All users of future autonomous systems with black box models should be included in the public since they will be most impacted by the effectiveness and trustworthiness of the system. Therefore, we should expect more voices to ask for safety guarantees in these critical systems in multiple channels including regulatory or legal actions. For example, The European Parliament recently adopted the General Data Protection Regulation (GDPR), which will become law in May 2018. An innovative aspect of the GDPR, which has been much debated, are the clauses on

automated (algorithmic) individual decision-making, including profiling, which for the first time introduce, to some extent, a right of explanation for all individuals to obtain "meaningful explanations of the logic involved" when automated decision making takes place. Despite divergent opinions among legal scholars regarding the real scope of these clauses, everybody agrees that the need for the implementation of such a principle is urgent and that it represents today a huge open scientific challenge (Guidotti et al., 2018). As the application of AI expands, regulatory requirements such as this will drive the need for more explainable AI models from the industry.

Given the legislation process of related issues could take much longer periods of time, it is also important to cultivate the awareness of sensitivity toward novel, safety-critical systems. Different individuals have their own attitudes, whether radical or conservative, towards state-of-the-art technologies, but all of them should have fair access to information about the status quo of these systems. That said, I suggest some kind of reading and discussion class on the facts about technologies should be required for all in their basic education. This additional part of the curriculum will better prepare future generations in the face of these new autonomous systems, letting them put appropriate amounts of trust into new products or services.

Industry

As demonstrated in the case of Boeing 737 Max crashes, the power dynamics in the industry regarding the interpretability of black-box models is an interesting one. On the one hand, any sane investors of such new autonomous systems will be happy to see an improvement in transparency and safety guarantee in the models, because this can greatly boost the trust from the users, decrease the likelihood of accidents, and hence increase the confidence from the

market. This is why major technology companies are pouring resources into supporting relevant works in academia. On the other hand, any investment asks for profits back, and a detailed cost-benefit analysis will be carried out when a model powerful but less so interpretable is used in a new system or product. If it is decided that the cost of building an interpretable model is not "worth it", the companies could choose not to "do the right thing". Even worse, corporations could avoid interpretable models deliberately because they can make profits from the intellectual property afforded to a black box.

Consider the COMPAS proprietary recidivism risk prediction tool discussed above that is in widespread use in the U.S. Justice System for predicting the probability that someone will be arrested after their release (Brennan et al., 2009). The COMPAS model is equally accurate for recidivism prediction as the very simple three rule interpretable machine learning model involving only age and number of past crimes. However, there is no clear business model that would suggest profiting from the simple transparent model. The simple model was created from an algorithm called Certifiably Optimal Rule Lists (CORELS) that looks for if-then patterns in data. Even though the model looks like a rule of thumb that a human may have designed without data, it is instead a full-blown machine learning model. Standard machine learning tools and interpretable machine learning tools seem to be approximately equally accurate for predicting recidivism, even if we define recidivism in many different ways, for many different crime types (Zeng et al., 2017; Tollenaar & van der Heijden., 2013). This evidence, however, has not changed the momentum of the justice system towards proprietary models. As of this writing, California has recently eliminated its cash bail system, instead enforcing that decisions be made by algorithms; it is unclear whether COMPAS will be the algorithm used for this, despite the fact that it is not known to be any more accurate than other models.

Transition into *phase two*?

The examples of COMPAS and other similar cases all illustrate a problem with the business model for machine learning. In particular, there is a conflict of responsibility in the use of black-box models for high-stakes decisions: the companies that profit from these models are not necessarily responsible for the quality of individual predictions. A prisoner serving an excessively long sentence due to a mistake entered in an overly-complicated risk score could suffer for years, whereas the company that constructed this complicated model is unaffected. On the contrary, the fact that the model was complicated and proprietary allowed the company to profit from it. In that sense, the model's designers are not incentivized to be careful in the model's design, performance, and ease of use.

Combining the fact of incidents of users over-trusting autonomous systems instead of holding a suspicious view, one has reason to question whether our society has already beginning to transition to *phase two* in Hughes' theory of technological momentum, that is we already entered or are about to enter a phase where the new technology, i.e. autonomous systems with black box models, modifies the society in an irreversible and irreparable way, and it self-propagates itself. As described by Frank Pasquale, we should be warned about a growing "black box society", governed by secret algorithms protected by industrial secrecy, legal protections, obfuscation, so that intentional or unintentional discrimination becomes invisible and mitigation becomes impossible." (Pasquale, 2015)

Conclusion

With the mass adoption of AI technologies, many autonomous systems with black-box models are created and to be embedded in our society, including autonomous driving, drones, and medical assistive technologies. The success of such systems is accompanied by rising concern over the issues of safety, transparency, and trustworthiness of them.

In this paper, I applied the theory of technological momentum to the problem and argued that we may be at a turning point in history where the society will gradually lose its shaping abilities on this technology from *phase one* to *phase two*. This paper also proposes a distinction between technically black-box and socially black-box models and exemplifies the problems, causes, and potential solutions to each through two case studies of autonomous cars loopholes and Boeing 737 Max crashes. Then, an analysis from multiple perspectives is carried out, discussing the root of the problem and future works to be done from the standpoints of relevant stakeholders: academia, the public, and industry.

My goal in this work is to highlight the status quo of the use of black-box models in our society and encourage the study of more interpretable models from academia, more responsible policy-making from the regulators, and more sustainable business models from invested corporations. Also, more frequent communications among these sectors will greatly help the society realize the problem in a bigger picture and promote more well-rounded solutions. It is my hope that people in different domains can all make efforts to build more responsible, and trustworthy future systems, and collectively contribute to a better society.

References

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. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.5a8a3a3d

Pasquale, F. (2015). The Black Box Society. In *Harvard University Press*. https://doi.org/10.4159/harvard.9780674736061

Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., & Houkes, W. (2011). A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems. *Synthesis Lectures on Engineers, Technology and Society*. https://doi.org/10.2200/s00321ed1v01y201012ets014

Schmelzer, R. (2019). What Happens When Self-Driving Cars Kill People? Retrieved from Forbes website: https://www.forbes.com/sites/cognitiveworld/2019/09/26/what-happens-with-self-driving-cars-kill-people/#571da96d405c

McCausland, P. (2019). Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk. *NBC News*. Retrieved from https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281

Kenton, W. (2020). Black Box Model. Retrieved from Investopedia website: https://www.investopedia.com/terms/b/blackbox.asp#:~:text=What Is a Black Box Model%3F&text=In science%2C computing%2C and engineering,knowledge of its internal workings.

Nassi, B., Nassi, D., Ben-netanel, R., Mirsky, Y., Drokin, O., & Elovici, Y. (2020). Phantom of the ADAS : Phantom Attacks on Driver-Assistance Systems. *Cryptology EPrint Archive*.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*.

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best?: A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. https://doi.org/10.1111/j.1467-985X.2012.01056.x

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. https://doi.org/10.1111/rssa.12227

Marco, B., Ribeiro, T., Singh, S., & Guestrin, C. (2016). Introduction to Local Interpretable Model- Agnostic Explanations (LIME).

Wang, Y., Huang, M., Zhao, L., & Zhu, X. (2016). Attention-based LSTM for aspect-level sentiment classification. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d16-1058

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-019-0048-x

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System. *Criminal Justice and Behavior*, *36*(1), 21–40.

Chui, M., Manyika, J., & Miremadi, M. (2018). Crossing the frontier: how to apply AI for impact. In *McKinsey Analytics*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–45. https://doi.org/10.1145/3236009