**An Ideological Exploration of Safe Machine Learning**

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

**Grant Dong**

Spring 2022

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

Advisor

Sean M. Ferguson, Department of Engineering and Society

## An Ideological Exploration of Safe Machine Learning

### Introduction

Machine learning (ML) is a popular buzzword for many big and influential companies, especially for companies that specialize in acquiring data. Thanks to powerful cloud technology being a catalyst for the degree of data availability we have today, there is now a plethora of scalable deployments and ideal training scenarios. Being able to predict outcomes accurately is a very important resource and a much-desired skill to have for any individual or company. Especially in this day and age where companies are constantly competing for the best algorithms, the popularity behind ML is well-founded. ML has reached a point in which it is almost critical to many kinds of production software systems (Rowan, 2020).

Bias is a chronic and ever-present issue in the field of ML; it is also a commonly used buzzword when people talk about controversy related to those fields. However, it is important to define bias, and to understand why it is such an issue to begin with. Some biases can be desired to achieve the correct system functionality. Then there are unimportant biases, where a bias is detectable by statistical tools as a systematic difference, but societal disclosure does not deem these differences as sensitive and problematic. Then we have undesired biases, where the stakeholders of the system or other persons are negatively impacted by the system (Balayn et al., 2021). These will be the harmful biases that we need to find a stable solution for in the discussion later.

ML technology has a lot of power, and with that span of influence, it can do a lot more harm than good if the risks are not fully realized. The issues plaguing ML today, if left unaddressed, could lead to damaging future technological developments that could have major consequences for the human rights and civil liberties of the subjects who depend on them.

Hence, with ML rapidly becoming integral to organizations' value proposition, the need for organizations to improve the security and robustness of ML systems is high (Boesch, 2021). Thus, it is important to step back and analyze the imaginary of how ML can be revolutionized to become safe and dependable in the future. In an effort to revive ML reputation, I will conduct a structured investigation using the imaginaries framework by synthesizing sentiments from various stakeholders in the form of three distinct ideologies that attempt to address the problems with ML technology and transition us towards Safe ML.

**Stakeholder Analysis: The General Populace**

Heidi Ledford (2019) presents a profound problem with racial bias in healthcare algorithms. A widely used algorithm has been systematically discriminating against the black population by referring fewer black people than white people who were equally sick to healthcare programs. Although, it should be noted that this type of study is rare because researchers often cannot gain access to proprietary algorithms and the amount of sensitive health data needed to thoroughly test them, according to epidemiologist Milena Gianfrancesco at the UC San Francisco, who has done a lot of work with sources of bias in electronic medical records. But regardless, the finding is certainly not dismissible, as the analysis reveals how the average black person was also substantially sicker than the average white person. Holistically, the data showed that the care provided to the black community cost an average of $1800 less per year than the care given to a white person with similar health problems (Ledord, 2019).

Another popular case of ML bias was brought to light by Natasha Lomas, a senior reporter for Tech Crunch who investigated FaceApp's popularity on Instagram and Twitter. The selfie app uses neural networks to change a person's appearance by plastering a fake smile onto their face or making them look older or younger. But the app's hot filter seemingly just makes

people look whiter. FaceApp's founder Yaroslave Goncharov claimed that the company was very apologetic and it was certainly not the intended behavior. The whitening was an unfortunate side-effect of the underlying neural network caused by training set bias. The hot filter was changed to the name "spark" to address the issue, but this workaround failed to solve the problem as the app was still sticking to an inherently racist system; it required expensive overhead to correct for the mistakes in the biased training dataset (Lomas, 2017). FaceApp chose not to waste resources on a cause that they did not prioritize. In short, FaceApp's issue stemmed from their algorithm not being trained on a diverse enough data set. This specific case illustrates the lurking problem of underlying algorithmic bias in such a visually impactful way.

James Vincent, a reporter who has a lot of experience in analyzing ML technology, highlighted the infamous cause of how Amazon scrapped an internal project that was trying to use ML to collect job applications after the software consistently downgraded female candidates. It turned out the models used by Amazon were perpetuating the data biases in a male-dominated working environment of the tech world. The program penalized applicants who attended all-women's colleges, as well as any resumes that contained female keywords. This case study demonstrates how an abstract concept such a gender bias can be operationalized into measurable features of text that can be computationally identified. Work within the field of stylistics on gender and language has identified recurring linguistic that are attributable to gender bias like terms used to describe groupings of men and women, how the term "girl" is used in more disparaging and sexual contexts, the ordering of male and female names, and how appearance is emphasized more than personality for women (Mehrabi et al., 2021). The team behind the project reportedly intended to speed up the hiring process with the ML recruiting system, according to an Amazon employee on the recruiting team. When the company realized the

software was not producing gender-neutral results it was immediately taken down for major changes (Vincent, 2018).

**Stakeholder Analysis: The Large Corporate Perspective**

Although the combination of scale and legibility makes ML systems uniquely attractive to large institutions seeking easy management and control over large populations, this doesn't necessarily mean that large companies are strictly benefitting from exploiting the use of ML. The rippling effects of biases can certainly affect these corporations negatively. DataRobot has conducted surveys for over 350 US companies who are major technology leaders. The survey found that 80% of US firms have problems despite having bias monitoring or algorithm tests already in place. These same organizations are already feeling the impact of this problem in the form of a 62% loss in revenue, 61% loss in customers, 43% loss in employees, and 35% incurred legal fees due to lawsuits and legal action (Combs, 2022). Despite their different perspective on ML technology, bias is just as major of a problem to them as it is to the individual.

Senior writer of Tech Republic, Veronica Combs, also examined survey data about how companies want to tackle these issues. 81% of survey responders believe in some sort of government intervention and regulation that would be helpful in addressing two particular components of the challenge: defining and preventing bias. Additionally, 45% of company leaders worry about the increase in costs and creation of barriers of adoption due to the new regulations, whereas only 32% of respondents said they are actually concerned with a lack of regulation which could end up hurting certain groups of people (Combs, 2022). This interaction between conflicting opinions on a solution to the ML problem is a direct result of how the issue is perceived via different ideologies. Many stakeholders are proponents for a future with ML at

the forefront of technological advancement, but what it takes to get there is where the discussion becomes much richer.

**Ideology 1: An Algorithmic Challenge to Overcome**

The vanilla ideology that most tech companies are major proponents for in regards to eliminating biases within their ML models is to think of the issue as an algorithmic barrier. Most technology that involves around improving models to reduce bias is relevant to the field of Adversarial ML. Adversarial ML is an ML method that aims to trick models by providing deceptive input, also known as an adversarial attack. Hence, it includes both the generation and detection of adversarial examples, which are inputs specially created to deceive models (Boesch, 2022). Adversarial attacks exploit the biases hidden within an ML model because bias is most easily classified as a mistake from the output. The information from these attacks allows for self-iteration and refinement of the model to prevent them from being susceptible to those mistakes in the future. There are also many use cases and applications of adversarial ML. In short, the main idea of adversarial ML revolves around a closed, black-box approach on how to improve an ML model when given differing training data. Adversarial ML is at a pivotal moment. As these systems become more widely deployed, theoretical attacks and defenses rooted in the academic literature will become the stuff of people's lives (Albert et al., 2020).

Conditional Generative Adversarial Networks (CGANs) are a type of adversarial ML that involves two partner models trained simultaneously, where one model is trying to create inputs to fool the other, and the other would learn and adapt to overcome them. This process is repeated until the resulting generative model can generate new unbiased data. This synthetic fair data has selective properties from the original data, and experimental results show that the proposed solution can efficiently mitigate different types of biases, while simultaneously enhancing the

prediction accuracy of the underlying ML model (Abusitta et al, 2019). Scholars behind the research of these CGANs have worked on a lot of developmental work for improving various types of ML models, so it is reliable to perceive CGANs as a state-of-the-art piece of adversarial technology. However, a major limitation is unpredictable bias. Information about the bias must be known ahead of time for the model algorithm to successfully mitigate it (Abusitta et al, 2019). Bias is a broad and undefined problem, which does not always target members of minority groups. Hence, CGANs do not fully solve this problem as there will always exist some kind of hidden biases that are unaccounted for.

Dynabench is a maturing research platform developed by Meta AI, that utilizes dynamic benchmarking. It measures how easily ML systems are fooled by humans, which is a better indicator of a model's quality than current static benchmarks provide. Ultimately, this metric will better reflect the performance of ML models in the circumstances when interacting with people, who behave and react in complex, changing ways that can't be reflected in a fixed set of data points. Dynamic benchmarking happens over multiple rounds, where each time the researcher or engineer using Dynabench selects models to serve as the target to be tested. Dynabench then collects examples using these models and periodically releases updated datasets to the ML community. When new models catch most of the examples that fooled previous models, they then start a new round with these better models in the loop. This fast cyclical process can be easily repeated, so that if biases appear over time, Dynabench can be used to identify them and create new examples that test whether the model has overcome them (Meta, 2022). Overall, Dynabench overcomes the main limitations of static benchmarking, and will be less prone to bias and artifacts. But it is important to also consider a limitation of time and money, as Dynabench's process can be quite slow and expensive with its brute force nature to identify and mitigating

biases. It is true that this method can eventually yield promising results, but sometimes it will take too many iterations to get to a point where the ML models are completely bias free. Not to mention the process needs to start from scratch for tasks in different contexts.

**Ideology 2: Data is the Root of All Bias**

Another well-supported ideology prioritizes the reformation of data collection processes, since data is just information that is a reflection of our biased society. Every ML algorithm operates wholly within the world defined by the data that were used to calibrate it. Limitations in the data set will magnify the bias of outcomes (Baer & Lofi, 2021). Most unfairness arises when certain categories of the population might have been discriminated against in the past, intentionally or unintentionally, and are therefore also discriminated against by a system trained on historical decision data (Balayn et al., 2021). If the training data have not been investigated for unfairness, such unfairness would only be discovered at deployment time, where the damage would have already been done.

To understand the ideology completely is to formally defined unfairness and how it stems from society. There are two major types of fairness: group fairness and individual fairness. Group fairness is achieved through statistical parity, which is verified if the records in both the protected and unprotected groups have an equal probability to receive a positive outcome. Individual fairness relies on the idea that similar individuals should be treated similarly independently of their membership to one of the groups (Gajane & Pechenizkiy, 2018). Ethical ML researcher, Andrew Burt, redefines unfairness as disparate impact, which is unintentional discrimination directly manifested from the data itself. Avoiding disparate impact is an altogether more complex undertaking because it occurs when a seemingly neutral variable in the data like income acts as a proxy for a protected variable like ethnicity. What makes avoiding

disparate impact so difficult in practice is that it is extremely challenging to truly remove all proxies for protected classes. In a society shaped by profound systemic inequities, disparities can be so deeply embedded that it oftentimes requires painstaking work to fully separate what variables operate independently from protected attributes (Burt, 2020).

The easiest issues to tackle regarding the data problem is just simply by collecting better data, and this is exemplified by case studies in healthcare. According to researchers at Harvard Medical School, the first steps in which bias can be introduced are data collection and data preparation. In many ML-based applications in medicine, data like risk factors and other clinical parameters, are heavy influencers of potential biases that show up in the output. If the training data is subject to sampling bias, the same bias may be replicated when the system is applied in the clinical setting. To address this, datasets are compiled to be as diverse and large as possible to have a better representation of all patient groups. Developers can also monitor the performance levels for improvement in further iterations (Vokinger et al., 2021). A research article published by several google employees proposed a new framework for a non-linear cycle of dataset development that involves documentation practices drawing from the best stages of a software lifecycle, diverse oversight processes, and robust maintenance mechanisms (Hutchinson et al., 2021). This ensures that often overlooked work and decisions that go into dataset creation remains clearly visible, which is a critical step in closing the accountability gap in ML systems and a necessary resource in targeting roots of biases.

Another layer of nuance is added to this issue of purifying data when one considers that disparate impact can also be a very helpful tool, which ties back directly to desired biases. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), is a software that measures the risk of a person to recommit another crime. Judges use COMPAS to

decide whether to release an offender or to keep him or her in prison. An investigation into the software found that COMPAS is more likely to have higher false positive rates for black offenders than white offenders in falsely predicting them to be at a higher risk of recommitting crime (Mehrabi et al., 2021). But patterns in the data that actually identify people with criminal intent may act in the same way as patterns that notice race, and it is very hard to separate the two without jeopardizing the effectiveness of the system. Ultimately, a possible solution is interpolating a type of fairness from the data while ignoring it for a different metric of fairness. Graduate students at the University of Southern California who've performed plentiful research in ML stated how it is nearly impossible to understand how one fairness solution would fare under a different definition of fairness. Synthesizing these definitions into one remains an open research problem, since it can make evaluation of these systems more unified and comparable (Mehrabi et al., 2021).

**Ideology 3: Affirmative Action in a Perpetually Biased System**

The third ideology embraces how the fundamental idea of many advanced and complex ML tasks is to emulate the mechanics of the human brain, as seen by deep learning with its artificial neural networks. Just as biases affect human intelligence, it will also affect artificial intelligence. ML systems are prone to incorporating the biases of their human creators, because at the end of the day, they are still designed and architected by people (Burt, 2021). As stated by Tobias Baer and Vishnu Nath, two seasoned consultants who worked with various large tech companies, ML will perpetuate and even amplify behavioral biases. A social-media site filtering news based on user preferences reinforces natural confirmation bias in readers. Furthermore, the site may even be systematically preventing perspectives from being challenged with

contradictory evidence. This self-fulfilling prophecy is a related, inevitable by-product of algorithms.

If bias is always inevitable, people with this perspective believe there must be some way we can address victims of ML discrimination, and this is where affirmative action becomes a major actor in mitigating ML damage. As mentioned in the first ideology, adversarial ML is a type of ML focusing on deceiving an ML system to achieve some kind of altered output. Using this process, it is possible to give aid and extra-support to individuals that belong to discriminated groups. Little attention has been paid to equity, which is the concept that each individual or group is given the resources they need to succeed. Operationalizing this definition and researching how it augments or contradicts existing definitions of fairness remains promising (Mehrabi, 2021). Although it is common within the security community to view those who wish to interfere with systems as "attackers," this framing ignores that those who resist such systems could easily be pro-democracy protesters or academics interested in evaluating the inclusiveness of the system as the biased system itself could be a "malicious" actor (Boesch, 2022). This dynamic is brilliantly illustrated via the adversarial arms race involving attackers attempting to break into the system, and the defenders attempting to build robust defenses.

Sauvik Das, a professor at the Georgia Institute of Technology who is an expert in cybersecurity and privacy policy developed a concept known as Subversive Artificial Intelligence (SAI). SAI is a form of adversarial ML where obfuscation filters are created for users to apply to their digital experience which inhibits algorithmic surveillance in reliable ways. This motivation stems from the social goal of empowering people, particularly those from communities who disproportionately bear the negative effects of algorithmic surveillance caused by ML bias, to fully use online services to connect with their communities and share their voices

without fear of being discriminated against (Das, 2020). SAI focuses on a human-centered design process where all diverse stakeholders are taken into consideration and the model designs are evaluated directly with these stakeholders in mind. Ideally, everyone should feel equally empowered to use the latest ML technology without having the fear of discrimination. Another application is EqualAI's project which runs a detection perturbation algorithm for the purpose of allowing individuals to make a certain image less likely to be detectable as a face. These types of obfuscations could be used by photographers who are documenting protests and would like to post the images without the danger for facial recognition software to identify protesters (Albert, 2020).

Through understanding lived experiences of resistance and applying the lessons of other disciplines, the adversarial ML community cannot just understand its work as political but take affirmative steps to ensure that it is used primarily for good. The existing ethos of ML research focuses on automated algorithmic surveillance infrastructure for the purposes of enhanced profits, security, and stability. The impact of this ubiquitous, expansive, and impersonal algorithmic surveillance can cultivate widespread barriers that stifle free expression and exacerbate systemic inequities (Das, 2020). In a more profound sense, the implementation of group fairness is the application of affirmative action. Most proponents claim group fairness is not meritocratic and reduces efficiency, but meritocracy itself leaves room for more bias. Therefore, race-conscious approaches would be significantly more efficient than race-blind approaches (Gajane & Pechenizkiy, 2018).

## Conclusion

After carefully reviewing the different ideologies for achieving a safe future with ML, it is clear that the necessary actions to achieve that future are quite divergent. One dominant belief

is in improving the algorithms of the ML models themselves to eliminate biases, another is to target data compilation and restructure how we record data to avoid societal biases, and the third is using new adversarial technology to give people commonly discriminated against by ML bias an advantage and special privileges. After evaluating all the scholarship, I believe all three of these modified visions are justified and can potentially contribute to a future where ML can be dependable and safer for the general population.

At first, I was unclear about the major distinction between the first and second ideologies. Being a computer science student who has experience designing and refining ML models, I believe that the processes of improving ML models computationally also entail examining training data and finding ways to mitigate biases in the data. But after my research, I have developed a clear distinction between the two ideologies. The algorithmic ideology is viewed as noticing biases in the data and self-correcting the ML model to avoid these biases, whereas in the data ideology, it is about defining what fair and unbiased data looks like and making sure data starts out unbiased from dataset creation. Hence, if there was a pipeline that illustrates the creation of ML bias, the data ideology emphasizes the existence of a bottleneck earlier in the pipeline than the algorithmic ideology.

It is important to acknowledge that when isolated, each of these ideologies for a vision of safe ML have impediments and dilemmas. Ideally, a realistic way to achieve that future of safe ML is to implement a combination of three of the ideologies. Most scholars think of a single perspective when addressing the issues of ML bias, but if more people actually embraced the different perspectives, a solution to ML bias will be more apparent. If we constantly improved ML models to notice and avoid biases, compiled datasets to reduce latent biases, and took an active approach to help individuals who suffer from ML discrimination, a future where the

negative impacts of ML bias are minimized effectively will certainly be on the horizon. In an almost trivial way, if all three ideologies were synthesized in considering a system that could reduce biases and their consequences efficiently, a future for safe ML will come to fruition.

## References


Abusitta, A., Aïmeur, E., & Wahab, O. A. (2019, May 23). *Generative adversarial networks for mitigating biases in machine learning systems*. arXiv.org. Retrieved March 3, 2022, from https://arxiv.org/abs/1905.09972

Albert, K., Penney, J., Schneier, B., & Kumar, R. (2020, March 27). *Politics of adversarial machine learning*. Social Science Research Network. Retrieved March 5, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547322

Baer, T., & Kamalnath, V. (2020, September 16). *Controlling machine-learning algorithms and their biases*. McKinsey & Company. Retrieved March 4, 2022, from https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases

Balayn, A., Lofi, C., & Houben, G.-J. (2021). Managing bias and unfairness in data for decision support. *The VLDB Journal*, *30*(5), 739–768. https://doi.org/10.1007/s00778-021-00671-8

Boesch, G. (2022, January 17). *What is adversarial machine learning?* viso.ai. Retrieved March 3, 2022, from https://viso.ai/deep-learning/adversarial-machine-learning/

Burt, A. (2020, October 8). *How to fight discrimination in Ai*. Harvard Business Review. Retrieved March 5, 2022, from https://hbr.org/2020/08/how-to-fight-discrimination-in-ai

Combs, V. (2022, January 11). *Guardrail failure: Companies are losing revenue and customers due to Ai Bias*. TechRepublic. Retrieved March 1, 2022, from https://www.techrepublic.com/article/guardrail-failure-companies-are-losing-revenue-and-customers-due-to-ai-bias/

Das, S. (2020). *Subversive AI: Resisting automated algorithmic surveillance with human-*

*centered adversarial machine learning*. Retrieved March 9, 2022, from https://sauvikdas.com/papers/27/serve

Gajane, P., & Pechenizkiy, M. (2018, May 28). *On formalizing fairness in prediction with machine learning*. arXiv.org. Retrieved March 5, 2022, from https://arxiv.org/abs/1710.03184

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021, March 1). *Towards Accountability for Machine Learning Datasets: Proceedings of the 2021 ACM Conference on Fairness, accountability, and transparency*. ACM Conferences. Retrieved April 11, 2022, from https://dl.acm.org/doi/abs/10.1145/3442188.3445918

Leavy, S. (2018, May 1). *Gender bias in Artificial Intelligence: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*. ACM Conferences. Retrieved March 3, 2022, from https://dl.acm.org/doi/pdf/10.1145/3195570.3195580

Ledford, H. (2019, October 24). *Millions of black people affected by racial bias in health-care algorithms*. Nature News. Retrieved March 3, 2022, from https://www.nature.com/articles/d41586-019-03228-6

Lomas, N. (2017, April 26). *FaceApp apologizes for building a racist AI*. TechCrunch. Retrieved March 1, 2022, from https://techcrunch.com/2017/04/25/faceapp-apologises-for-building-a-racist-ai/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Meta. (2020, September 24). *Introducing dynabench: Rethinking the way we benchmark ai*. Meta AI. Retrieved March 4, 2022, from https://ai.facebook.com/blog/dynabench-rethinking-ai-benchmarking

Rowan, I. (2020, July 18). *The state of AI in 2020*. Medium. Retrieved March 1, 2022, from https://towardsdatascience.com/the-state-of-ai-in-2020-1f95df336eb0

Vincent, J. (2018, October 10). *Amazon reportedly scraps internal AI recruiting tool that was biased against women*. The Verge. Retrieved March 4, 2022, from https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021, August 23). *Mitigating bias in machine learning for medicine*. Communications Medicine. Retrieved March 1, 2022, from https://www.nature.com/articles/s43856-021-00028-w