Ethics and Bias in Machine Learning: A Technical Study of What Makes Us "Good"

A Thesis Presented in Partial Fulfillment of the

Requirements for the Master of Science in

Digital Forensics and Cyber

Security John Jay College

of Criminal Justice City

University of New York

Ashley Nicole Shadowen

Fall:
February

TABLE OF CONTENTS

ABSTRACT

The topic of machine ethics is growing in recognition and energy, but bias in machine learning algorithms outpaces it to date. Bias is a complicated term with good and bad connotations in the field of algorithmic prediction making. Especially in circumstances with legal and ethical consequences, we must study the results of these machines to ensure fairness. This paper attempts to address ethics at the algorithmic level of autonomous machines. There is no one solution to solving machine bias, it depends on the context of the given system and the most reasonable way to avoid biased decisions while maintaining the highest algorithmic functionality. To assist in determining the best solution, we turn to machine ethics.

MOTIVATION

The ubiquity of autonomous machines in our world means algorithmic decisions affect nearly every aspect of our lives. Academic, government, military, and commercial interest has grown around the subject in recent years due to potential reduced costs and increased productivity. Although this topic has fascinated sci-fi enthusiasts for centuries, experts agree we have barely scratched the surface in understanding the impact of artificial intelligence (AI). Challenges related to artificial intelligence may seem complex and abstract, but we learn most from studying its current form - the algorithms that increasingly determine our everyday reality. For every modern person whose life involves the use of a device to access the Internet, machine learning (ML) algorithms have become the fabric of their experience. Using machine learning, search engines like Google find us the “best” results, email hosts like Office 365 filter our spam, and social networks like Facebook track and tag our best friends and family for us.

We are building algorithmic computations into our everyday experience. Some consequences of machine learning can seem innocuous with a hypothetical long-term impact that can incur financial or mission loss. These machine learning applications are identified as “Type B” by researchers of cyber-physical safety at IBM [1]. For example, a person could apply for a loan and get denied because of a decision made by a machine. The Type A application of machine learning is more insidious, with real-time or near term impact [1]. Consider, for example, the use of an algorithm to calculate the risk of a criminal to commit a second crime or “recidivate”. The algorithm’s programmers must choose the right machine learning algorithm, the right metrics to weigh in making predictions, and an appropriately representative dataset to train the algorithm. Once all of this is complete, proper use of the algorithm must be outlined and implemented. Only then should validation testing of the entire machine in the intended application be conducted. Unfortunately, this is the ideal and nowhere near the standard. Machine learning algorithms are systematically deployed before proper vetting to be used in a myriad of consequential circumstances today. The propensity for errors resulting in ethical concerns like programmed machine bias *must* be considered from initial construction of an algorithm to final implementation.

The importance of machine learning ethical considerations will only grow as our reliance on technology increases. For the purposes of this paper, we focus solely on bias as an ethical consideration of machine learning. Just this year, a team from the Alan Turing Institute determined that if machine learning is to be used with legal and ethical consequences in our society, responsible use of this tool means pursuing the accuracy of predictions *and* fair resulting social implications [2]. Intuition leads us to predict that an ethical challenge like bias,

can only be solved by means of an ethical solution. The question is, what kind of solution is best in what circumstance? We explore the options below.

BACKGROUND

Machine Learning

Artificial Intelligence is the ability of machines to exhibit human-like decision making in context with their surroundings. According to the Turing test developed by Alan Turing in 1950, if a human is interacting with a machine and cannot tell if the machine is human or robot, artificial intelligence has been achieved. Machine learning is a branch of AI in which machines are given the ability to learn without being programmed [3]. Machine learning accomplishes an aspect of artificial intelligence in which some data from an external context may be ingested, "understood," and integrated into the algorithmic function to make predictions. Parametric and nonparametric algorithms make up the two main groups within machine learning. Parametric machine learning entails algorithms that analyze data according to preset parameters defined by human creators. Nonparametric algorithms in machine learning allow for more freedom to learn any functional form from input data. [4] An example of nonparametric machine learning is deep learning, in which an algorithm starts with human-created code but slowly learns from trends found in training data input. Machine learning is a subset of artificial intelligence that allows for efficient solutions to complex and data-heavy problems that could take lifetimes for humans to achieve manually.

Machine Bias

As machine learning algorithms proliferate in everyday life, the scientific community has become increasingly aware of the ethical challenges, both simple and complex, that arise with the technology. Machine bias is one such ethical challenge. For the purposes of this paper, machine bias is defined as the oftentimes unintended algorithmic preference for one prediction over another that results in legally or ethically inappropriate implications. Said more concisely, machine bias is programming that assumes the prejudice of its creators or data [5]. The word "bias" has many meanings in a machine learning context, so it is necessary to define this term explicitly. In fact, bias is a required function in predictive algorithms. As Dietterich and Kong pointed out over 20 years ago, bias is implicit in machine algorithms, a required specification to determining desired behavior in prediction making. Bias is a widely used term in machine learning and statistics and can have many meanings [6]. Here, we refer to machine bias as the skewing of data to be biased according to accepted normative, legal, or moral principles.

Typically, any algorithm that harms human life in an unfair capacity due to protected attributes has a machine bias problem. This can range from subtle bias with ambiguous impact to outright discrimination with quantifiable financial, psychological, and life or death repercussions. Every company that uses technology in any capacity is at risk of being influenced by the bias inherent in their programming. As one of the most forward-thinking corporations of our time, Google creates wide-ranging machine learning projects that users have exposed as biased. In July 2015, Jacky Alcine posted to Twitter a Google Photo that used facial recognition to tag him and another friend as gorillas [5]. Ethnicity can confound machine learning algorithms, just as gender can. A study by AdFisher revealed that men were six times more likely than women to see Google ads for high paying jobs [5] [7]. In some cases, multiple groups are impacted.

Researchers found that commercially available emotion recognition software like Google's Vision API and Microsoft Emotion API (often used by developers for their applications) lag significantly when it comes to minorities, such as those outside of adult middle-age or the ethnic majority [8] . Biases like these influence the way users of all types of experience technology differently.

Scientists at the Center for Information Technology Policy at Princeton University, used a machine learning tool called "word embedding" to expose complex trends of ingrained bias in the words we use. The tool interprets speech and text by mathematically representing a word with a series of numbers derived from other words that occur most frequently in text alongside it. When over 840 billion words from the "wild" internet were fed into the algorithm, the experiment resulted in positive word association for European names and negative for African American. Additionally, women were more commonly associated with arts and humanities while men were more commonly associated with engineering and math [9]. This networked word identification system results in a deeper "definition" of a word than a typical dictionary, incorporating social and cultural context into the word's meaning.

The tools above are examples of type B applications in which hypothetical long-term impacts are increasingly likely. Type A applications have more direct and immediate consequences. Machine learning algorithms are being used to automatically produce credit scores, conduct loan assessments, and inform admissions decisions [10] [5]. Even further, robotic uses of machine learning include auto-pilot cars and aircraft [10] [11] [12], and "slaughter bots" [13] like those created by the US Army Future Combat systems program [12]. These systems necessitate bias, due to the moral nature of their function. These artificial intelligent agents use

machine learning to ingest contextual data, and determine whom is right to protect and why. In our case study, a commercially available machine learning algorithm called COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions, scores defendants on their likelihood to recidivate. As we will see, these scores can be used in a variety of circumstances that impact a defendant's access to bail or even ultimate sentencing decision [14]. Machine bias can result in mild to severe negative impact on a person's life. No matter how intelligent, these examples show that machines are still only a product of their creators. If we are biased, our machines will be biased, unless we train them otherwise.

Machine Ethics

Though the term is relatively new, machine ethics have existed since the birth of artificial intelligence. But not until recently have computer scientists and artificial intelligence researchers acknowledged the need for ethics. The sheer diversity and the far reach of machine learning applications requires an examination of machine ethics. One of the pioneers of machine ethics, James Moor has been discussing the topic since 1985 [15] . Implicit ethical agents are defined by Moor as ethical because of how they are programmed and what they are used to do. For example, the auto-pilot feature on a commercial airplane is "implicitly ethical" because when it works properly, humans aboard are transported safely from point A to point B [16]. Explicit ethical agents are machines given ethical principles or data for ethical decision making even in unfamiliar circumstances. Moor called these machines explicit ethical agents, and they are known by another name now, as Artificial Ethical Agents (or AEA) [11]. AEA can be programmed to exhibit ethical behavior using principles from a combination of ethical theories. Three primary ethical theories are: deontic logic, or obligation; epistemic logic, based on belief and knowledge; or action logic, based on action [16]. Some methodologies have bucked the trend of using ethical theories, and instead rely on majority rule ethics to inform morally sound

decision making machine criteria [12]. More recent is the work of Headleand and Teahan, who take the idea of Breitenburg's "emotional" vehicles, and apply it to ethics. The team claims that almost all machine ethics approaches are top-down in methodology. The project's ethical vessels make "ethical" decisions from the ground-up instead, through the use of layered rules with exceptions based on Asimov's "Three Laws of Robotics" [11]. In this paper, we focus primarily on how machine ethics can be used to solve for bias in machine learning, but the challenges and possibilities in the relatively nascent field require attention from experts of all disciplines.

Essential to AEA, according to Moor and other scholars, is the machines' ability to make explicit ethical decisions and provide evidence for justifying these decisions [16]. Integrating ethical decision making into machines can ultimately solve for issues that arise in algorithmic computing, like machine bias.

WHERE DOES BIAS COME FROM?

Challenges that result in machine bias

There are three primary questions we must ask when bias results from a machine learning algorithm: (1) How is the model made? (2) How is it used? (3) What ethical considerations are in place? As many machine learning and AI experts say "garbage in, garbage out." [17] Without a sufficiently effective algorithm and representative data to train on, a machine learning model will not output useful predictions.

How is it made?

A machine learning model is architected from the programmers that create it, the algorithm and metrics used, and the data it takes as input. When a development team programs a machine

learning model they must choose carefully: what type of algorithm is used, how the algorithm is set up, what metrics and parameters are used, and on what data the algorithm is trained and tested. Creators' influence can show up in unexpected ways. In the case of Pokémon Go, the gaming application that swept the U.S. and other countries, the game had fewer pokémon characters to catch in low-income, black neighborhoods. To choose locations, coders of Pokémon Go used location data crowd-sourced from gamers of "Ingress," informally surveyed as primarily young white men in 2013 - 2014 [18] [7]. In places like the United States, transparency around the exact details of proprietary machine learning algorithms is not often available [8]. In our case study, Northpointe, the company that developed the COMPAS algorithm, does not publicly disclose the calculations or algorithms they use. This makes it harder to verify the resulting predictions [19]. If users or outsiders are not able to immediately understand how decisions are being made by an algorithm, the "who" behind the curtains becomes even more important.

The algorithm in a machine learning model is a powerful way to control for bias, because it is the primary source of how the data fits to a model. A simple linear regression algorithm will typically fit well to data of varying dimensions with a strong linear form. In contrast, for complex, non-linear data, a support vector machine (SVM) or kernel SVM is preferable [20]. Knowing and choosing an appropriate algorithm is important, but even more so, is the data available to the algorithm [21].

Input data determines how a machine learning algorithm trains on future data, and therefore influences the overall functionality of the algorithm. This is a particular problem in 2017 because big data, gleaned from billions of Internet consumers, is an enticing bounty. How data

is gathered and preprocessed makes a big difference for fair results, and not all data is created equal. Take, for example, data gathered online to represent emotion recognition in the human face. Typically, online data comes from people who own a computer and use the Internet: primarily middle-aged adults, who may not represent all ethnic diversities proportionally. If a machine learning algorithm takes in this "wild" data to train, it may not test properly for children, the elderly, or ethnic minorities [8]. When machine bias results from a machine learning algorithm, oftentimes the dataset will be the first culprit to verify.

How is it used?

Though admittedly challenging to quantify, determining if bias is present in the results of a machine learning algorithm depends not only on how the model is built, but for what purpose it is used. As an example, in our case study, the COMPAS algorithm was developed and trained on data from a system, some may argue, that is already biased - the American criminal justice system. And, years following its creation, a public defender appealed the use of the COMPAS risk score in a defendant's ultimate sentencing. When the creator of the algorithm, Tim Brennan, was called to testify, he indicated that he never intended the algorithm to be used in sentencing, but had softened his opinion with time [19]. What does it mean for a machine learning algorithm to be created for one purpose and used for another? What if the original intended purpose has less moral implications than the unintended?

What are the ethical considerations?

Unfortunately, there is no one-size-fits-all ethical formula we can insert into a machine learning model and ensure unbiased results. If we are to combat bias using ethics, it will have to be on a case-by-case basis. One of the best methods to inform "ethical corrections" for a given machine learning algorithm is to conduct extensive contextual validity testing, particularly on algorithms

with high legal and ethical impact. Validity testing research of machine learning algorithms is currently very limited, and much of what has been conducted is by the original creators of the algorithm [14]. Indeed, in our case study, the COMPAS algorithm was validity tested in several states *after* already being in use to score defendants for several years. Even then, the algorithm was tested by the company that created it (Northpointe) and by a university on behalf of the Sheriff's office that uses it [19] [22] [23]. ProPublica, a third-party organization conducting a larger analysis of bias in machine learning did a validity test of the algorithm in 2016, and determined it to be racially discriminatory [14] [19]. Similarly to validation testing, studies are conducted to determine the number of defects in software. Oftentimes, these studies are done by the same developers who created the software. The most strongly significant factor in predicting defects are the metrics that research groups use in conducting their tests, because they tend to use the same or similar metrics [24]. This study is particularly encouraging because it shows a level of depth that is gaining traction in the field that can only come with time and experience. Experts are surely on their way to learning more about where we go wrong, why, and what we can do to correct for machine bias.

CASE STUDY

Background

To better understand how machine bias can occur and the ethical considerations that could help reduce bias in the application of machine learning, we will consider the findings of ProPublica, a nonprofit team of investigative journalists, regarding COMPAS, a risk scoring algorithm in use in the American criminal justice system today [25]. The group chose COMPAS because it is one of the most widely used algorithms in pretrial and sentencing in America today [14]. There are other tools in use, that pre-date COMPAS, one of which is the Level of Service Inventory, or

LSA, developed in Canada. Both tools, among others, are still used to varying levels around the country.

COMPAS was first made commercially available by Northpointe, founded by Tim Brennan and Dave Wells in 1989. Brennan, a then professor of Statistics at the University of Colorado, wanted to make an improved assessment tool to use in law enforcement. Northpointe defines recidivism as “a finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code” [14]. The algorithm is trained by compiling the answers to 137 questions from defendants that *do not* include questions about race [19]. Northpointe does not publicly disclose COMPAS calculations and exact algorithmic code. However, in their validity report in 2006, they review the most recent version of their model “4G,” explaining that COMPAS risk and classification models use logistic regression, survival analysis, and bootstrap classification to identify defendant risk scores [22].

Many court systems and law enforcement departments throughout the country use risk assessment tools like COMPAS to help make decisions about defendants. In some locations, the scores are used to determine bail terms, if a defendant should be released or held for bail pretrial. In others, the scores are used in sentencing itself. Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin all use algorithmic risk assessment tools during criminal sentencing. Notably, in many locations, risk assessment tools are implemented *before* testing to be sure they work properly in a given population. [19]

In Broward County, Florida, COMPAS risk scores are used to determine pretrial conditions such as bail or release. ProPublica chose this community because of Florida’s dedication to publicly

releasing criminal data [14]. ProPublica studied over 6,000 defendants with a risk score between 2013 - 2014. The researchers built profiles for each defendant using identifying data like first and last name to match original COMPAS risk scores of recidivism and violent recidivism to Broward County criminal records within the two years following scoring [14].

ProPublica's Results

Overall, ProPublica found that the percentage of medium or high risk scores actually resulting in recidivism were relatively equal between black defendants and white defendants (63% and 59% respectively). It was the way in which incorrect predictions were incorrect that resulted in racial bias [14]. First, in a simple distribution histogram plotting white and black risk of recidivism scores among the ~6,000 defendants, ProPublica found that white defendants' risk scores were skewed lower while black defendants' scores were evenly distributed. Black defendants were twice as likely as white defendants to be misclassified as high risk when incorrectly predicted to recidivate (45% vs. 23%). Similarly, when white defendants *did* recidivate, it was two times more likely that they were misclassified as low risk than black defendants (48% compared to 28% respectively) [14]. Although predictions were roughly even for overall recidivism, when they were incorrect, they were very differently incorrect, most significantly between white and black defendants.

COMPAS models had been validity tested before ProPublica's research by several sources, including researchers at Florida State University and the creators of the algorithm themselves. Both a study by Brennan and the Center for Criminology and Public Policy Research in Florida found that the COMPAS models resulted in significantly accurate predictions [19] [22] [23]. Importantly, ProPublica is not, in fact, disputing this claim, but instead brings to light bias in the *way in which* the algorithm outputs when it gets it wrong. The risk scores are by nature

predictions and not fact. But when they are used without understanding what underlying patterns are present in resulting errors, individual lives can suffer.

Our Results

The question is: how do we correct for the bias in an algorithmic methodology? For the purposes of this paper, we conducted a simple analysis of the ProPublica results. Several more categories of solutions are offered in the following paragraphs, that we identify as a starting point for future research.

As indicated in Figure 1 below, ProPublica identified the false positives and negatives for both black and white defendants in predicted recidivism. It appears that the rates are almost flipped for black versus white defendants. Black defendants are two times more likely to be falsely identified as high risk to recidivate, while white defendants are two times more likely to be falsely identified as low risk when they do recidivate. Typically the difference in higher "allowance" for false positives over negatives, means that there is something about predicting *true* versus *false* that is less costly or problematic. For example, with machine learning algorithms used to help diagnose cancer in patients, if the algorithm is going to get it wrong, less loss of life occurs when the machine errs on the side of predicting patients have cancer (even if they don't). In our case, COMPAS seems to associate higher cost with identifying white defendants as high risk incorrectly, and the opposite for black defendants. Determining the "utility judgement" made by COMPAS that results in these false positive and negative rates, could provide more information about why the bias is occuring in the algorithm.

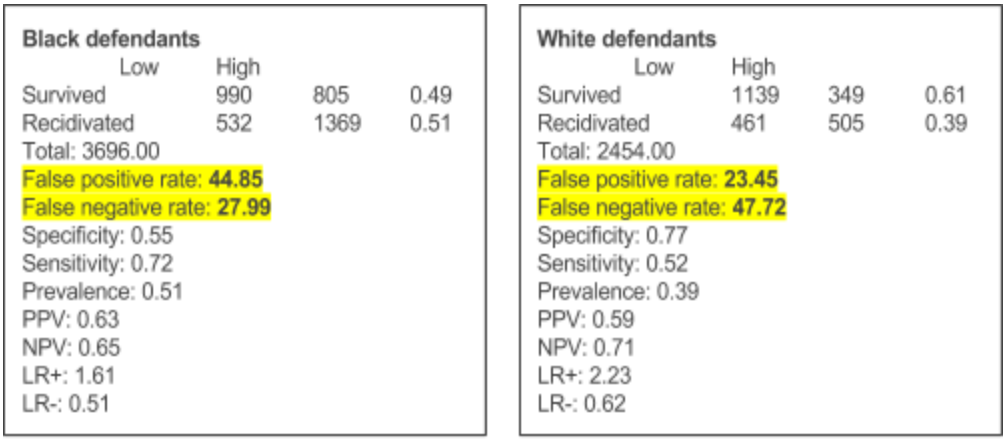

Figure 1

Difference in Black and White Defendants False Negative and Positives

The efforts of ProPublica to understand the COMPAS algorithm and bias that appears in its results are comprehensive despite the lack of transparency around the calculations and data used by Northpointe. In practice, after analyzing a commercially available algorithm like COMPAS and finding reason to believe it is biased, the next step is to reduce the bias for future use. There are technical, political, social, and philosophical solutions to making this possible. Using the right intersection of all pertinent disciplines will accelerate our path to finding trends in ethical solutions for machine bias of all types. More research is needed to test the effectiveness of these solutions, and ultimately, we can identify trends in ethical solutions and categorize them for machine bias of similar types.

THE ETHICAL SOLUTION

Using Ethics to Solve Bias in Machine Learning

Any method used to solve for unfair predictions with legal and ethical consequences, is an example of machine ethics. A perfect solution to bias output in machine learning may never be

possible because of our obligation to rely on imperfect data and the fundamentally prejudiced real-world we live in. Additionally, optimizing total societal good over accuracy in an isolated context is a complicated task. However, it is a moral obligation for computer scientists and those who purchase and employ the use of machine learning algorithms to pursue output distributions that most closely reflect an ideal state of fairness for all impacted individuals. A number of solutions have been either tested or conceptualized in the field of machine ethics to solve for bias in four main categories: technical, social, political, and philosophical.

Technical Solutions to Machine Bias

Programming ethics into machine learning algorithms is not straightforward The source of the bias must first be determined, and options for adjustment can be considered. Research suggests that statistical bias and variance can relate strongly to appropriateness of machine bias. To reduce variance, for algorithms like decision trees, “softer splits” can be used. For example, in a Markov Tree model, data goes down both sides of the split and the leaves of the tree are longer. Additionally, multiple hypotheses may be used to “vote” on the classification of test cases through ensemble and randomization methods. In a study in 1995 by Dietterich and Kong, this resulted in near perfect or perfect reduction of error in test cases. When necessary, statistical bias can be reduced using error-correcting output coding (ECOC) which uses “bit-position hypotheses” and reduces overall error. [6]

When trying to reduce bias in their emotion recognition study, Howard, Zhang, and Horvitz, discovered that a hierarchical classification model worked well to fit to both their majority and minority data. They found that the inclusiveness of a classifier can be manipulated depending on the ethics of the resulting decisions by using a generalized learning algorithm and a specialized learner to correct for bias towards one or more minority class [8]. When testing for

emotion recognition in children's faces, one emotion in particular seemed biased towards misclassification - Fear. Instead, the machine was outputting Surprise far more often than was accurate. So, the researchers broke down the emotion of fear into calculable facial characteristics, and created a specialized learning algorithm with more detailed information to decipher the difference between fear and surprise. This resulted in a 41.5% increase in the recognition rate for the minority class (children), while also increasing the recognition rate of the majority by 17.3% [8].

Kusner and other researchers at the Alan Turing Institute, have identified the difficulty of implementing quantitative changes to algorithms to address ethical issues. The group outlines the "first explicitly causal approach" [10] dealing with ethics in an algorithm that predicts the success of law students in completing law school. Their approach weighs protected attributes (such as ethnicity and gender) differently than other descriptive data, *but does not exclude them from the algorithm*. The method takes social bias into account and then compensates effectively, so as not to ignore protected attributes, but to handle them in a way that is fair.

One of the simplest ways to learn from and ultimately rectify bias in machine learning algorithms is to have rigorous external validation testing conducted on algorithms used in high impact, moral contexts. Experimenting with different datasets and metrics, especially when validity testing their own models, will help researchers find the best balance for internal and external validity in empirical research [24]. For optimized validation testing, machine learning source code should be as transparent as possible, if not open source. Periodic audits of machine learning code will help encourage and maintain accuracy and fairness.

Political Solutions to Machine Bias

Though much slower than the pace of technology, the pervasive impacts of machine learning have started to spur some nations to take regulatory action. In April 2016, the EU passed the General Data Protection Regulation (GDPR) to take effect in 2018, which outlines the requirement of corporations to oblige a "right to explanation" for citizens in the EU. This means that when an algorithmic decision is made about an individual, they have a right to know why. Parts of the regulation also seem to disallow the use of "personal data which are by their nature particularly sensitive" to profile an individual [5]. Although this may not catch on as quickly in the U.S., efforts were made by former President Obama to acknowledge the risk of machine bias. The Obama administration released a report demonstrating the importance of investigating big data and machine learning algorithms to ensure fairness in 2016 [26]. There have been other American political figures supporting this effort, such as former U.S. Attorney General Eric Holder, who called for the U.S. sentencing commission to study risk assessments like COMPAS in 2014 [19].

Political momentum on the subject of machine ethics will only build as the ramifications increase. Creating policies, standards, best practices, and certifications pertaining to machine learning creation and use will help progress the ethical implementation of these technologies.

Social Solutions to Machine Bias

Ethical solutions to machine bias can come, simply, from social awareness. As Cathy O'Neil says in her book, "Weapons of Math Destruction," algorithms may be unduly respected as authoritative and objective because they are mathematical and, therefore, "impenetrable" to lay-users [27]. Aware consumers are better able to demand ethical standards and transparent practices. This, in addition to focus on curating multifaceted technical teams, groups that foster

diverse talent, and “community policing” can all reduce bias. Following a flip in tradition, in 2014, tech giants like Google and Apple released statistics on their workforce diversity. Higher diversity among team members has been shown to increase innovation and profit [5]. Groups such as the Algorithmic Justice League (AJL), founded by Joy Buolamwini, work to promote crowd-sourced reporting and study of bias in machine learning and other technologies [28]. Similarly, the Fairness, Accountability, and Transparency in Machine Learning (FAT ML) workshop, was created by researchers from Google and Microsoft to analyze algorithmic bias and its impacts. When it comes to social pressure, consumers can demand greater transparency in machine learning, by supporting those companies that provide it. Crowdsourcing ethics may be another way to combat the problem. In some gaming communities, moral conflicts find resolution through community voting systems [5]. Involvement from diverse populations in the ethical creation and consumption of machine learning predictions will lead to further progress in ethics that include *all* users.

Philosophical Solutions to Machine Bias

It is imperative to discuss tactical answers to this ethical problem, because it is here now, and impacting individual lives today. But, as with all moral conundrums, there are higher level questions experts continue to posit. For example, Headleand and Teahan’s insightful research delivers an option for bottom-up ethical machines [11]. Though not feasible as an immediate solution to bias today, continuing this research could lead to the opportunity to build more complex and *innately* ethical intelligent agents. Some experts are at complete odds at where humans belong in the mix. On one hand, perhaps we have reached the peak of moral functioning in the machine, and the human is the required final piece to make ethical decisions with all contextual divergences considered [29]. Or, instead, machines give us an opportunity to discover where we truly stand in matters of ethics. Training from human ethicists may be all that

is necessary for the models to do better than an average human when making moral decisions [12]. Exploring these quandaries further will contribute to the body of study, and a deeper understanding of how machine ethics can solve challenging moral problems.

CONCLUSION

We have reviewed the background, challenges, and solutions present in the interdependent fields of machine bias and ethics. Using our case study of the COMPAS algorithm and ProPublica analysis, we developed several courses of further research to expose machine bias in detail to best solve for it. In today's climate, easy access to big data entices many to formulate machine learning algorithms with the intent of solving consumer problems, but not ethical problems. To avoid hard-coding centuries of bias, bigotry, and prejudice into our machines, we *must* pay attention to the programs we develop and train on our input. If we don't, as Laura Weidman Powers, the founder of Code2040 says, "we are running the risk of seeding self-teaching AI with the discriminatory undertones of our society in ways that will be hard to rein in because of the often self-reinforcing nature of machine learning" [5]. Machine ethics is a complicated and multifaceted problem. But if we get it right, we will unleash the full benefit of machine learning for humankind.

REFERENCES

[1] Varshney, K. R., & Alemzadeh, H. (2016, October 5). On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products. Retrieved December 04, 2016, from https://arxiv.org/abs/1610.01256

[2] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. (2017). Counterfactual fairness. arXiv preprint arXiv:1703.06856

[3] Ng, A. (n.d.). What is Machine Learning? - Stanford University. Retrieved December 09, 2017, from https://www.coursera.org/learn/machine-learning/lecture/Ujm7v/what-is-machine-learning

[4] Brownlee, J. (2016, September 21). Parametric and Nonparametric Machine Learning Algorithms. Retrieved December 09, 2017, from https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/

[5] Garcia, M. (2017, January 07). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. Retrieved December 03, 2017, from http://muse.jhu.edu/article/645268/pdf

[6] Dietterich, T. G. & Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.Technical Report, Department of Computer Science, Oregon State University, Corvallis, Oregon. Available from ftp://ftp.cs.orst.edu/pub/tgd/papers/tr-bias.ps.gz.

[7] Reese, H. (2016). Bias in Machine Learning, and How to Stop It. TechRepublic. Retrieved October 9, 2017, from http://www.techrepublic.com/google-amp/article/bias-in-machine-learning-and-how-to-stop-it/

[8] Howard, A., Zhang, C., & Horvitz, E. (2017). Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*. doi:10.1109/arso.2017.8025197

[9] Devlin, H. (2017, April 13). AI programs exhibit racial and gender biases, research reveals. Retrieved October 09, 2017, from https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals

[10] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. (2017). Counterfactual fairness. arXiv preprint arXiv:1703.06856

[11] Headleand, C. J., & Teahan, W. (2016). Towards ethical robots: Revisiting Braitenbergs vehicles. *2016 SAI Computing Conference (SAI),* 469-477. doi:10.1109/sai.2016.7556023

[12] Anderson, Michael & Anderson, Susan. (2007). Machine Ethics: Creating an Ethical Intelligent Agent.. AI Magazine. 28. 15-26.

[13] May, P. (2017, November 21). Watch out for 'killer robots,' UC Berkeley professor warns in video. Retrieved December 09, 2017, from http://www.mercurynews.com/2017/11/20/watch-out-for-killer-robots-uc-berkeley-professor-warns-in-video/

[14] Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How We Analyzed the COMPAS Recidivism Algorithm. Retrieved December 03, 2017, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[15] SIGCAS - Computers & Society. (2003). Retrieved December 05, 2017, from http://www.sigcas.org/awards-1/awards-winners/moor

[16] J.H. Moor. (2006). "The Nature Importance and Difficulty of Machine Ethics", *Intelligent Systems IEEE*, vol. 21, pp. 18-21, 2006, ISSN 1541-1672.

[17] Petrasic, K., Saul, B., & Greig, J. (2017, January 20). Algorithms and bias: What lenders need to know. Retrieved December 05, 2017, from https://www.lexology.com/library/detail.aspx?g=c806d996-45c5-4c87-9d8a-a5cce3f8b5ff

[18] Akhtar, A. (2016, August 09). Is Pokémon Go racist? How the app may be redlining communities of color. Retrieved December 09, 2017, from https://www.usatoday.com/story/tech/news/2016/08/09/pokemon-go-racist-app-redlining-communities-color-racist-pokestops-gyms/87732734/

[19] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. Retrieved December 03, 2017, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[20] Eremenko, K., & De Ponteves, H. (2017, November 02). Machine Learning A-Z™: Hands-On Python & R In Data Science. Retrieved December 05, 2017, from https://www.udemy.com/machinelearning/

[21] Rajaraman, A. (2008, March 24). More data usually beats better algorithms. Retrieved December 09, 2017, from http://anand.typepad.com/datawocky/2008/03/more-data-usual.html

[22] Brennan, T., Dieterich, W., & Ehret, B. (2008). Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior, 36*(1), 21-40. doi:10.1177/0093854808326545

[23] Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). Validation of the COMPAS risk assessment classification instrument. Retrieved from the Florida State University website: http://www.criminologycenter.fsu.edu/p/pdf/pretrial/Broward%20Co.%20COMPAS%20 Validation%202010.pdf

[24] Tantithamthavorn, C., Mcintosh, S., Hassan, A. E., & Matsumoto, K. (2016). Comments on "Researcher Bias: The Use of Machine Learning in Software Defect Prediction". *IEEE Transactions on Software Engineering, 42*(11), 1092-1094. doi:10.1109/tse.2016.2553030

[25] About Us. (n.d.). Retrieved December 05, 2017, from https://www.propublica.org/about/

[26] Smith, M., Patil, D., & Muñoz, C. (2016, May 4). Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights. Retrieved December 03, 2017, from https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights
[27] ONeil, C. (2017). *Weapons of math destruction: how big data increases inequality and threatens democracy*. London: Penguin Books.
[28] AJL -ALGORITHMIC JUSTICE LEAGUE. (n.d.). Retrieved December 05, 2017, from https://www.ajlunited.org/
[29] Steusloff, H. (2016). Humans Are Back in the Loop! Would Production Process Related Ethics Support the Design, Operating, and Standardization of Safe, Secure, and Efficient Human-Machine Collaboration? *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW),* 348-350. doi:10.1109/w-ficloud.2016.76